

ENCYCLOPEDIA OF QUANTITATIVE RISK ANALYSIS AND ASSESSMENT

Editors-in-Chief

EDWARD L. MELNICK

New York University, New York, NY, USA

BRIAN S. EVERITT

Institute of Psychiatry, King's College, London, UK



Contents

VOLUME 1

Absolute Risk Reduction	1	Bonus–Malus Systems	165
Accelerated Life Testing	2	Burn-in Testing: Its Quantification and Applications	170
Actuary	8	1,3-Butadiene	176
Adverse Selection	14	Cancer Risk Evaluation from Animal Studies	181
Air Pollution Risk	15	Canonical Modeling	189
Alternative Risk Transfer	21	Case–Control Studies	192
Arsenic	28	Causal Diagrams	205
As Low as Reasonably Practicable/ As Low as Reasonably Achievable	38	Causality/Causation	211
Asbestos	43	Censoring	221
Assessment of Risk Association with Beryllium Exposure	52	Change Point Analysis	225
Asset–Liability Management for Life Insurers	60	Chernobyl Nuclear Disaster	227
Asset–Liability Management for Nonlife Insurers	63	Clinical Dose–Response Assessment	233
Association Analysis	67	Coding: Statistical Data Masking Techniques	240
Attributable Fraction and Probability of Causation	73	Cohort Studies	243
Availability and Maintainability	77	Collective Risk Models	255
Axiomatic Measures of Risk and Risk-Value Models	84	Combining Information	259
Axiomatic Models of Perceived Risk	94	Common Cause Failure Modeling	264
Basic Concepts of Insurance	105	Comonotonicity	274
Bayes' Theorem and Updating of Belief	107	Comparative Efficacy Trials (Phase III Studies)	279
Bayesian Analysis and Markov Chain Monte Carlo Simulation	111	Comparative Risk Assessment	283
Bayesian Statistics in Quantitative Risk Assessment	119	Compensation for Loss of Life and Limb	292
Behavioral Decision Studies	136	Competing Risks	301
Benchmark Analysis	145	Competing Risks in Reliability	307
Benchmark Dose Estimation	149	Compliance with Treatment Allocation	312
Benzene	154	Computer Security: A Historical Perspective	321
Bioequivalence	159	Condition Monitoring	341
Blinding	163	Conflicts, Choices, and Solutions: Informing Risky Choices	346
		Confounding	353
		Considerations in Planning for Successful Risk Communication	362
		Continuous-Time Asset Allocation	366

Copulas and Other Measures of Dependency	372	Environmental Health Risk	574
Correlated Risk	374	Environmental Health Risk Assessment	580
Cost-Effectiveness Analysis	375	Environmental Health Triage, Chemical Risk Assessment for	589
Counterterrorism	387	Environmental Monitoring	591
Credibility Theory	390	Environmental Performance Index	602
Credit Migration Matrices	401	Environmental Remediation	606
Credit Risk Models	407	Environmental Risk Assessment of Water Pollution	611
Credit Scoring <i>via</i> Altman Z-Score	412	Environmental Risk Regulation	613
Credit Value at Risk	416	Environmental Risks	617
Cross-Species Extrapolation	419	Environmental Security	622
Cumulative Risk Assessment for Environmental Hazards	425	Environmental Tobacco Smoke	627
Cyber Risk Management	429	Epidemiology as Legal Evidence	630
VOLUME 2		Equity-Linked Life Insurance	637
Data Fusion	441	Estimation of Mortality Rates from Insurance Data	643
Decision Analysis	447	Ethical Issues in Using Statistics, Statistical Methods, and Statistical Sources in Work Related to Homeland Security	647
Decision Conferencing/Facilitated Workshops	451	Evaluation of Risk Communication Efforts	650
Decision Modeling	459	Experience Feedback	656
Decision Trees	462	Expert Elicitation for Risk Assessment	661
Default Correlation	470	Expert Judgment	667
Default Risk	476	Extreme Event Risk	676
Degradation and Shock Models	481	Extreme Value Theory in Finance	678
Dependent Insurance Risks	489	Extreme Values in Reliability	686
Design of Reliability Tests	493	Extremely Low Frequency Electric and Magnetic Fields	691
Detection Limits	498	Failure Modes and Effects Analysis	695
Digital Governance, Hotspot Detection, and Homeland Security	502	Fair Value of Insurance Liabilities	700
Dioxin Risk	510	Fate and Transport Models	705
Disease Mapping	513	Fault Detection and Diagnosis	714
Distributions for Loss Modeling	516	Federal Statistical Systems as Data Sources for Intelligence, Investigative, or Prosecutorial Work	718
Dose-Response Analysis	523	Fraud in Insurance	723
Dynamic Financial Analysis	528	From Basel II to Solvency II – Risk Management in the Insurance Sector	727
Early Warning Systems (EWSs) for Predicting Financial Crisis	535	Game Theoretic Methods	733
Ecological Risk Assessment	539	Gene-Environment Interaction	737
Economic Criteria for Setting Environmental Standards	542	Geographic Disease Risk	744
Effect Modification and Interaction	553	Global Warming	748
Efficacy	556		
Engineered Nanomaterials	558		
Enterprise Risk Management (ERM)	559		
Environmental Carcinogenesis Risk	566		
Environmental Exposure Monitoring	567		
Environmental Hazard	572		

Group Decision	760	Managing Foodborne Risk	1029
Hazard and Hazard Ratio	771	Managing Infrastructure Reliability, Safety, and Security	1033
Hazardous Waste Site(s)	777	Managing Risks of Consumer Products	1040
Hazards Insurance: A Brief History	784	Markov Modeling in Reliability	1045
Health Hazards Posed by Dioxin	786	Mathematical Models of Credit Risk	1049
Hexavalent Chromium	796	Mathematics of Risk and Reliability: A Select History	1054
History and Examples of Environmental Justice	808	Mercury/Methylmercury Risk	1062
History of Epidemiologic Studies	817	Meta-Analysis in Clinical Risk Assessment	1064
Homeland Security and Transportation Risk	830	Meta-Analysis in Nonclinical Risk Assessment	1075
Hormesis	838	Methylmercury	1084
Hotspot Geoinformatics	844	Microarray Analysis	1089
Human Reliability Assessment	851	Mobile Phones and Health Risks	1092
Imprecise Reliability	875	Model Risk	1094
Individual Risk Models	881	Modifiable Areal Unit Problem (MAUP)	1098
Inequalities in Risk Theory	888	Molecular Markers and Models	1101
Inferiority and Superiority Trials	895	Monte Carlo Methods	1104
Influence Diagrams	897	Moral Hazard	1107
Informational Value of Corporate Issuer Credit Ratings	910	Multiattribute Modeling	1109
Insurability Conditions	915	Multiattribute Utility Functions	1114
Insurance Applications of Life Tables	916	Multiattribute Value Functions	1119
Insurance Pricing/Nonlife	922	Multistate Models for Life Insurance Mathematics	1128
Integration of Risk Types	933	Multistate Systems	1139
Intensity Modeling: The Cox Process	937	Multivariate Reliability Models and Methods	1144
Intent-to-Treat Principle	940		
Interpretations of Probability	943		
VOLUME 3		Natural Resource Management	1149
Kriging	953	Near-Miss Management: A Participative Approach to Improving System Reliability	1154
Large Insurance Losses Distributions	961	No Fault Found	1163
Latin Hypercube Sampling	969	Non- and Semiparametric Models and Inference for Reliability Systems	1167
Lead	973	Nonlife Insurance Markets	1176
Lévy Processes in Asset Pricing	978	Nonlife Loss Reserving	1180
Life Insurance Markets	983	Nonparametric Calibration of Derivatives Models	1187
Lifetime Models and Risk Assessment	987	Nuclear Explosion Monitoring	1191
Linkage Analysis	991	Numerical Schemes for Stochastic Differential Equation Models	1197
Logistic Regression	1003		
Longevity Risk and Life Annuities	1014		
Low-Dose Extrapolation	1019		
Maintenance Modeling and Management	1023	Occupational Cohort Studies	1203
		Occupational Risks	1214

Odds and Odds Ratio	1219	Randomized Controlled Trials	1415
Operational Risk Development	1220	Recurrent Event Data	1419
Operational Risk Modeling	1224	Reinsurance	1425
Optimal Risk-Sharing and Deductibles in Insurance	1230	Relative Risk	1429
Optimal Stopping and Dynamic Programming	1236	Reliability Data	1430
Options and Guarantees in Life Insurance	1244	Reliability Demonstration	1438
Ordering of Insurance Risk	1250	Reliability Growth Testing	1443
		Reliability Integrated Engineering Using Physics of Failure	1448
Parametric Probability Distributions in Reliability	1255	Reliability of Consumer Goods with "Fast Turn Around"	1462
Persistent Organic Pollutants	1261	Reliability of Large Systems	1466
Pesticide Risk	1271	Reliability Optimization	1471
Pharmaceuticals in the Environment	1273	Remote Sensing	1477
Players in a Decision	1277	Repair, Inspection, and Replacement Models	1481
Point Source Modeling	1278	Repairable Systems Reliability	1493
Polychlorinated Biphenyls	1281	Repeated Measures Analyses	1500
Potency Estimation	1285	Risk and the Media	1508
Precautionary Principle	1290	Risk Attitude	1512
Precautionary Principles: Definitions, Issues, and Implications for Risky Choices	1293	Risk Characterization	1524
Premium Calculation and Insurance Pricing	1302	Risk Classification in Nonlife Insurance	1530
Pricing of Life Insurance Liabilities	1314	Risk Classification/Life	1535
Privacy Protection in an Era of Data Mining and Record Keeping	1322	Risk from Ionizing Radiation	1540
Probabilistic Design	1325	Risk in Credit Granting and Lending Decisions: Credit Scoring	1546
Probabilistic Risk Assessment	1349	Risk Management of Construction Defects	1553
Product Risk Management: Testing and Warranties	1356	Risk Measures and Economic Capital for (Re)insurers	1556
Professional Organizations in Risk Management	1360	Risk-Benefit Analysis for Environmental Applications	1566
Protection of Infrastructure	1364	Risk-Neutral Pricing: Importance and Relevance	1569
Public Health Surveillance	1371	Role of Alternative Assets in Portfolio Construction	1574
Public Participation	1381	Role of Risk Communication in a Comprehensive Risk Management Approach	1584
Quantitative Reliability Assessment of Electricity Supply	1393	Ruin Probabilities: Computational Aspects	1587
		Ruin Theory	1591
VOLUME 4			
R&D Planning and Risk Management	1399	Safety	1599
Radioactive Chemicals/Radioactive Compounds	1403	Sampling and Inspection for Monitoring Threats to Homeland Security	1600
Radon	1409	Scenario Simulation Method for Risk Management	1603
Radon Risk	1411		

Scenario-Based Risk Management and Simulation Optimization	1611	Threshold Models	1769
Scientific Uncertainty in Social Debates Around Risk	1618	Toxic Torts: Implications for Risk Management	1771
Securitization/Life	1623	Toxicity (Adverse Events)	1777
Simulation in Risk Management	1627	Truncation	1779
Social Networks	1634	Ultrahigh Reliability	1783
Societal Decision Making	1640	Uncertainty Analysis and Dependence Modeling	1789
Software Testing and Reliability	1642	Uncertainty and Variability Characterization and Measures in Risk Assessment	1797
Soil Contamination Risk	1653	Understanding Large-Scale Structure in Massive Data Sets	1804
Solvency	1657	Use of Decision Support Techniques for Information System Risk Management	1811
Spatial Risk Assessment	1663	Utility Function	1818
Spatiotemporal Risk Analysis	1665	Value at Risk (VaR) and Risk Measures	1823
Stakeholder Participation in Risk Management Decision Making	1668	Value Function	1830
Statistical Arbitrage	1672	Volatility Modeling	1832
Statistics for Environmental Justice	1678	Volatility Smile	1839
Statistics for Environmental Mutagenesis	1682	Vulnerability Analysis for Environmental Hazards	1845
Statistics for Environmental Teratogenesis	1687	Warranty Analysis	1849
Statistics for Environmental Toxicity	1692	Water Pollution Risk	1857
Stochastic Control for Insurance Companies	1697	Wear	1862
Stratospheric Ozone Depletion	1702	Weather Derivatives	1868
Stress Screening	1704	What are Hazardous Materials?	1873
Structural Models of Corporate Credit Risk	1707	Zero Failure Data	1879
Structural Reliability	1712		
Structured Products and Hybrid Securities	1715		
Subjective Expected Utility	1719		
Subjective Probability	1724		
Supra Decision Maker	1734		
Survival Analysis	1736		
Syndromic Surveillance	1743		
Systems Reliability	1753		

Foreword

Risk means different things to different people, and even within the technical realm of statistics it has a multiplicity of meanings. Different dictionaries and encyclopedias define it differently. A common definition found in multiple places on the WWW and elsewhere is that risk is “a potential negative impact to an asset or some characteristic of value that may arise from some present process or future event”. In everyday usage, risk is often used synonymously with the probability or expectation of a (known) loss. Implicitly, for risk to exist, there must be uncertainty about a known quantity or about an event that has an unknown outcome (e.g., see the links in statistics among risk, utility, loss, and decision-making under uncertainty).

Risk plays a strong role in many of today’s most pressing problems such as the protection of privacy, bio-surveillance, and the detection and control of environmental hazards. Statistical notions are inherent in almost all aspects of risk analysis and risk assessment. The *Encyclopedia of Quantitative Risk Analysis and Assessment* brings together a diverse collection of articles dealing with many of these different perspectives of risk and its assessment. Virtually every topic from modern statistical methodology has found its way into the entries ranging from Causation and Confounding, through Bayesian Analysis and Markov Chain Monte Carlo Simulation, Randomized Controlled Trials, Social Networks, and Survival Analysis. Each is examined from the perspective of risk assessment as well as from the perspective of specific health and other hazards, such as those associated with substances such as arsenic, asbestos, and lead, as well as with global warming, ozone depletion, and water pollution. The result is a fascinating collection of articles that spans the field of statistics and its application to human affairs, including work in cognate fields such as epidemiology, ethics, finance, insurance, and the legal theory of toxic torts.

These topics and especially their application to risk assessment cannot be found in a single or even a handful of standard sources, and readers would normally need to consult multiple books and encyclopedias to gain access to authoritative descriptions of the technical topics and their relevance. The Editors of the *Encyclopedia of Quantitative Risk Analysis and Assessment* have assembled a series of entries from distinguished experts on an interlocking set of topics and it should prove to be a major reference in this domain for years to come as members of society collectively attempt to cope with the multiplicity of aspects of risk, its measurement, and its assessment.

Stephen E. Fienberg

Department of Statistics, Machine Learning Department, and Cylab
Carnegie Mellon University
Pittsburgh, PA, USA

July 2008

Preface

Exposure to risk permeates our daily experiences. The assessment of risk and strategies to reduce the likelihood of suffering harm or loss is the concern of all scientific endeavors. However, the study of risk has never evolved into its own area with its own language and methodologies. Instead, risk analysis is a cross-cutting topic combining areas that might include such diverse subjects as engineering, medicine, finance theory, public policy, and the military. Independent of the specific areas of applications, the core ideas behind risk assessment and analysis are essentially the same. A solution depends upon the probabilities of the occurrence of a set of potential problems, the probabilities of different levels of catastrophes being realized if a particular problem occurs, and a loss function associated with the cost of each catastrophe. The challenge is the quantification of the probabilities and costs to specific problems for setting policies that minimize costs while maximizing benefits.

Different disciplines have met those challenges in a variety of ways. A few have explicitly built upon the large body of statistical work subsumed in probabilistic risk assessment, but most have not. Many have developed alternative strategies that are robust to specific kinds of uncertainty, or handle adversarial situations, or deal with dynamically changing action spaces such as decision making in an economic environment. These kinds of diverse settings have broadened risk analysis beyond the traditional mathematical formulations.

Currently, the relevant literature on risk assessment is scattered in professional journals and books, and is not readily accessible to those who would profit from it. The aim of the encyclopedia is to draw together these varied intellectual threads in the hope that risk analysts in one area can gain from the experience and expertise of those in other disciplines. Corporate risk assessment, for example, may learn from military solutions; the work on monitoring for adverse health events might help to inform the early detection of unsafe automobiles; and portfolio management is very likely to be relevant to public policy investments.

Quantitative risk assessment is an important growing component of the larger field of risk assessment that includes priority setting and the management of risk. The statistical theory within the encyclopedia is designed to unify the study of risk by presenting the underpinnings of risk management within the context of the special features of particular areas. Applications from such diverse areas as drug safety, investment theory, public policy applications, transportation safety, public perception of risk, epidemiological risk, national defense and security, critical infrastructure, and program management are included to illustrate this unification.

Further, the need to understand the risks of an activity has spawned new classes of mathematical techniques for hazard identification, dose-response assessment, exposure assessment, and risk characterization. These concepts are discussed in the encyclopedia and illustrated with applications that require a more general description of variability and uncertainty inherent in the risk identification process than were available in the classical statistical literature.

The need for a unifying authoritative reference work on risk assessment was recognized in 2005 by John Wiley and Sons and the editors-in-chief. The project began with the identification

of 10 major areas that have contributed to the literature of developing strategies for studying risk. For each area, a section editor was recruited with known expertise in his or her field. These categories, and the section editors, are:

1. Risk management (Tony Cox)
2. Environmental risk (Walt Piegorsch)
3. Insurance/actuarial risk (Michel Denuit)
4. Financial/credit risk (Ngai Chan)
5. Toxic substances/chemical risk (Dennis Paustenbach and Jennifer Roberts)
6. Reliability (Frank Coolen and Leslie Walls)
7. Bayesian methods/decision analysis (Simon French)
8. Clinical risk (Susan Sereika)
9. Epidemiology/public health (Susan Sereika)
10. Homeland security (Edward Melnick)

Once recruited, editors-in-chief and section editors developed a list of entries within each topic, and from there the section editors solicited authors. The entries were chosen to provide a broad coverage of methods and applications of risk quantification and analysis. Some entries were chosen to cover material at a basic level, while others were at a sophisticated mathematical level. The goal was to make the encyclopedia meet the needs of a wide readership with articles that differed in technical level and mathematical content.

The section editors did an outstanding job of soliciting leading authors to write authoritative articles on selected key topics. Out of approximately 400 potential authors, they were able to recruit all but a handful. We would like to thank all of them for their very substantive and insightful contributions.

Once the first draft of manuscripts was received, the editors-in-chief read and made suggestions to the section editors. The section editors then worked with the authors as supporters and editors. Once the section editors accepted the final version of the manuscripts, the editors-in-chief set up the cross-referencing and compiled all the articles into the encyclopedia.

Any success that the encyclopedia may have will be due, in no small measure, to the efforts of the nearly 360 authors. They all worked very hard, often through multiple drafts, to produce this work. With the editors-in-chief on different sides of the Atlantic, and section editors and authors spread all over the world, this work could not have been carried out without the Internet. We want to thank all who responded promptly to messages sent in the middle of the night (their time) or just after they went home for the weekend.

This encyclopedia uses a system of cross-referencing to make the material more accessible to the reader. Firstly, other articles that the reader might be interested in are cross-referenced using the words “see” or “see also” in parentheses at the appropriate location within the text of an article. For example, ‘...of the air pollutant in ambient air (*see* **Air Pollution Risk**)’. Secondly, related articles are listed at the end of many of the articles, under the heading “Related Articles”. Finally, there are a number of “blind entries” that refer the reader to a full-length article. For example, ‘**Polypatterns: See Remote Sensing,**’ or ‘**Epistemic Communities: See Scientific Uncertainty in Social Debates around Risk.**’

Jill Hawthorne, who was our Project Editor at Wiley, did a superb job leading us along, providing encouragement where it was needed, and somehow keeping complete track of over 300 articles in various stages of completion. We could not have done this without her. We are also grateful for the assistance offered by her colleagues at Wiley, including Daniel Finch, Layla Harden, Debbie Allen and Tony Carwardine all of whom did sterling work on the project.

We also acknowledge the contribution of Sangeetha and the team at Laserwords Private Ltd., who were responsible for the copyediting and typesetting of all the articles.

John Wiley & Sons, Ltd was supportive of this work throughout. They provided funds for several meetings of the editors-in-chief, and made much of the editorial process almost completely transparent to us.

Brian S. Everitt
Edward L. Melnick

July 2008

Absolute Risk Reduction

Absolute Risk

Risk is defined in the *Shorter Oxford English Dictionary* (fifth edition, 2002) as “the possibility of incurring misfortune or loss”. Quantifying and assessing risk involve the calculation and comparison of probabilities, although most expressions of risk are compound measures that describe both the probability of harm and its severity. Americans, for example, run a risk of about 1 in 4000 of dying in an automobile accident. The probability for lethally severe injuries is 1 out of 4000.

Perceptions of risks are influenced by the way the risks are presented. You might be worried if you heard that occupational exposure (*see* **Occupational Cohort Studies**) at your job doubled your risk of serious disease compared to the risk entailed while working at some other occupation: you might be less worried if you heard that your risk had increased from one in a million to two in a million. In the first case, a relative risk is presented, in the second an absolute risk. Relative risk is generally used in medical studies investigating possible links between a risk factor and a disease – it is an extremely important index of the strength of the association between the factor and the disease (*see* **Logistic Regression**), but it has no bearing on the probability that an individual will contract the disease. This may explain why airplane pilots, who presumably have relative risks of being killed in airplane crashes that are of the order of a 1000-fold greater than the rest of us occasional flyers, can still sleep easy in their beds. They know that their absolute risk of being a victim of a crash remains extremely small.

A formal definition of *absolute risk* in the context of medicine is provided by Benichou [1];

The probability that a disease-free individual will develop a given disease over a specified time interval, given current age and individual risk factors, and in the presence of competing risks.

Absolute risk is a probability and therefore, lies between 0 and 1.

Absolute Risk Reduction

Absolute risk can be used as the basis of a measure of effective size of an intervention in a clinical trial. The

measure is called the *absolute risk reduction* (ARR), which is simply the absolute difference in outcome rates between the control and treatment groups. For example, in a clinical trial of mammography, it was found that out of 129 750 women who were invited to begin having mammograms in the late 1970s and early 1980s, 511 died of breast cancer over the next 15 years, a death rate of 0.4%. In the control group of 117 260 women who were not invited to have regular mammograms, there were 584 breast cancer deaths over the same period, a death rate of 0.5% (*see* [2]). So here the estimated ARR is 0.1%. Using the same example, the *relative risk reduction* (RRR), the proportional reduction in the relative risk amongst treated and controls, is 20%. RRR is often more impressive than ARR and the lower the event rate in the control group, the larger will be the difference between the two measures. But for the individual patient, it is ARR that often matters most.

Number Needed to Treat

Over the last 10 years or so, it has become an increasingly popular practice among clinicians to express the ARR in a way that they see as being easier for their patients to understand: one such approach is to use *number needed to treat* (NNT) which is simply the reciprocal of ARR. It gives the estimated number of patients who needed to undergo the new treatment rather than the standard one, for one additional patient to benefit. For example, in a study of the effectiveness of intensive diabetes therapy on the development and progression of neuropathy, 9.6% of patients randomized to usual care and 2.8% of patients randomized to intensive therapy suffered from neuropathy. Consequently, the ARR is 6.8% leading to an NNT of 14.7 ($= 1/6.8\%$). Rounding this figure up, the conclusion is that 15 diabetic patients need to be treated with intensive therapy to prevent one from developing neuropathy (this example comes from the web site of the Centre for Evidence Based Medicine: <http://www.cebm.net/>).

Altman [3] shows how to calculate a confidence interval for NNT, although this is not considered helpful if the 95% confidence interval for ARR includes the value zero, as this gives rise to a nonfinite confidence interval for NNT. There have been some criticisms of NNT (*see*, for example, Hutton [4]), but Altman and Deeks [5] defend the concept as a useful communications tool when presenting results from

clinical studies to patients. (The concept of NNT can equally well be applied to harmful outcomes as well as those that are beneficial, when instead it is referred to as the *number needed to harm*.)

References

- [1] Benichou, J. (2005). Absolute risk, in *Encyclopedia of Biostatistics*, 2nd Edition, P. Armitage & T. Colton, eds, John Wiley & Sons, Chichester.
- [2] Nystrom, L. (2002). Long term effects of mammography screening: updated overview of the Swedish randomized trials, *Lancet* **359**, 909–919.
- [3] Altman, D.G. (1998). Confidence intervals for the number needed to treat, *British Medical Journal* **317**, 1309–1312.

- [4] Hutton, J.L. (2000). Number needed to treat: properties and problems, *Journal of the Royal Statistical Society, Series A* **163**(3), 381–402.
- [5] Altman, D.G. & Deeks, J.J. (2000). Number needed to treat: properties and problems, *Journal of the Royal Statistical Society, Series A* **163**, 415–416.

Related Articles

Axiomatic Measures of Risk and Risk-Value Models

Odds and Odds Ratio

BRIAN S. EVERITT

Accelerated Life Testing

General Description of Accelerated Life Testing

Accelerated testing (*see Reliability Growth Testing; Reliability Demonstration; Reliability of Consumer Goods with “Fast Turn Around”; Reliability Integrated Engineering Using Physics of Failure*) is a set of methods intended to ensure product reliability during design and manufacturing and after product fielding. In all accelerated testing approaches, stress is applied to promote failure. The applied stresses can be environmental (e.g., temperature, humidity, vibration, shock, corrosive gasses, and electromagnetic), application loading (e.g., duty cycle and configuration), or interface loading (e.g., varying line voltage, repeated electrical connector removal and installation, and mounting base flexure). By the term accelerated, a shortened time to failure is implied. However, accelerated tests are commonly used for two very different purposes. If the intention is to use accelerated testing to predict product lifetime [1], the stress must produce the same degradation mechanism and failure types that would be encountered during the product's intended use. If the intention is to use accelerated testing for new product qualification or for manufacturing quality verification, then the stress applied may be unrelated to expected field stress.

In order to make a valid inference about the normal lifetime of the system from the accelerated data, it is necessary to know the relationship between time to failure and the applied stress. This will, of course, depend upon the nature of the system and the nature of the stress applied. Typically, a parametric statistical model of the time to failure and of the manner in which stress accelerates aging is used.

Statistical Models of Accelerated Life Testing

Analysis of accelerated life test data requires acceleration models (or stress relationship models (*see Hazard and Hazard Ratio*)) that relate time acceleration to stress variables like voltage, temperature, physical stress, and chemical environment. A statistical lifetime distribution is commonly used to model

the lifetime to failure mechanism. Frequently used distributions are lognormal, Weibull, exponential, and logistic. Published successful applications should be used as reference for initial model selection considerations. Validation of the model can only be performed after the distribution of time to failure is observed under normal (unstressed) conditions.

Life Stress Relationship Models

Arrhenius. The Arrhenius lifetime relationship (*see Comparative Risk Assessment*) is most commonly used to model product lifetime as a function of temperature. It is one of the earliest and most successful acceleration models that predict how time-to-failure changes with temperature. The Arrhenius rate law states the relationship between the rate of chemical reaction and temperature as

$$R(T) = A \exp \left[\frac{-E}{kT} \right] \quad (1)$$

where R is the chemical reaction rate and T is the temperature in absolute scale [K], k is Boltzmann's constant (8.63×10^{-5} eV K⁻¹), and E is the activation energy in electron volts (eV). The parameter A is a scaling factor determined by product or material characteristics. Activation energy, E , is a critical parameter in the model that depends on the failure mechanism and materials involved. The Arrhenius acceleration factor is given by

$$AF = \exp \left[\frac{E}{k} \left(\frac{1}{T_1} - \frac{1}{T_2} \right) \right] \quad (2)$$

where $T_2 > T_1$ (in Kelvin).

Inverse Power Law. An inverse power relationship is commonly used to model product lifetime as a function of an acceleration stress. The relationship is also called the *inverse power rule* or *inverse power law* or *simply power law*. The model can be written as

$$R(V) = AV^{-\beta} \quad (3)$$

where V is the stress.

Voltage Models. Temperature and voltage stress can be modeled with the following relationships:

$$R(V) = Ae^{\frac{E}{kT}} V^{-\beta} \quad (4)$$

$$R(V) = Ae^{\frac{E}{kT}} e^{-BV} \quad (5)$$

Outcome Variables

Failure is defined as a preset amount of degradation of a system as well as the catastrophic failure of the unit. In accelerated life testing, catastrophic failure modes such as fire and mechanical collapse are often avoided and replaced by an outcome variable passing a certain threshold. This outcome variable has to be materially linked to the final catastrophic failure mode. A few examples of this concept for accelerated life testing are as follow. For crack accelerated fatigue-crack growth, the crack reaching a specified length may be considered a failure. For fire risk, the system elevating to the autoignition temperature of the material would be defined as failure. A system where the failure may not be characterized by a measured characteristic might be death by disease.

The failure criterion variable used must be directly related to the failure mode at hand, and should be selected for both relevance and ease of measurement during testing. The failure criterion variable must be different from the acceleration variable, i.e., if the time to reach auto ignition temperature is the failure criterion, temperature cannot be the accelerating variable; another relevant variable such as current or chemical exposure must be chosen.

Once the failure criterion variable is decided on, the appropriate range of acceleration variables must be selected. This process is not trivial and must generally include a few quick experiments. The first consideration is to whether there are any natural limitations on the degree of acceleration that is possible while maintaining the same physical mechanisms. For example an accelerated thermal test cannot exceed the melting temperatures of the materials involved and still be relevant. The same would be true for glass transition temperatures in polymers and current levels that would begin to melt conductors. Once these natural limitations on acceleration are defined, it is advisable to run a short experiment with the acceleration variable just under this limit in order to define the fastest feasible accelerated time to failure. It is imperative that an accelerated failure has actually been forced before choosing the range of acceleration variables for the testing.

Once the time of the fastest feasible time to failure is established, models of acceleration can be used to design an experiment that will be expected to develop a failure within the time frame of testing, generally

between 1 day and 1 month. It is desirable to use as little acceleration as possible to develop high-fidelity results within the time period available for testing.

Once the failure criterion variable has been defined, the times to failure for the population are generally defined by one of the following methods that allow the entire population to be characterized by a single outcome variable.

Mean Time between Failure. For constant failure rate of nonrepairable components or systems, mean time between failure (MTBF) (*see Reliability Growth Testing; Imprecise Reliability; Availability and Maintainability*) is the inverse of the failure rate. For example, if a component has a failure rate of 20 failures per million hours, the MTBF would be the inverse of that failure rate.

$$MTBF = (1\,000\,000\text{ h})/(20\text{ failures}) = 50\,000\text{ h} \quad (6)$$

Percent Failed at Specified Time. Percent failed at specified time (*see Comparative Risk Assessment*) is often computed as the cumulative probability of failure. Manufacturers or product developers are often more interested in knowing percent failed at specified time rather than MTBF. For example, one may want to know what proportion of the products will fail at the end of 1, 2, or 3 years. Once a parametric model, such as lognormal or Weibull is fit, the percent of product failed at a specified time can be obtained through the model. For example, an exponential failure distribution with MTBF of $\theta = 13.58$ years would give a cumulative percent of failure of 7.1% at the end of the first year. The calculation can be written as

$$P = 1 - \exp\left(\frac{-1}{13.58}\right) = 0.071 \quad (7)$$

Time to Specified Percent Failure. Time to specified percent failure (also known as *percentile of failure*) is commonly calculated from the inverse function of a probability function. For example, the inverse function of exponential distribution is written as

$$\text{Percentile} = -\theta \ln(1 - p) \quad (8)$$

where p is the cumulative probability or percent of failure, and θ is the MTBF. When $\theta = 13.58$ years

and $p = 1\%$, the corresponding percentile is 0.14 year or 1.64 months. For specified values of the parameters, percentiles from lognormal or Weibull distributions or other time-to-failure distributions can be computed.

Regression Methods

Regression methods have become increasingly common tools for analyzing accelerated failure test data, as more statistical software packages for personal computers have made these options available to analysts in recent years. Typically, the lifetime distribution is accelerated by making a linear transformation to time to event in which the accelerating factors are independent variables in the regression. More than one accelerating factor can be used. The acceleration factors may be controlled (e.g., temperature and voltage) in the experiment, but may also include uncontrolled factors (e.g., manufacturing line, operator, and batch of raw materials). Regression methods allow us to build lifetime models to account these factors and make failure predictions properly using more information. Regression methods are generally classified into two approaches: parametric models and the proportional hazards model.

Parametric. In these models, the lifetime distribution under normal operating conditions (unaccelerated) is assumed to follow a parametric distribution. Typical lifetime regression software [2] will support Weibull, lognormal, and other common lifetime distributions.

If T_0 is the time to failure under normal operating conditions, and $\mathbf{X}$ is a vector of factors that accelerate the time to failure and $\boldsymbol{\beta}$ is a vector of unknown coefficients, then the accelerated lifetime model is conveniently specified as $\Pr\{T > t|\mathbf{X}\} = \Pr\{T_0 > \exp(-\mathbf{X}'\boldsymbol{\beta})t\}$. In other words, conditional upon the values of the acceleration variables $\mathbf{X}$, $T = \exp(\mathbf{X}'\boldsymbol{\beta})T_0$. This model will allow multiple acceleration variables, e.g., stress and temperature and their interaction. The model formulation is quite general; covariates to describe differences in device design or differences in manufacturer can be added to the model.

Proportional Hazards. The proportional hazards model or Cox proportional hazards regression model

(see **Lifetime Models and Risk Assessment; Competing Risks**) is widely used in biomedical applications and especially in the analysis of clinical trial data. The model can also be used as an accelerated life testing model. The model does not assume a form for the distribution, because the form of hazard function is unspecified. It cannot be used to extrapolate lifetime in the time scale. This limits the model's ability to make early or late lifetime prediction beyond the range of the data. However, it can extrapolate lifetime in stress. The distribution-free feature makes it attractive when other parametric distributions fail to model the data adequately [3].

Bayesian Methods for Accelerated Life Testing

Meeker and Escobar [4] provide an interesting example of the use of Bayesian methods in accelerated life testing (see **Mathematics of Risk and Reliability: A Select History**) using the Arrhenius model of accelerated lifetime. When activation energy is estimated from the experimental data, the confidence bounds on the estimated distribution of time to failure are generally much larger than the case when the activation energy is known. Bayesian analysis provides a useful middle ground. A prior distribution for the activation energy is provided, and the usual Bayesian mechanics are applied. Even a diffuse prior distribution over a finite range will produce tighter confidence bounds than non-Bayesian methods with no assumption about the activation energy.

History of Accelerated Life Testing

Since accelerated testing is expensive, a cost-benefit justification is often required [5]. Justification of reliability work is not new. In 1959, US Air Force Major General William Thurman's presentation about cold war ballistic missile reliability said that some people at that time felt that "reliability was a bunch of hokum and that the Air Force was making a fetish out of it". The general provided a useful operative definition of reliability, as "the probability that the product or system will perform its required action under operational conditions for a given period of time" [6]. In 1959, General Thurman was concerned about achieving 99% reliability of an intercontinental ballistic missile, but today there are nonmilitary products such as medical products and data centers that require

even higher reliability than what General Thurman required. Although products such as consumer entertainment products do not inherently require high reliability, warranty costs often drive large investments in reliability. Technology has progressed since 1959, but designing an accelerated testing program as part of an overall reliability program still faces challenges similar to those outlined by General Thurman in 1959.

Reliability and Quality

Reliability can be briefly summarized for the purpose of the discussion of accelerated testing. Throughout the history of accelerated testing, two fundamentally different top-level approaches to reliability prediction and management for complex systems have developed. One approach divides product lifetime into three curves forming “the bathtub curve” (see **Burn-in Testing: Its Quantification and Applications**), a curve with the product’s early failure rate dropping to a steady level followed by a rapid rise. The life phases in the bathtub curve are “infant mortality”, a “useful lifetime”, and “wear out”. Much of the reliability program emphasis is placed on understanding where these phases begin and end in the product lifetime. Strategies are applied to each phase. The typical goal is to run through infant mortality in the factory using screening methods such as burn-in. The useful life phase is assumed to be governed by the exponential distribution, with a constant called the *hazard rate*, the inverse of MTBF, to predict the number of failures in the field. Finally, field data is monitored to determine when wear out begins. This approach to reliability is fundamentally different from an approach based on a combination of physical testing and modeling sometimes referred to as the *physics of failure (PoF)* (see **Stress Screening**) in which lifetime is described by a deterministic model plus a stochastic term [7]. The two approaches to reliability need not be viewed as being in conflict, as long as each is applied conscientiously, but accelerated testing definitely fits into the PoF point of view.

Economic Issues

Reliability practitioners have invented many approaches to accelerated testing [8] because of the large economic impact. Each of the popular accelerated testing approaches for consumer and business products offers a trade-off between accuracy

and cost, and some approaches target different parts of the product life cycle. However, most popular accelerated practices do not apply well to industries with rugged products, such as the petroleum and avionics industries. In any case, testing can include physical testing, such as temperature cycling, and “virtual testing” using mathematical simulation.

A priori knowledge of the expected failure mechanism is required in the design of an effective accelerated life testing plan. One way to look at failure mechanisms is to assign failures to one of the two basic classes: overstress or wear out mechanisms. With the exception of highly accelerated life testing (HALT) (see **Product Risk Management: Testing and Warranties**), accelerated life testing focuses on wear out failure mechanisms. (HALT is a design procedure defined by some proponents [9] as a test-analyze-fix-test (TAFT) (see **Reliability Growth Testing**) strategy and by Munikoti and Dhar [10] and Donahoe [11] as simply an extreme accelerated test).

Although there are a finite number of failure mechanisms, the list of failure mechanisms is long. Some of these failure mechanisms include vibration-driven loosening of mechanical fasteners, damaging flexure of components due to mechanical shock, wear of items such as electrical connectors from repeated operation, wear of mechanical bearings, shorting due to voltage-driven conductive filament formation or dendrite growth in printed wiring boards (PWBs), electric arcing due to rarefied atmosphere at altitude, electrostatic discharge, solder fatigue due to temperature cycling, fretting corrosion of electrical connectors due to normal temperature cycling, loss of volatile constituents over time and temperature in lubricants and interfaces, delamination of plastic built-up parts due to swelling from high humidity, and ultraviolet damage. Therefore, the following paragraphs provide some examples of how accelerated testing by stressing improves understanding of some types of underlying failure mechanisms.

Examples of Accelerated Life Testing

Temperature Cycling – Printed Wiring Boards

Temperature cycling is a common accelerated test used for PWBs. The primary failure mechanism in PWBs is driven by the large mismatch of coefficients of thermal expansion between the PWB base material

(e.g., a glass-epoxy laminate) [12] and either the solder (typically tin–lead or tin–silver–copper solder) used to attach electrical components or ceramic components. Given the geometry of a PWB layout and the components selected, it is possible to predict the fatigue life of solder joints. Therefore, it is possible to use greater temperature excursion while cycling to accelerate the time to solder joint failure. However, solder is near its melt temperature during specified operational temperatures (and even closer to melting during accelerated testing). As a result, analysis is difficult owing to creep and nonlinear solder material properties. Furthermore, a PWB typically has hundreds to thousands of components with many differing geometries. This example of the design of an accelerated test using only temperature cycling for a PWB shows how test design and interpretation of results can be confounding. As a result, industry has produced standard test profiles.

Separable Electrical Connectors – Insertion–Removal Cycles

Estimating the lifetime of a separable electrical connector in a consumer product is another practical example of accelerated testing. In a personal computer, for example, there are a number of electrical devices that the user plugs in and removes over the lifetime. Devices include parts inside the computer chassis such as daughter boards (e.g., graphics cards or memory modules) and external parts such as Personal Computer Memory Card International Association (PCMCIA) cards (and their progeny), the printer connector, the display connector, the mouse connector, Universal Serial Bus devices (USB), etc. During lifetime, separable electrical connectors typically suffer added electrical resistance as mechanical plating wears and the electrical circuit has a threshold electrical resistance beyond which the system will stop performing. In all of these examples, the accelerated test challenge is to determine how many insertion–removal cycles are required in product life, i.e., how many insertion–removal cycles should be tested. Also, as the connector wears, there is a concern whether environmental corrosion will damage the electrical interface during its lifetime.

Mixed Flowing Gas – Accelerated Corrosion

Mixed flowing gas (MFG) is an example of an accelerated test for corrosion due to common atmospheric

gasses. The MFG test combines several gasses with humidity to generate corrosion on common metals. Although there is some dispute about the correct acceleration factor, the test is in widespread use today. Recent applications include studies to determine the robustness of materials used to replace those substances prohibited by the European Union Restriction of Hazardous Substances Regulation [13]. The development of the test is primarily attributed to W. Abbott at Battelle Labs and to IBM [14]. A number of the major standards bodies describe the method.

Limitations of Temperature Acceleration for Mature Products

Static temperature testing is especially important for products based on digital microelectronics. Semiconductor devices are temperature dependent by their nature. Continuing miniaturization of the integrated circuit along Moore’s Law [15] has created a cooling problem [16–18]. Therefore, thermal design of modern electronics pushes limits, and, as a result, accelerated testing has little thermal headroom (the difference between the design temperature and the temperature at which the component ceases to operate or is irreversibly damaged). The thermal strategy often used today is throttling, a reduction in power consumption during normal operation. However, users may select higher operational settings that challenge the design margins. As a result, creating an accelerated test by simply increasing the ambient temperature, as was common practice, will result in the product shutting itself off. In these cases, reliability testing may be forced to working closely with the part manufacturer who, without doubt, would have performed accelerated testing during design. However, the supplier (especially if that supplier enjoys a monopoly on a particular component) may be unwilling to provide information. This puts many manufacturers into a difficult spot, as they are forced to believe unsupported reliability claims.

Highly Accelerated Stress Testing (HAST)

Highly accelerated stress testing (HAST) is a common accelerated test used for electronics (see Further Reading). HAST is described in industry standards and by test equipment manufacturers and refers to

a specific test combining temperature, pressure, and moisture. In the so-called autoclave test, the environment is the same as a household pressure cooker (121 °C, 100% relative humidity, and 2 atm [19]). Since this is an extreme condition, the test designer must ensure that extraneous failure modes are not being introduced. Examples of test issues are the temperature of material phase changes, glass transition temperatures, corrosion propensity, surface moisture due to adsorption or condensation, and changes in electrical properties. HAST is routinely used for testing hermetic integrated circuit packages, popcorning (propensity for delamination and cracking of nonhermetic integrated plastic packages due to the formation of steam within plastic voids during soldering [20]), and moisture effects in many other types of devices such as multilayer ceramic capacitors [10, 11, 21].

References

- [1] Kraisch, M. (2003). Accelerated testing for demonstration of product lifetime reliability, *Annual Reliability and Maintainability Symposium 2003 Proceedings*, 27–30 January 2003, Tampa.
- [2] *SAS Proceedings of LIFEREG*, SAS Institute, Cary Version 9.1, 2003, Cary, N.C.
- [3] Nelson, W. (2004). *Accelerated Testing*, Wiley-Interscience.
- [4] Meeker, W. & Escobar, L. (1998). *Statistical Methods for Reliability Data*, Wiley-Interscience.
- [5] Misra, R. & Vyaas, B. (2003). Cost effective accelerated, *Annual Reliability and Maintainability Symposium 2003 Proceedings*, 27–30 January 2003, Tampa.
- [6] Thurman, W. (1959). Address to the fifth national symposium on reliability and quality control, *IRE Transactions Reliability and Quality Control* **8**(1), 1–6.
- [7] Pecht, M. & Dasgupta, A. (1995). Physics of failure: an approach to reliable product development, *International Integrated Reliability Workshop*, Final Report, 22–25 October 1995, pp. 1–4.
- [8] Pecht, M. (2004). *Parts Selection and Management, Using Accelerated Testing to Assess Reliability*, John Wiley & Sons, pp. 194–199.
- [9] Hobbs, G. (2002). *HALT and HASS, The New Quality and Reliability Programme*.
- [10] Munikoti, R. & Dhar, P. (1988). Highly accelerated life testing (HALT) for multilayer ceramic capacitor qualification, *IEEE Transactions Components and Packaging Technology* **11**(4), 342–345.
- [11] Donahoe, D. (2005). Moisture in multilayer ceramic capacitors, Ph.D. Dissertation, University of Maryland, Maryland.
- [12] Englemaier, W. (1983). Fatigue life of leadless chip carriers solder joints during power cycling, *IEEE Transactions on Components, Hybrids and Manufacturing Technology* **6**, 232–237.
- [13] European Commission (2006). *Restriction on the Use of Certain Hazardous Substances in Electrical and Electronic Material*, Directive 2002/95/EC.
- [14] Abbott, W. (1988). The development and performance characteristics of flowing mixed gas test environments, *IEEE Transactions on Components, Hybrids and Manufacturing Technology* **11**(1), 22–35.
- [15] Noyce, R. (1977). Microelectronics, *Scientific American* **237**(3), 63–69.
- [16] Ning, T. (2000). Silicon technology directions in the new millennium, *38th Annual Reliability Physics Symposium*, IEEE International, pp. 1–6.
- [17] Chu, R., Simons, R., Ellsworth, M., Schmidt, R. & Cozzolino, V. (2004). Review of cooling technologies for computer products, *IEEE Transactions on Device and Materials Reliability* **4**(4), 568–585.
- [18] Millman, J. (1979). *Microelectronics, Digital and Analog Circuits and Systems*, McGraw-Hill, pp. 673–676.
- [19] Lindeburg, M. (1980). *Saturated Steam, Mechanical Engineering Review Manual*, 5th Edition, Professional Engineering Review Program, San Carlos, pp. 7–37.
- [20] Gallo, A. & Munamarty, R. (1995). Popcorning: a failure mechanism in plastic-encapsulated microcircuits, *IEEE Transactions on Reliability* **44**(3), 362–367.
- [21] Donahoe, D., Pecht, M., Lloyd, I. & Ganesan, S. (2006). Moisture induced degradation of multilayer ceramic capacitors, *Journal of Microelectronics Reliability* **46**, 400–408.

Further Reading

- JEDEC Solid State Technology Association (2000). *Accelerated Moisture Resistance – Unbiased HAST, JESD22-A118*.
 JEDEC Solid State Technology Association (2000). *Accelerated Moisture Resistance – Unbiased HAST, JESD22-A102-C*.

DANIEL DONAHOE, KE ZHAO, STEVEN
MURRAY AND ROSE M. RAY

Actuarial Journals

Introduction

The actuarial profession consists of over 40 000 credentialed members working in over 100 countries throughout the world. In addition, there are likely to be many other professionals working in this field without formal certification. Despite the common name, there are actually various types of actuaries, whose work and interests can diverge significantly. One basic classification is life *versus* general actuaries. Life actuaries deal with life insurance, pensions, and health insurance issues; general (or nonlife) actuaries deal with property and liability insurance concerns. Another widely quoted classification was proposed by Hans Bühlmann [1] in an ASTIN Bulletin (AB) editorial published in 1987 in which he described three kinds of actuaries. Actuaries of the first kind emerged in the late 1600s and used deterministic methods to develop mortality tables and other calculations related to life insurance, and later to pension plans. Actuaries of the second kind appeared in the early 1900s and applied probabilistic methods to the developing lines of workers' compensation and automobile insurance, as well as to other property and liability insurance areas. Actuaries of the third kind, the ones who were the focus of Bühlmann's article, emerged in the 1980s to deal with investment issues for both insurers and other organizations, applying stochastic models to both assets and liabilities. Since the appearance of Bühlmann's article, a fourth kind of actuary has been identified, one who deals with enterprise risk management, the combination of all risks facing an organization, not just the insurance and financial risks on which actuaries had previously focused. Other classifications of actuaries are into lay (practical) and pure (theoretical), and into practicing and academic actuaries. The different types of actuaries, regardless of the classification system used, do have different interests, but share a common concern for advancing the profession through developing and sharing new actuarial techniques. To this end, a number of actuarial journals have been developed to stimulate research into actuarial problems and to share advances within the profession.

While there are several journals that have been very influential in actuarial science for many decades, the actuarial journal landscape has evolved over

time. A study by Colquitt [2] evaluates the relative influence that many of these journals have had in the field of actuarial science in recent years.

The Colquitt study evaluates the influence of 8 of the top actuarial science journals by reviewing the citations to each of these journals found in 16 of the top risk, insurance, and actuarial journals published in the years 1996–2000. The actuarial journals included in the study are the *AB*, the *British Actuarial Journal (BAJ)*, the *Casualty Actuarial Society Forum (CASF)*, *Insurance: Mathematics and Economics (IME)*, the *Journal of Actuarial Practice (JAP)*, the *North American Actuarial Journal (NAAJ)*, the *Proceedings of the Casualty Actuarial Society (PCAS)*, and the *Scandinavian Actuarial Journal (SAJ)*. The Colquitt study did not include all actuarial journals, only those cited in the other journals in his study. Some other leading actuarial journals include the *Australian Actuarial Journal*, *Belgian Actuarial Bulletin*, *South African Actuarial Journal*, *Bulletin of the Swiss Actuarial Association*, *German Actuarial Bulletin (Blätter of the German Actuarial Association)*, *French Actuarial Bulletin*, and *Italian Actuarial Journal (Giornale dell'Instituto Italiano degli Attuari)*. Actuarial articles also appear in many other journals, including the *Geneva Risk and Insurance Review*, the *Journal of Risk and Insurance*, and many other journals on probability or stochastic finance.

Actuarial Journal Descriptions

The *AB* (<http://www.casact.org/library/ASTIN/>) began publishing articles on general (nonlife) insurance in 1958. ASTIN, which stands for Actuarial Studies in Nonlife Insurance, is a section of the International Actuarial Association (IAA). In 1988, the *AB* began to accept a broader array of articles by including the Actuarial Approach for Financial Risks (AFIR) section of the IAA and incorporating financial risk issues. Recently, the editorial guidelines were revised to be open to papers on any area of actuarial practice. Articles published in the *AB* generally apply advanced mathematics and often are theoretical in nature. The editorial board is a combination of academic and practicing actuaries from a number of different countries so that it has an international perspective.

The *BAJ* (http://www.actuaries.org.uk/Display_Page.cgi?url=/maintained/resource_centre/baj.html)

has been published jointly by the Institute of Actuaries and the Faculty of Actuaries since 1995. Previously, each organization had its own journal, the *Journal of the Institute of Actuaries* and the *Transactions of the Faculty of Actuaries*. This refereed journal published papers on all areas of insurance research, life, pensions, and general insurance, with a focus on articles dealing with insurance issues applicable in Britain. The editorial board consists of Fellows of the Institute or Faculty of Actuaries, both practicing and academic. The goal of the *BAJ* is to provide the permanent record of advances in actuarial science. The journal includes, in addition to scientific papers, association studies and reports on meetings of the sponsoring societies. In 2007, the UK actuarial profession began a new journal, *Annals of Actuarial Science* (http://www.actuaries.org.uk/Display_Page.cgi?url=/maintained/resource_centre/annals.html) that publishes peer-reviewed articles on all aspects of insurance. The *BAJ* will continue to publish papers presented at meetings.

The Casualty Actuarial Society (CAS) publishes two journals. The *PCAS* (<http://www.casact.org/pubs/proceed/index.cfm?fa=pastind>) was the Society's refereed journal from 1914 through 2005. This journal tended to publish practical applications of actuarial research by members of the CAS, although membership was not a requirement for authors. Starting in 2007, the CAS changed the name of its refereed journal to *Variance* (<http://www.variancejournal.org/>), and actively encouraged nonmembers to submit articles. The goal of this journal is to foster and disseminate practical and theoretical research of interest to general actuaries worldwide. The other CAS journal is the *CASF* (<http://www.casact.org/pubs/forum/>). This journal is not refereed, and serves to encourage the exchange of current research. The complete texts of all papers that have been published in either the *PCAS* or the *CASF* are available at no charge through the CAS website.

IME (http://www.elsevier.com/wps/find/journaldescription.cws_home/505554/description#description), first published in 1982, is an international journal publishing articles in all areas of insurance. This journal seeks to bring together researchers in insurance, finance, or economics with practitioners interested in developing or applying new theoretical developments. The editorial board consists entirely of academics from a variety of disciplines related to

risk, including actuaries, economists, and mathematicians. The goal of this journal is to publish articles on the theory of insurance mathematics, economics, or innovative applications of theory.

The *JAP* (<http://www.absalompress.com/>), first published in 1993, aims to meet the needs of both practitioners and academics by publishing refereed articles on advanced topics of actuarial science in a format that strives to make the mathematics understandable to all readers. The editorial board includes both practicing and academic actuaries primarily from the United States. The journal publishes articles on all aspects of actuarial practice. The goal of the journal is to bridge the gap between the "art" and "science" of actuarial work on an international basis.

The *NAAJ* (<http://www.soa.org/news-and-publications/publications/journals/naaj/naaj-detail.aspx>) is the current refereed journal of the Society of Actuaries. This journal has been published since 1997 and seeks to attract both authors and readers beyond the membership of the Society. This journal publishes articles on a broad array of insurance issues, with a focus on life, health, pension, and investment topics. The editorial board consists primarily of members of the Society drawn from both academia and practicing actuaries. The Society's prior journals were the *Transactions of the Society of Actuaries*, published since 1949, and the journals of its predecessors, the *Record of the American Institute of Actuaries* (first published in 1909), and the *Transactions of the Actuarial Society of America*, published since 1891. The scientific articles published in these journals tended to be written by members, and to focus on practical applications of life, health, and pension actuarial issues. The Society of Actuaries also publishes a non-refereed journal, Actuarial Research Clearing House (ARCH), which seeks to facilitate the rapid sharing of advances within the actuarial field.

The *SAJ* (<http://www.ingentaconnect.com/content/routledge/sact>), published since 1918, is jointly sponsored by the Danish Society of Actuaries, the Actuarial Society of Finland, the Norwegian Society of Actuaries, and the Swedish Society of Actuaries, with editors from each society. This journal publishes theoretical and practical research using mathematical applications to insurance and other risk areas. This journal draws articles from authors internationally, and covers all areas of insurance.

Actuarial Journal Influence

First, the findings of the Colquitt study suggest that the journal most frequently cited by many of the citing journals is the citing journal itself. Given that the authors of articles published in a journal are obviously familiar with the journal in which the article is published, there is a resulting increase in self-citations. Another likely cause of a journal's increased self-citation rate is that a journal editor tends to appreciate an author's recognition of the influence that his/her journal has on a particular stream of literature.

Regarding the overall influence of the sample journals, *IME* was either the first or second most frequently cited journal by six of the eight actuarial journals. The only two actuarial journals where the *IME* was not the first or second most frequently cited journal were the *CASF* and the *PCAS*. In the cases of the *CASF* and the *PCAS*, the two most frequently cited journals were the *CASF* and the *PCAS*. Both of these journals are sponsored by the CAS and those who publish articles in these two journals are likely very familiar with and greatly influenced by the research published by this organization.

Colquitt also determines each journal's self-citation index. The higher the self-citation index, the higher a journal's frequency of self-citations relative to the frequency with which it is cited by the other sample journals. The lower the self-citation index, the more influential the journal is presumed to be. A relatively high self-citation index could indicate that a journal is inclined toward self-promotion or perhaps the journal publishes research on topics that are of a specialized nature and, as a result, is most frequently referenced by other articles within that same journal. Among the sample actuarial journals, the *NAAJ* (.40) has the lowest self-citation index, with *IME* (0.67), the *AB* (0.84), and the *SAJ* (0.92) following close behind. The remaining four actuarial journals and their self-citation indices are the *PCAS* (1.39), the *CASF* (1.51), the *BAJ* (1.52) and the *JAP* (5.15).

When looking at total citations, *IME* is the most frequently cited actuarial journal with 854, followed by the *AB* (658), the *PCAS* (626), the *SAJ* (453), and the *BAJ* (410). The remaining three actuarial journals were the *CASF* (194), the *NAAJ* (148), and the *JAP* (26). One reason for the low citation totals for the *NAAJ* and the *JAP* is likely the relative newness of these journals. In addition, the pedagogical nature of

some of the articles in the *JAP* and the relatively low number of *JAP* subscribers are also likely reasons for its low number of citations. When excluding self-citations, the only changes in the order is a switch in the first and second positions between *IME* (413) and the *AB* (481) and the switch in the sixth and seventh positions between the *CASF* (92) and the *NAAJ* (101).

While the total number of citations for the sample journals provides a measure of the total impact that each journal has on actuarial research, the total number of citations is greatly affected by the number of citable articles published by the sample journals. For example, if Journal A publishes twice as many articles as Journal B, then Journal A should receive twice as many citations, even if the research published in Journal A is not any more influential than that published in Journal B. To control for the difference in the number of articles that the journals publish, Colquitt creates an impact factor that captures the relative research impact of a journal on a per article basis. The higher the impact factor, the more influential a journal's articles are.

When evaluating the research impact of a journal on a per article basis, the *AB* is ranked first among actuarial journals with an impact factor of 2.0175. This essentially means that the *AB* articles published during the sample period were cited an average of 2.0175 times per article by the sample risk, insurance, and actuarial journals analyzed. Following the *AB* is the *PCAS* (1.9825), *IME* (1.6336), the *SAJ* (1.5656), the *BAJ* (1.3892), the *NAAJ* (1.1746), the *CASF* (0.6078), and the *JAP* (0.2766). When looking at an impact factor adjusted to exclude self-citations, there is a considerable difference in the rankings. The *AB* (1.4561) has the highest impact factor, followed by the *SAJ* (1.1475), the *PCAS* (1.1404), the *NAAJ* (0.8016), *IME* (0.7466), the *BAJ* (0.4162), the *CASF* (0.2778), and the *JAP* (0.1170).

Colquitt also determined the top 12 most frequently cited articles in any of the sample actuarial journals (there were actually 17 articles in all with five being tied for the twelfth spot). All actuarial journals except the *CASF*, the *JAP*, and the *NAAJ* are represented on this list. The *AB* and *IME* lead the list with five articles each. Close behind the *AB* and *IME* is the *PCAS* with four of the top actuarial articles and the *BAJ* and *SAJ* have two and one on the list, respectively. All but one of the articles on the list are from the 1980s and 1990s. Finally, two themes

common to several of the 17 most influential articles published in the sample actuarial journals in recent years are pricing and financial distress, which are the subject of over a third of the articles published.

Actuarial Journal Accessibility

Given the wide-ranging extent of the actuarial profession, and the variety of different specialties within the profession, it is not surprising that such an array of actuarial journals has developed to serve the profession. Advances in mathematics, the recent development of technology that allows quantitative methods that were previously impractical, and shifts in the type of risks that the profession is seeking to address, makes the ability to share advances in the field more important than ever. The actuarial journals help advance the development of actuarial science throughout the world by encouraging the dissemination of research and facilitating the exchange of published articles.

Several of these journals provide free access to all published articles online. Articles published in the *Journal of the Institute of Actuaries* or the *Transactions of the Faculty of Actuaries* are available through the association's website, although articles published in the *BAJ* are only available to members. The complete texts of all articles that have been published in *AB*, the *PCAS* or the *CASF* are available at no charge through the CAS website. Abstracts of articles published in the *JAP* are available on their website. Articles in other journals are available to subscribers of online journal services.

References

- [1] Bühlmann, H. (1987). Actuaries of the Third Kind? *Astin Bulletin* **17**, 137–138.
- [2] Colquitt, L. (2005). An Examination of the Influence of Leading Actuarial Journals, *Proceedings of the Casualty Actuarial Society*, Vol. 92, p. 1–25.

L. LEE COLQUITT AND STEPHEN P. D'ARCY

Actuary

The word *actuary* derives from the Latin word *actuarius*, who was the business manager of the Senate of Ancient Rome. It was first applied to a mathematician of an insurance company in 1775 in the Equitable Life Insurance Society of London, United Kingdom. By the middle of the nineteenth century, actuaries were active in life insurance, friendly societies, and pension schemes. As time has gone on, actuaries have also grown in importance in relation to general insurance, investment, health care, social security, and also in other financial applications in banking, corporate finance, and financial engineering.

Over the times there have been made several attempts to give a concise definition of the term *actuary*. No such attempted definition has succeeded in becoming universally accepted. As a starting point, reference is made to The International Actuarial Association (IAA)'s description of what actuaries *are*:

Actuaries are multiskilled strategic thinkers, trained in the theory and application of mathematics, statistics, economics, probability and finance

and what they *do*:

Using sophisticated analytical techniques, actuaries confidently make financial sense of the short term as well as the distant future by identifying, projecting and managing a spectrum of contingent and financial risks.

This essay will adopt a descriptive approach to what actuaries are and what they do, specifically by considering the actuarial community from the following angles:

- actuarial science, the foundation upon which actuarial practice rests;
- actuarial practice;
- some characteristics of the actuarial profession.

Actuarial Science

Actuarial science provides a structured and rigid approach to modeling and analyzing uncertain outcomes of events that may impose or imply financial losses or liabilities upon individuals or organizations. Different events with which actuarial science

is concerned – in the following called *actuarial events* – are typically described and classified according to specific actuarial practice fields, which will be discussed later.

A few examples of actuarial events are as follows:

- An individual's remaining lifetime, which is decisive for the outcome of a life insurance undertaking and for a retirement pension obligation.
- The number of fires and their associated losses within a certain period of time and within a certain geographical region, which is decisive for the profit or loss of a fire insurance portfolio.
- Investment return on a portfolio of financial assets that an insurance provider or a pension fund has invested in, which is decisive for the financial performance of the provider in question.

Given that uncertainty is a main characteristic of actuarial events, it follows that probability must be the cornerstone in the structure of actuarial science. Probability in turn rests on pure mathematics.

In order to enable probabilistic modeling of actuarial events to be a realistic and representative description of real-life phenomena, understanding of the “physical nature” of the events under consideration is a basic prerequisite. Pure mathematics and pure probability must therefore be supplemented with and supported by the sciences that deal with such “physical nature” understanding of actuarial events. Examples are death and disability modeling and modeling of financial market behavior.

It follows that actuarial science is not a self-contained scientific field. It builds on and is the synthesis of several other mathematically related scientific fields: pure mathematics, probability, mathematical statistics, computer science, economics, finance, and investments. Where these disciplines come together in a synthesis geared directly toward actuarial applications, terms like actuarial mathematics and insurance mathematics are often adopted.

To many actuaries, both in academia and in the business world, this synthesis of several other disciplines is “the jewel in the crown”, which they find particularly interesting, challenging, and rewarding.

Actuarial Practice

The main practice areas for actuaries can broadly be divided into the following three categories:

- life insurance and pensions (*see* **Life Insurance Markets; Longevity Risk and Life Annuities**)
- general/nonlife insurance (*see* **Nonlife Insurance Markets**)
- financial risk (*see* **Distributions for Loss Modeling**).

There are certain functions in which actuaries have a statutory role. Evaluation of reserves in life and general insurance and in pension funds is an actuarial process, and it is a requirement under the legislation in most countries that this evaluation is undertaken and certified by an appointed actuary. The role of an appointed actuary has long traditions in life insurance and in pension funds (*see* **Longevity Risk and Life Annuities**). A similar requirement has been introduced by an increasing number of countries since the early 1990s.

The involvement of an actuary can be required as a matter of substance (although not by legislation) in other functions. An example is the involvement of actuaries in corporations' accounting for occupational pensions. An estimate of the value of accrued pension rights is a key figure that goes into this accounting, and this requires an actuarial valuation. Although the actuary who has undertaken the valuation does not have a formal role in the general auditing process, auditors would usually require that the valuation be undertaken and reported by a qualified actuary. In this way, involvement by an actuary is almost as it was legislated.

Then there are functions where actuarial qualifications are neither a formal nor a substantial requirement, but where actuarial qualifications are perceived to be a necessity.

Outside of the domain that is restricted to actuaries, they compete with professionals with similar or tangent qualifications. Examples are statisticians, operations researchers, and financial engineers.

Life Insurance and Pensions

Assessing and controlling the risk of life insurance and pension undertakings is the origin of actuarial practice and the actuarial profession. The success in managing the risk in this area comprised the following basics:

- understanding lifetime as a stochastic phenomena, and modeling it within a probabilistic framework;
- understanding, modeling, and evaluating the diversifying effect of aggregating lifetime of several individuals into one portfolio;
- estimating individual death and survival probabilities from historic observations.

This foundation is still the basis for actuarial practice in the life insurance and pensions fields.

A starting point is that the mathematical expectation of the present value of (the stochastic) future payment streams represented by the obligation is an unbiased estimate of the (stochastic) actual value of the obligation. Equipped with an unbiased estimate and with the power of the law of large numbers, the comforting result that actual performance in a large life insurance or pension portfolio can "almost surely" be replaced by the expected performance.

By basing calculations on expected present values, life insurance and pensions actuaries can essentially do away with frequency risk in their portfolios, provided their portfolios are of a reasonable size. Everyday routine calculations of premiums and premium reserves are derived from expected present values (*see* **Premium Calculation and Insurance Pricing**).

Since risk and stochastic is not explicitly present in the formulae that life insurance and pensions actuaries develop and use for premium and premium reserve calculations, their valuation methods are sometimes referred to as deterministic. Using this notion may in fact be misleading, since it disguises the fact that the management process is based on an underlying stochastic and risk-based model.

If there was no risk other than frequency risk in a life insurance and pensions portfolio, the story could have been completed at this point. However, this would be an oversimplification, and for purposes of a real-world description life insurance and pensions actuaries need to take also other risks (*see* **Axiomatic Measures of Risk and Risk-Value Models; Basic Concepts of Insurance**) into consideration. Two prominent such risks are model risk and financial risk.

Model risk represents the problem that a description of the stochastic nature of different individuals' lifetimes may not be representative for the actual nature of today's and tomorrow's actual behavior of that phenomenon. This is indeed a very substantial risk since life insurance and pension undertakings usually involve very long durations (*see* **Estimation of Mortality Rates from Insurance Data**).

The actuarial approach to solving this problem has been to adopt a more pessimistic view of the future than one should realistically expect. Premium and premium reserves are calculated as expected present values of payment streams under the pessimistic outlook. In doing so, frequency risk under the pessimistic outlook is diversified. Corresponding premiums and premium reserves will then be systematically overstated under the realistic outlook.

By operating in two different worlds at the same time, one realistic and one pessimistic, actuaries safeguard against the risk that what one believes to be pessimistic *ex ante* will in fact turn out to be realistic *ex post*. In this way, the actuary's valuation method has been equipped with an implicit safety margin against the possibility of a future less prosperous than one should reasonably expect. This is safeguarding against a systematic risk, whereas requiring large portfolios to achieve diversification is safeguarding against random variations around the systematic trend.

As time evolves the next step is to assess how experience actually has been borne out in comparison with both the pessimistic and the realistic outlook, and to evaluate the economic results that actual experience gives rise to. If actual experience is more favorable than the pessimistic outlook, a systematic surplus will emerge over time. The very purpose of the pessimistic outlook valuation is to generate such a development.

In life insurance it is generally accepted that, when policyholders pay premiums as determined by the pessimistic world outlook, the excess premium relative to the realistic world outlook should be perceived as a deposit to provide for their own security. Accordingly, the surplus emerging from the implicit safety margins should be treated differently from ordinary shareholders' profit. The overriding principle is that when surplus has emerged, and when it is perceived to be safe to release some of it, it should be returned to the policyholders. Designing and controlling the dynamics of emerging surplus and its reversion to the policyholders is a key activity for actuaries in life insurance.

The actuarial community has adopted some technical terms for the key components that go into this process:

- pessimistic outlook: first-order basis
- realistic outlook: second-order basis

- surplus reverted to policyholders: bonus.

By analyzing historic data and projecting future trends, life insurance actuaries constantly maintain both their first-order and their second-order bases. Premium and premium reserve valuation tariffs and systems are built on the first-order basis. Over time emerging surplus is evaluated and analyzed by source, and in due course reverted as policyholders' bonus.

This dynamic and cyclic process arises from the need to protect against systematic risk for the pattern of lifetime, frequency and timing of death occurrences, etc. As mentioned above, another risk factor not dealt with under the stochastic mortality and disability models is financial risk.

The traditional actuarial approach to financial risk has been inspired by the first-order/second-order basis approach. Specifically, any explicit randomness in investment return has been disregarded in actuarial models, by fixing a first-order interest rate that is so low that it will "almost surely" be achieved. This simplified approach has its shortcomings, and we describe a more modernized approach to financial risk under the section titled "Finance".

A life insurance undertaking typically is a contractual agreement of payment of nominally fixed insurance amounts in return for nominally fixed premiums. Funds for occupational pensions, on the other hand, can be considered as prefunding vehicles for long-term pension payments linked to future salary and inflation levels. For employers' financial planning it is important to have an understanding of the impact that future salary levels and inflation have on the extent of the pension obligations, both in expected terms and in terms of the associated risk. This represents an additional dimension that actuaries need to take into consideration in the valuation of pension fund's liabilities. It also means that the first-order/second-order perspective that is crucial for life insurers is of less importance for pension funds, at least from the employer's perspective.

General Insurance

Over the years actuaries have attained a growing importance in the running of the nonlife insurance operations. The basis for the insurance industry is to accept economic risks. An insurance contract may give rise to claims. Both the number of claims

and their sizes are unknown to the company. Thus insurance involves uncertainty and here is where the actuaries have their prerogative; they are experts in insurance mathematics and statistics.

Uncertainty is perhaps even more an issue in nonlife than in the life insurance industry, mainly due to the catastrophic claims such as natural disasters. Even in large portfolios substantial fluctuations may occur.

In some countries it is required by law for the nonlife insurance companies to have an appointed actuary to approve the level of reserves and premiums and report to the supervisory authorities.

The most important working areas for nonlife actuaries in addition to statutory reporting are as follows:

- reserving
- pricing
- reviewing reinsurance program
- profit analyses
- budgeting
- product development
- produce statistics.

Reserving

It takes time from the occurrence of a claim until it is being reported to the company and even more time until the claim is finally settled. For instance, in accident insurance the claim is usually reported rather quickly to the company but it may take a long time until the size of the claim is known since it depends on the medical condition of the insured. The same goes for fire insurance where it normally takes a short time until the claim is reported but it may take a long time to rebuild the house. In liability insurance it may take a very long time from the occurrence of a claim until it is reported. One example is product liability in the pharmaceutical industry where it may take a very long time until the dangerous side effects of a medicine are revealed. The insurance company has to make allowance for future payments on claims already incurred. The actuary has a key role in the calculation of these reserves. The reserves may be split into IBNR reserves (incurred but not reported) and RBNS reserves (reported but not settled) (*see Securitization/Life*). As the names suggest the former is a provision for claims already occurred but have not yet been reported to the company. The latter is a provision for claims already

reported to the company but have not yet been finally settled, that is, future payments will occur. The actuary is responsible for the assessment of the IBNR reserves and several models and methods have been developed. The RBNS reserves are mainly fixed on an individual basis by the claims handler based on the information at hand. The actuary is, however, involved in the assessment of standard reserves, which is used when the claims handler has too scarce information to reach a reliable estimate. In some lines of business where claims are frequent and moderate in size, for instance, windscreen insurance in motor, the actuary may help in estimating a standard reserve that is used for all these claims in order to reduce administration costs. In some cases the RBNS reserves are insufficient and it is necessary to make an additional reserve. This reserve is usually called an *IBNER reserve (incurred but not enough reserved)*. However, some actuaries denote the sum of (pure) IBNR reserve and the insufficient RBNS reserve as the IBNER reserve.

Pricing

Insurance companies sell a product whose exact price is unknown to them. Actuaries can therefore help in estimating the price of the insurance product. The price will depend on several factors; the insurance conditions, the geographical location of the insured object, the age of the policyholder, etc. The actuary will estimate the impact on the price of the product from each of the factors and produce a rating table or structure. The salesmen will use the rating table to produce a price for the insurance product. The actuaries will also assist in reviewing the wording of insurance contracts in order to maintain the risks at an acceptable level. The actuaries may also assist in producing underwriting guidelines. The actuary will also monitor the overall premium level of the company in order to maintain profitability.

Reinsurance (see Reinsurance)

For most insurance companies, it will be necessary to reduce their risk by purchasing reinsurance. In this way a part of the risk is transferred to the reinsurance company. The actuary may help in assessing the necessary level of reinsurance to purchase.

Profit analyses

Analyzing the profitability of an insurance product is a complex matter involving taking into account

the premium income, investment income, payments and reserves of claims, and finally the administration costs. Thus the actuary will be a prime contributor to such analyses.

Budgeting

Actuaries may also help in developing profit and loss accounts and balance sheets for future years.

Product development

New risks or the evolvement of existing risks may require development of new products or alteration of existing products. Actuaries will assist in this work.

Statistics

To perform the above analyses, reliable and consistent data are required. Producing risk statistics is therefore an important responsibility for the actuary.

It is important for the actuary to understand the underlying risk processes and to develop relevant models and methods for the various tasks. In most lines of business the random fluctuation is substantial. This creates several challenges for the actuary. One major challenge is the catastrophic claims. Taking such claims into full consideration would distort any rating table. On the contrary, the insurance company would have to pay the large claims as well and this should be reflected in the premium.

Finance

Financial risk (*see* **Model Risk; Integration of Risk Types**) has grown to become a relatively new area for actuarial practice. Actuaries who practice in this field are called “actuaries of the third kind” within the actuarial community, perhaps also among nonactuaries.

Financial risk has always been present in all insurance and pensions undertakings, and in capital accumulating undertakings as a quite dominant risk factor. As mentioned under the section titled “Life Insurance and Pension”, the “traditional approach” to this risk has been disregarded by assumption, by stipulating future liabilities with a discount rate that was so low that it would “almost surely” be realized over time.

Financial risk is different from ordinary insurance risk in that increasing the size of a portfolio does not in itself provide any diversification effect. For

actuaries it has been a disappointing and maybe also a discouraging fact that the law of large numbers does not come into assistance in this regard.

During the last half-century or so, new perspectives on how explicit probabilistic approaches can be applied to analyze and manage financial risk have developed. The most fundamental and innovative result in this theory of financial risk/mathematical finance is probably that (under certain conditions!) risk associated with contingent financial claims can in fact be completely eliminated by appropriate portfolio management.

This theory is the cornerstone in a new practice field that has developed over the last decades and is called *financial engineering*. Activities in this field include quantitative modeling and analysis, funds management, interest rate performance measurement, asset allocation, and model-based scenario testing. Actuaries may practice financial engineering in their own right, or they may apply financial engineering as an added dimension to traditional insurance-orientated actuarial work.

A field where traditional actuarial methods and methods relating to financial risk are beautifully aligned is asset liability management (ALM). The overriding objective of ALM is to gain insight into how a certain amount of money is best allocated among given financial assets, in order to fulfill specific obligations represented by a future payment stream. The analysis of the obligation’s payment stream rests on traditional actuarial science, the analysis of the asset allocation problem falls under the umbrella of financial risks, and the blending of the two is a challenge that requires insight into both and the ability to understand and model how financial risk and insurance risk interact. Many actuaries have found this to be an interesting and rewarding area, and ALM is today a key component in the risk management of insurance providers, pension funds, and other financial institutions around the world.

A tendency in the design of life insurance products in the recent decades has been unbundling. This development, paralleled by the progress in the financial derivatives theory, has disclosed that many life insurance products have in fact option- or option-like elements built into them. Examples are interest rate guarantees, surrender options, and renewal options. Understanding these options from a financial risk perspective, pricing and managing them is an area of active actuarial research and where actuarial practice

is also making interesting progress. At the time of writing this article, meeting long-term interest rate guarantees is a major challenge with which the life and pensions insurance industry throughout the world is struggling.

A new challenge on the horizon is the requirement for insurers to prepare financial reports on a market-based principle, which the International Accounting Standards Board has had under preparation for some time. In order to build and apply models and valuation tools that are consistent with this principle, actuaries will be required to combine traditional actuarial thinking with ideas and methods from economics and finance.

With the financial services industry becoming increasingly complex, understanding and managing financial risk in general and in combination with insurance risk in particular could be expected to expand the actuarial territory in the future.

Characteristics of Profession

Actuaries are distinguished from other professionals in their qualifications and in the roles they fill in business and society. They also distinguish themselves from other professionals by belonging to an actuarial organization.

The first professional association, The Institute of Actuaries, was established in London in 1848, and by the turn of the nineteenth century, 10 national actuarial associations were in existence. Most developed countries now have an actuarial profession and an association to which they belong.

The role and the activity of the actuarial associations vary substantially from country to country. Activities that an association may or may not be involved in include

- expressing public opinion on behalf of the actuarial profession;
- providing or approving basic and/or continued education;
- setting codes of conduct;
- developing and monitoring standards of practice;
- involvement in or support to actuarial research.

There are also regional groupings of actuarial association, including the grouping of associations from the European Union (EU) European Economic Area (EEA) member countries, the Southeast Asian grouping, and the associations from Canada, Mexico, and United States.

The IAA, founded in 1895, is the worldwide confederation of professional actuarial associations (also with some individual members from countries that do not have a national actuarial association). It is IAA's ambition to represent the actuarial profession globally, and to enhance the recognition of the profession and of actuarial ideas. IAA focuses on professionalism, actuarial practice, education and continued professional development, and the interface with governments and international agencies. Actuarial professions, which are full members of IAA, are required to have in place a code of conduct, a formal disciplinary process and a due process for adopting standards of practice, and to comply with the IAA educational syllabus guidelines.

Related Articles

Basic Concepts of Insurance

Correlated Risk

Hazards Insurance: A Brief History

ARNE EYLAND AND PÅL LILLEVOLD

Adverse Selection

If an insurer sets a premium based on the average probability of a loss in an entire population, those at higher-than-average risk for a certain hazard will benefit most from coverage, and hence will be the most likely to purchase insurance for that hazard. In an extreme case, the poor risks will be the only purchasers of coverage, and the insurer can expect to lose money on each policy sold. This situation, referred to as *adverse selection*, occurs when the insurer cannot distinguish between members of good- and poor-risk categories in setting premiums. This article discusses the implications of adverse selection for insurance.

An Example of Adverse Selection

The assumption underlying adverse selection is that purchasers of insurance have an informational advantage over providers because they know their own true *risk types*. Insurers, on the other hand, must collect information to distinguish between risks.

Example: Private Information about Risk Types Creates Inefficiencies

Suppose some homes have a 10% probability of suffering damage (the “good” risks) and others have a 30% probability (the “poor” risks). If the loss in the event of damage is \$100 for both groups and if there are an equal number of potentially insurable individuals in each risk class, then the expected loss for a random individual in the population is $0.5 \times (0.1 \times \$00) + 0.5 \times (0.3 \times \$100) = \$20$. If the insurer charges an actuarially fair premium across the entire population, then only the poor-risk class would normally purchase coverage, since their expected loss is \$30 ($= 0.3 \times \100), and they would be pleased to pay only \$20 for the insurance. The good risks have an expected loss of \$10 ($= 0.1 \times \100), so they probably would not pay \$20 for coverage. If only the poor risks purchase coverage, the insurer will suffer an expected loss of $-\$10$, $(\$20 - \$30)$, on each policy it sells.

Managing Adverse Selection

There are two main ways for insurers to deal with adverse selection. If the company knows the probabilities associated with good and bad risks, it can raise the premium to at least \$30 so that it will not lose money on any individual. This is likely to produce a partial market failure, as many individuals who might want to purchase coverage will not do so at this high rate. Alternatively, the insurer can design and offer a “separating contract” [1]. More specifically, it can offer two different price-coverage contracts that induce different risk “types” to separate themselves in their insurance-purchasing decisions. For example, contract 1 could offer price = \$30 and coverage = \$100, while contract 2 might offer price = \$10 and coverage = \$40. If the poor risks preferred contract 1 over contract 2, and the good risks preferred contract 2 over contract 1, then the insurer could offer coverage to both groups while still breaking even. A third approach is for the insurer to collect information to reduce uncertainty about true risks, but this may be expensive.

Conclusion

In summary, the problem of adverse selection only emerges if the persons considering the purchase of insurance have more accurate private information on the probability of a loss than do the firms selling coverage. If the policyholders have no better data than the insurers, coverage will be offered at a single premium based on the average risk, and both good and poor risks will want to purchase policies.

Reference

- [1] Rothschild, M. & Stiglitz, J. (1976). Equilibrium in competitive insurance markets: the economics of markets with imperfect information, *Quarterly Journal of Economics* **90**, 629–650.

HOWARD KUNREUTHER

Air Pollution Risk

Concern about air pollution risk takes two main forms. The first is the greenhouse effect – the collective contribution of a group of gases (known as *greenhouse gases*), which results in global warming and has potentially catastrophic consequences for our climate. The best-known greenhouse gas, and the one on which most emission-reduction attempts are focused, is carbon dioxide (CO_2). However, since this encyclopedia contains a separate entry on *global warming* (see **Global Warming**), we shall not consider it any further here. The second major risk, and the focus of this article, is the effect of air pollution on human health.

As an illustration of one of the major recent studies of this phenomenon, Figure 1 (taken from [1]) shows the results of a time series study based on 88 US cities. For each city, there is a plot of the regression coefficient and 95% confidence interval for the estimated percentage increase in mortality corresponding to a $10\mu\text{g m}^{-3}$ rise in particulate matter of aerodynamic diameter less than $10\mu\text{m}$ (PM_{10}), a size at which particles are capable of penetrating directly into the lungs. Other studies have focused on $\text{PM}_{2.5}$, which has the same definition with a maximum diameter of $2.5\mu\text{m}$.

The cities are grouped into seven regions, and the figure also shows a posterior mean and 95% posterior interval of the pooled effect across each region. Finally, at the right-hand side of the figure the estimated national effect is shown: this shows a posterior mean increased mortality of 0.21% with a posterior standard deviation of 0.06%. Other results from the so-called National Morbidity and Mortality Air Pollution Study (NMMAPS) have included a similar study of ozone [2] and the effect of $\text{PM}_{2.5}$ on hospital admissions [3]. These and other results have been extensively cited in recent years as evidence for tightening the U.S. air pollution standards.

The remainder of this article covers the background and history of this subject, followed by a detailed description of time series studies. Other study designs are also covered, followed by some of the caveats that have been expressed about this whole area of research.

Background and History

The first studies of the impact of air pollution on human health were done in the 1950s, as a result of several dramatic incidents of extremely high air pollution causing widespread death. Possibly the best known of these incidents was the London “smog” of December 5–8, 1952, during which the level of “British smoke” rose to over $3000\mu\text{g m}^{-3}$ and the result was around 4000 excess deaths over what would normally have been expected during this period. Similar incidents in other places led to global concern about the consequences of high air pollution, and motivated the introduction of legislation such as the (British) Clean Air Act of 1956 and the (US) Clean Air Act of 1970, which were the first attempts to deal with the issue by regulation.

Despite the success of these early attempts at eliminating very high pollution events, concern persisted that even at much lower levels pollution was still responsible for adverse health effects, including premature death. Analysis of long-term data records from London ([4, 5], amongst others) prompted researchers to start compiling and analyzing time series from several U.S. cities (e.g. [6–9]). Most of these showed that, after adjusting for effects due to seasonal variation and meteorology, a strong correlation remained between PM and mortality. Other studies showed similar associations with various measures of morbidity, for example, hospital admissions or asthma attacks among children. However, some authors focused on the sensitivity of these results to modeling assumptions and suggested they were not statistically reliable [10, 11].

This background led to a number of large-scale research efforts, the best known of which is NMMAPS. In the next section, we outline the methodology behind these studies.

Time Series Analyses

Although there are many variants on the basic methodology, most of these are close to the following method.

The analyses depend on multiple regressions in which the dependent variable y_t , $t = 1, \dots, n$ is either the death count or some measure of morbidity (e.g., hospital admissions) on day t . Typically the death counts exclude accidental deaths and they may

an explanation were true, there should be observed negative correlations to account for the temporary decrease in the population of susceptible individuals. Studies have repeatedly failed to demonstrate such correlations [12, 13].) Sometimes pollutants other than the main one of interest are included as possible “copollutants”, e.g., in a study of PM_{10} , we may include SO_2 as a copollutant to adjust for the possible confounding of those two effects.

For (b), temperature is almost always included, as well as at least one variable representing humidity, and there may be lagged values as well. The NMMAPS papers have used temperature and dewpoint as the two variables of interest, both of current day and the average of the three previous days to accommodate lagged effects. Other authors have used either specific or relative humidity instead of dewpoint, and some have also included atmospheric pressure.

For (c), it is conventional to assume that one component of the regression function is some nonlinear function of time that has sufficient degrees of freedom to incorporate both seasonal and long-term trend effects. The nonlinear effect may be modeled as a linear sum over K spline basis functions [14]; here K is the number of “knots” and is the most critical parameter. Typically authors use between 4 and 12 knots/year. Similar representations are sometimes used to treat other variables, such as temperature and dewpoint, nonlinearly, though typically with much smaller K (in the range 3–6).

In addition to the above covariates, the NMMAPS analyses have typically included a day-of-week effect and additional nonlinear terms to represent the interaction of long-term trend with age-group.

The alternative “generalized additive model” or (GAM) approach [15] has also been used for nonlinear effects. Some erroneous results were reported owing to inappropriate use of default convergence criteria and standard error formulae [16], though subsequent research resolved these difficulties and strengthened the methodology [17].

Combining Estimates across Cities

Although the initial application of time series regression analysis was to one city at a time, it has been generally recognized that, to obtain definitive results, it is necessary to combine analyses across many

cities. A typical assumption is that the critical parameter of interest (for example, the regression coefficient relating mortality to PM_{10}) is a random effect for each city, say, θ_c in city c , drawn independently from a normal distribution with mean θ^* and variance τ^2 . However, the *estimate* in city c , denoted $\hat{\theta}_c$, is also treated as random with mean θ_c with a presumed known standard error. On the basis of these assumptions, we could, for example, estimate the national parameters θ^* and τ^2 by restricted maximum likelihood, followed by smoothed (or “shrinkage”) estimates of the individual θ_c ’s. Alternatively, researchers have taken a Bayesian approach (see **Bayesian Statistics in Quantitative Risk Assessment**) to the whole analysis, for example using the TLNISE software of Everson and Morris [18]. Some attempts have been made to extend the basic random-effects model to allow for spatially dependent effects [19].

The results of Figure 1 result from application of this methodology to data on 88 US cities from 1987 to 2000. The air pollution variable was daily PM_{10} , lagged 1 day. Other covariates at each city include long-term trend, temperature and dewpoint (current day plus average of the three previous days, using splines to allow for a nonlinear effect), day of week, and an interaction term between the long-term trend and age-group. Most of the attention has been focused on the regional and national “overall” results, where point and interval estimates are given for θ^* .

Alternative Study Designs

Prospective Studies

Apart from time series analysis, there are two other commonly used study designs. *Prospective studies* take a specific cohort of individuals and follow them through a long time period (see **Cohort Studies**). This has the advantage of allowing researchers to measure long-term effects, which is not possible in time series studies. However, unlike time series studies in which regression parameters are computed for each city, and only later combined across cities to achieve greater precision, in prospective studies the regressions themselves rely on between-city comparisons, typically estimating a standardized mortality rate for each city and regressing on some citywide measure of air pollution. This raises issues associated with ecological bias, or in other words, the possibility

that between-city variations may be owing to effects that have nothing to do with air pollution.

Reference [20] presents results from the Harvard Six Cities study, a long-term study of over 8000 individuals in six U.S. cities. Survival rates were conducted using the Cox regression model and showed that, after adjusting for smoking and other known risk factors, there was a statistically significant association between air pollution and mortality. A subsequent paper [21] showed similar results based on a much larger study (the American Cancer Society or (ACS) study), which involved over 500 000 individuals from 154 U.S. cities. Although the study involved many more participants, in other respects it was inferior to the Six Cities study, for example in using participants recruited by volunteers rather than a randomized sample, and in relying essentially on air pollution measures at a single point in time. A third study is the Adventist Health Study of Smog (AHSMOG), which showed similar results for a cohort of over 6000 non-smoking California Seventh-day Adventists [22].

Given the importance of these studies for regulation, the Health Effects Institute commissioned an independent reanalysis of the Six Cities and ACS studies [23]. This study recreated the datasets and largely confirmed the correctness of the original analyses. However, they also conducted many sensitivity analyses, some of which raised doubts about the interpretation of results. We refer to these in more detail in the section titled “Issues and Controversies”.

Case-Crossover Studies

A third paradigm for the design of air pollution–mortality studies is the *case-crossover design*. The idea is to compare the exposure of an individual to a pollutant immediately prior to some catastrophic event (e.g., death and heart attack) with the exposure of the same individual to the same pollutant at other control or “referent” times. Making plausible assumptions about how the risk of the catastrophic event depends both on time and covariates, it is possible to write down likelihood estimating equations (for a regression coefficient between the pollutant and the risk of the catastrophic event) that look very similar to the Poisson-regression equations that arise in time series studies. However, a source of bias is the time interval between the catastrophic event and the selected referent times: if it is too long the analysis may be biased owing to trend, and if it is too

short it could be affected by autocorrelation. References [24, 25] used (respectively) simulation and theoretical arguments to examine the bias issue. The case-crossover methodology was applied [26] to out-of-hospital sudden cardiac arrest in the Seattle region, finding no significant relationship between high air pollution and mortality, which the authors attributed to the lack of prior history of coronary artery disease in the subjects under study, in contrast with other studies that have included patients with such history.

Issues and Controversies

Despite the enormous amount of research that has been done on air pollution and health, the scientific community is by no means unanimous about the interpretation of these studies, especially in the context of regulations about air quality standards. Extended commentaries have been provided [27, 28]; here we summarize a few of the issues that have been raised.

None of the study designs we have discussed are controlled, randomized studies of the sort that are common in, for instance, drug testing. Therefore, they are all vulnerable to possible confounders or “effect modifiers”. Despite serious efforts to include such effects as covariates in the regression analyses, the results typically remain sensitive to exactly which covariates are included or certain *ad hoc* decisions about how to include them (for example, when long-term trends are modeled nonlinearly using splines, how many degrees of freedom to include in the spline representation). See [11, 29] for issues related to model selection or model averaging; the recent paper [30] contains a particularly comprehensive discussion of the degrees of freedom issue.

Most studies have assumed a linear relationship between dose and response (possibly after transformation, e.g., $\log \mu_t$ in the case of Poisson-regression time series analysis). But this is arguably inappropriate for regulatory decisions in which it is critical to assess the likely benefit of a specific reduction in pollution (for example, if the 8-h ozone standard were reduced from 80 to 60 parts per billion). Bell *et al.* [31] presented nonlinear models for ozone; the earlier authors did the same for PM [32–34] with varying conclusions.

The question of whether to focus on fine particles or coarse particles has been the cause of much debate.

Much of the research and regulatory effort over the past decade has been focused on fine particles ($PM_{2.5}$), which penetrate deeper into the lungs and are therefore widely believed to have a more significant health effect. However, to consider one example, Smith *et al.* [34] reached the opposite conclusion while analyzing epidemiological data from Phoenix, Arizona.

The criticisms that have been raised regarding cohort studies are somewhat different, but ultimately the basic issue is whether the associations found in studies are indicative of a true causal effect. Krewski *et al.* [23] introduced a number of “ecological covariates” on a citywide scale to determine whether the intercity PM effects that had been observed in earlier studies could be due to other sources. In the case of the ACS dataset, they examined some 20 possible ecological covariates; all but two were not statistically significant, but one of those that was significant was gaseous sulfur dioxide (SO_2). The picture was further clouded when spatial correlations were introduced into the model; in one analysis, involving both SO_2 and sulfate particles in a model with spatial dependence, the “particles” effect was not statistically significant, though the SO_2 effect still was significant. It has been speculated [35] that these inconsistencies in the results of different cohort studies may be due to an inappropriate assumption of proportional hazards in the Cox regression model.

Summary and Conclusions

The Environmental Protection Agency has recently finalized a new $PM_{2.5}$ standard – controversially from the point of view of some epidemiologists, it did not lower the long-term average level permitted from the standard of $15 \mu g m^{-3}$ that was introduced in 1997. A possible lowering of the ozone standard, from its present value of 80 parts per billion, is still under consideration. Other countries have similar standards in force that in some cases are lower than in the United States. Both advocates and opponents of tightened standards draw heavily on the epidemiological studies that have been discussed in this article, so their interpretation has significant political and societal implications. In the view of this author, new research over the past decade has added enormously to the information available about health effects, but there remain fundamental controversies that may never be fully resolved.

References

- [1] Dominici, F., McDermott, A., Daniels, M., Zeger, S.L. & Samet, J.M. (2003). Mortality among residents of 90 cities, *Revised Analyses of the National Morbidity, Mortality and Air Pollution Study, Part II*, Health Effects Organization, Cambridge, pp. 9–24.
- [2] Bell, M.L., McDermott, A., Zeger, S.L., Samet, J.M. & Dominici, F. (2004). Ozone and short-term mortality in 95 US urban communities, 1987–2000, *The Journal of the American Medical Association* **292**, 2372–2378.
- [3] Dominici, F., Peng, R., Bell, M., Pham, L., McDermott, A., Zeger, S. & Samet, J. (2006). Fine particles air pollution and hospital admission for cardiovascular and respiratory diseases, *The Journal of the American Medical Association* **295**, 1127–1135.
- [4] Mazumdar, S., Schimmel, H. & Higgins, I.T.T. (1982). Relation of daily mortality to air pollution: an analysis of 14 London winters, 1958/59–1971/72, *Archives of Environmental Health* **37**, 213–220.
- [5] Schwartz, J. & Marcus, A. (1990). Mortality and air pollution in London: a time series analysis, *American Journal of Epidemiology* **131**, 185–194.
- [6] Schwartz, J. & Dockery, D.W. (1992). Increased mortality in Philadelphia associated with daily air pollution concentrations, *The American Review of Respiratory Disease* **145**, 600–604.
- [7] Schwartz, J. & Dockery, D.W. (1992). Particulate air pollution and daily mortality in Steubenville, Ohio, *American Journal of Epidemiology* **135**, 12–19.
- [8] Pope, C.A., Schwartz, J. & Ransom, M. (1992). Daily mortality and PM_{10} pollution in Utah Valley, *Archives of Environmental Health* **42**, 211–217.
- [9] Schwartz, J. (1993). Air pollution and daily mortality in Birmingham, Alabama, *American Journal of Epidemiology* **137**, 1136–1147.
- [10] Styer, P., McMillan, N., Gao, F., Davis, J. & Sacks, J. (1995). The effect of outdoor airborne particulate matter on daily death counts, *Environmental Health Perspectives* **103**, 490–497.
- [11] Smith, R.L., Davis, J.M., Sacks, J., Speckman, P. & Styer, P. (2000). Regression models for air pollution and daily mortality: analysis of data from Birmingham, Alabama, *Environmetrics* **11**, 719–743.
- [12] Dominici, F., McDermott, A., Zeger, S.L. & Samet, J.M. (2003). Airborne particulate matter and mortality: time-scale effects in four US cities, *American Journal of Epidemiology* **157**, 1055–1065 (reply to commentary: pp. 1071–1073).
- [13] Smith, R.L. (2003). Commentary on Dominici *et al.* (2003). Airborne particulate matter and mortality: time-scale effects in four US cities, *American Journal of Epidemiology* **157**, 1066–1070.
- [14] Green, P.J. & Silverman, B.J. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall, London.

- [15] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*, Chapman & Hall, London.
- [16] Dominici, F., McDermott, A., Zeger, S.L. & Samet, J.M. (2002). On the use of generalized additive models in time series of air pollution and health, *American Journal of Epidemiology* **156**, 193–203.
- [17] Dominici, F., McDermott, A. & Hastie, T. (2004). Improved semi-parametric time series models of air pollution and mortality, *Journal of American Statistical Association* **468**, 938–948.
- [18] Everson, P.J. & Morris, C.N. (2000). Inference for multivariate normal hierarchical models, *Journal of the Royal Statistical Society, Series B* **62**, 399–412.
- [19] Dominici, F., McDermott, A., Zeger, S.L. & Samet, J.M. (2003). National maps of the effects of PM on mortality: exploring geographical variation, *Environmental Health Perspectives* **111**, 39–43.
- [20] Dockery, D.W., Pope, C.A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G. & Speizer, F.E. (1993). An association between air pollution and mortality in six U.S. cities, *The New England Journal of Medicine* **329**, 1753–1759.
- [21] Pope, C.A., Thun, M.J., Namboodiri, M.M., Dockery, D.W., Evans, J.S., Speizer, F.E. & Heath, C.W. (1995). Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults, *American Journal of Respiratory and Critical Care Medicine* **151**, 669–674.
- [22] Abbey, D., Nishino, N., McDonnell, W.F., Burchette, R.J., Knutsen, S.F., Beeson, W.L. & Yang, J.L. (1999). Long-term inhalable particles and other air pollutants related to mortality in nonsmokers, *American Journal of Respiratory and Critical Care Medicine* **159**, 373–382.
- [23] Krewski, D., Burnett, R.T., Goldberg, M.S., Hoover, K., Siemiatycki, J., Jerrett, M., Abrahamowicz, M. & White, W.H. (2000). *Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality*, A Special Report of the Institute's Particulate Epidemiology Reanalysis Project, Health Effects Institute, Cambridge.
- [24] Levy, D., Lumley, T., Sheppard, L., Kaufman, J. & Checkoway, H. (2001). Referent selection in case-crossover analyses of health effects of air pollution, *Epidemiology* **12**, 186–192.
- [25] Janes, H., Sheppard, L. & Lumley, T. (2005). Overlap bias in the case-crossover design, with application to air pollution exposures, *Statistical Medicine* **24**, 285–300.
- [26] Levy, D., Sheppard, L., Checkoway, H., Kaufman, J., Lumley, T., Koenig, J. & Siscovick, D. (2001). A case-crossover analysis of particulate matter air pollution and out-of-hospital primary cardiac arrest, *Epidemiology* **12**, 193–199.
- [27] Smith, R.L., Guttorp, P., Sheppard, L., Lumley, T. & Ishikawa, N. (2001). Comments on the criteria document for particulate matter air pollution, NRCSE Technical Report Series 66, available from <http://www.nrcse.washington.edu/research/reports.html>.
- [28] Moolgavkar, S.H. (2005). A review and critique of the EPA's rationale for a fine particle standard, *Regulatory Toxicology and Pharmacology* **42**, 123–144.
- [29] Clyde, M. (2000). Model uncertainty and health effect studies for particulate matter, *Environmetrics* **11**, 745–763.
- [30] Peng, R., Dominici, F. & Louis, T. (2006). Model choice in multi-site time series studies of air pollution and mortality (with discussion), *Journal of the Royal Statistical Society Series A* **169**, 179–203.
- [31] Bell, M., Peng, R. & Dominici, F. (2006). The exposure-response curve for ozone and risk of mortality and the adequacy of current ozone regulations, *Environmental Health Perspectives* **114**, 532–536.
- [32] Daniels, M.J., Dominici, F., Samet, J.M. & Zeger, S.L. (2000). Estimating particulate matter-mortality dose-response curves and threshold levels: an analysis of daily time series for the 20 largest US cities, *American Journal of Epidemiology* **152**, 397–406.
- [33] Schwartz, J. & Zanobetti, A. (2000). Using meta-smoothing to estimate dose-response trends across multiple studies, with application to air pollution and daily death, *Epidemiology* **11**, 666–672.
- [34] Smith, R.L., Spitzner, D., Kim, Y. & Fuentes, M. (2000). Threshold dependence of mortality effects for fine and coarse particles in Phoenix, Arizona, *Journal of the Air and Waste Management Association* **50**, 1367–1379.
- [35] Moolgavkar, S.H. (2006). Fine particles and mortality, *Inhalation Toxicology* **18**, 93–94.

Related Articles

Environmental Health Risk

Environmental Monitoring

Environmental Performance Index

RICHARD L. SMITH

Alternative Risk Transfer

Active management of risk is an important element of any corporate strategy designed to increase enterprise value (the discounted value of future after-tax profits), and traditional loss financing mechanisms such as full insurance/reinsurance contracts (*see* **Reinsurance**), are widely used in support of this goal. These conventional techniques have been supplemented since the mid- to late 1990s by solutions from the alternative risk transfer (ART) market, which we define as the combined risk-management marketplace for innovative loss financing programs. The ART market comprises various products, vehicles, and solutions, including the following:

- Partial insurance and finite risk contracts, which finance, rather than transfer, risk exposures.
- Multirisk contracts, which transfer multiple risks simultaneously.
- Insurance-linked securities, which securitize, and then transfer, insurable risks to capital markets investors.
- Contingent capital, which provides a company with securities or bank financing in the aftermath of an insurable loss event, at a cost determined *ex ante*.
- Insurance derivatives, which allow hedging of insurable risk through a direct transfer to capital markets institutions.
- Captives, which facilitate and economize a company's self-directed risk financing or risk transfer strategies.
- Enterprise risk-management programs, which combine disparate risks, time horizons, and instruments into a single, multiyear risk plan.

The ART marketplace is considered “alternative” because it expands conventional risk-management horizons, allowing companies to consolidate risks, alter cash flows and time horizons, utilize a full range of “hybrid” insurance and banking mechanisms, and attract capital from the insurance sector and global capital markets at the lowest possible cost.

Partial Insurance and Finite Risk Contracts

Full insurance can be considered a maximum risk transfer contract where the ceding company shifts

to an insurer as much exposure as possible at an economically fair premium. Full insurance is characterized by small deductibles, large policy caps, limited coinsurance, and narrow exclusions. Partial insurance contracts, in contrast, involve a far greater amount of retention. By increasing the deductible, lowering the policy cap, and increasing coinsurance and exclusions the ceding company retains more risk and pays the insurer a lower premium (*see* **Premium Calculation and Insurance Pricing**). This is equivalent to partial financing rather than transfer, and a simple cost/benefit exercise will reveal whether the strategy actually serves to increase enterprise value.

Finite risk contracts, such as loss portfolio transfers and adverse development cover, are partial insurance mechanisms that are commonly used to manage the risks associated with loss exposures or the rate of loss accrual. They serve primarily as cash flow timing, rather than loss transfer, mechanisms, offering balance sheet, and cash flow protection rather than capital protection.

Consider the example of a company that wishes to create more stable cash flows while transferring a small amount of property and casualty (P&C) risk exposure. The company enters into a 3-year finite policy where it pays \$2 million of premium/year into an experience account that earns 5% interest (taxable at a marginal corporate rate of 34%). The company favors this arrangement because the \$2 million reflects a certain, predictable cash outflow. In order to establish and maintain the program, the company pays an insurer an annual fee equal to 10% of premium. Any P&C losses that occur over the 3-year period are funded through the experience account and any shortfall is split between the company and insurer on a 90%/10% basis. Table 1 reflects assumed loss experience of \$1, \$2, and \$5 million and the resulting cash flows under the finite contract.

The account ends with a deficit of nearly \$2.4 million at the end of year 3, meaning the company will fund approximately \$2.2 million while the insurer will cover the balance. In fact, the company may arrange to fund the shortfall in additional installments over another 3-year period in order to continue smoothing its cash flows, which is its primary goal in creating the finite program.

Table 1 Finite contract cash flows^(a)

	Year 1	Year 2	Year 3
Previous balance	\$0	\$860 000	\$747 000
Premium deposit	\$2 000 000	\$2 000 000	\$2 000 000
Fee	–\$200 000	–\$200 000	–\$200 000
Beginning balance	\$1 800 000	\$2 660 000	\$2 547 000
Claims (loss experience)	–\$1 000 000	–\$2 000 000	–\$5 000 000
After-tax interest	\$60 000	\$87 000	\$85 000
Ending balance	\$860 000	\$747 000	–\$2 368 000

^(a) Reproduced from [1] with permission from John Wiley & Sons, Inc., © 2004

Multirisk Contracts

A multirisk contract (*see* **Risk Measures and Economic Capital for (Re)insurers**), such as the multiple peril or multiple trigger products, combines several exposures or events into a single risk-management instrument, giving a company an efficient, and often cost-effective, risk solution. Since multirisk contracts are based on multiple risks/events, the effects of correlation, and joint probabilities can lead to a more favorable cost of protection, which can strengthen enterprise value.

Multiple peril products, which include commercial umbrella policies and multiline policies, act as risk consolidation programs, combining designated insurable risks into a single, 3- to 7-year policy with an aggregate premium, deductible, and cap. Since coverage is conveyed through a consolidated document, transaction costs and premiums decline, while the risk of overinsurance is reduced. In addition, a company with a multiple peril policy need not be concerned about the specific source of a loss; as long as a loss-inducing peril is named in the policy, indemnification occurs. The scope of insurable risk coverage has broadened in recent years; while the earliest versions of the contract focused on similar risks, disparate risks are now often included (e.g., workers' compensation, environmental liability, and P&C business interruption).

Multiple trigger products, a second type of multirisk contract, provide a company with loss financing only if two or more risk events occur. For instance, a power company supplying electricity to industrial and retail customers might be severely impacted if its generators suffer from mechanical failure (first trigger event) and the price of electricity rises at the same time (second trigger event); under this scenario the higher the power price during interruption, the greater

the loss. If the company is primarily concerned about the financial impact of the joint event, which may be regarded as catastrophic in nature (i.e., high severity/low probability), it can combine the two in a multiple trigger contract and achieve considerable cost savings (*see* **Dependent Insurance Risks; Risk Measures and Economic Capital for (Re)insurers**). For instance, if each of the two events has a 10% likelihood of occurrence, the joint event has only a 1% probability of occurrence, which leads to a reduction in the cost of protection.

Insurance-Linked Securities

Securitization which is the process of removing assets, liabilities, or cash flows from the corporate balance sheet and conveying them to third parties through tradable securities, has been a feature of the financial markets for several decades. In the 1990s, banks began applying securitization techniques, which they had used to good effect in securitizing mortgages, receivables, and credits to the insurance market, creating notes and bonds based on insurance-related events. Most insurance-linked securities issuance is arranged on behalf of insurers and reinsurers that are seeking to transfer risk exposure or create additional risk capacity within their catastrophic P&C portfolios (e.g., hurricane, windstorm, and earthquake). The fundamental mechanism provides access to the global capital markets, which are significantly larger than the insurance/reinsurance markets. This can provide important cost savings, particularly during hard market cycles in insurance/reinsurance.

Under a typical insurance-linked securitization (Figure 1) a special purpose reinsurer issues securities to investors and writes a matching reinsurance

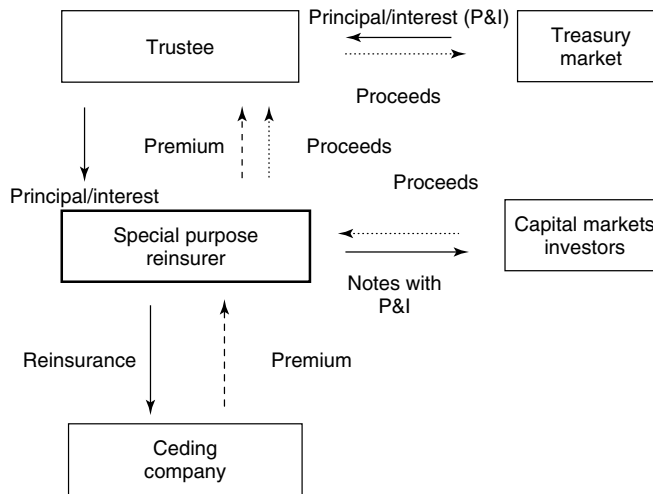


Figure 1 Insurance securitization [Reproduced from [1] with permission from John Wiley & Sons, Inc., © 2004.]

(or retrocession) contract to the insurer (reinsurer). Proceeds from the issue are invested in government securities by the trustee, and the yield on the government collateral together with the premium on the reinsurance combine to form the interest coupon. Actual payment of interest and/or principal is based on losses arising from a defined trigger event: if losses exceed a predetermined threshold as a result of a named event (e.g., earthquake or hurricane), the special purpose reinsurer withholds interest and/or principal payments from investors; if no event occurs, investors receive full interest and principal cash flows. The insurer or reinsurer thus protects itself from losses created by the defined event. Securities can be issued with an indemnity trigger (where cash flows are suspended if the insurer's actual book of business loses a certain amount), a parametric trigger (where cash flows are suspended if a location/severity damage metric exceeds a particular value), or an index trigger (where cash flows are suspended if a loss index exceeds a particular value).

Contingent Capital

Contingent capital is a contractually agreed financing facility that provides a company with funding in the aftermath of a defined insurable loss event. Since the facility is arranged in advance of any loss leading to financial distress, the company's financing cost does not reflect the risk premium that might otherwise be demanded by capital suppliers after a loss. For

instance, if a company that can normally raise 12-month funds at 50 basis points over government rates suffers a very large loss as a result of an insurable event that impairs its financial condition, it may be forced to pay +250 basis points to raise funds, reducing its enterprise value in the process. The contingent capital facility eliminates this incremental financing burden.

In a generic form of the structure (Figure 2) a company identifies an amount of capital that it wishes to raise if a loss occurs, determines the event that can trigger the loss, and defines the specific form of financing it wishes to raise (e.g., debt securities, bank loan, and equity securities). If the event occurs, the capital provider supplies funds at the *ex ante* price. In return, the company pays the capital provider a periodic (or up front), nonrefundable commitment fee (payable whether or not the financing is raised) as well as an underwriting/financing fee (payable only if financing is raised).

Consider the following example. An insurance company arranges a \$500 million 5-year note issue that will be triggered in the event losses in its P&C portfolio exceed \$500 million over the next 2 years. The arranging bank identifies investors that "prefund" a \$500 million trust in exchange for an all-in yield equal to the commitment fee plus the return on 5-year government bond. The \$500 million in prefunding proceeds are used to purchase the bonds. Assume that 1 year from now the insurer's actual P&C loss experience is greater than expected. The

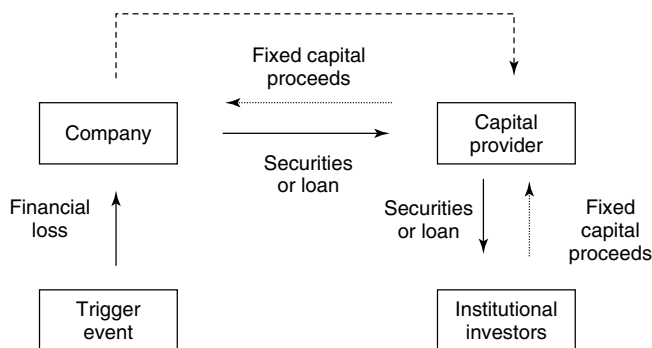


Figure 2 Contingent financing [Reproduced from [1] with permission from John Wiley & Sons, Inc., © 2004.]

insurance company issues \$500 million of 5-year notes to the trust, the trust liquidates its government bonds and uses the proceeds to acquire the notes; the trust now holds the insurance company's notes, the insurer receives \$500 million of cash to help manage its financial position and end investors continue to receive the enhanced yield on the trust-issued notes.

Insurance Derivatives

Derivatives, which can be broadly defined as financial contracts that derive their value from a market reference, permit users to transfer the economics of specific risk references such as equity indexes and interest rates (*see Stochastic Control for Insurance Companies*). The contracts have been actively used by hedgers and speculators for decades to exchange or assume risks of various financial and commodity indexes and can be structured in exchange-traded (listed and standardized) or over-the-counter (bespoke) form. Derivatives have also been applied selectively to insurance risks since the 1990s. The most common, in what is still a relatively small (though growing) market, include catastrophe swaps and noncatastrophic weather contracts. As with other instruments, the intent is to manage risk exposures as efficiently as possible.

An insurer can manage its risk portfolio using the catastrophe reinsurance swap, a synthetic over-the-counter transaction where it exchanges a commitment fee for a contingent payment from its counterparty based on a catastrophic loss. By doing so, the insurer obtains many of the same benefits provided by reinsurance or securitization (e.g., portfolio diversification, exposure reduction, and increased capacity), but without the structural complexities and costs. For

instance, an insurer might pay a reinsurer a funding rate plus a spread over a multiyear period in exchange for \$100 million of contingent catastrophe exposure capacity. If the catastrophe named under the swap occurs and creates a loss, the counterparty provides the ceding insurer with compensation of up to \$100 million and assumes claim rights through subrogation. An insurer can also alter its portfolio through the pure catastrophe swap, a synthetic transaction that provides for the exchange of uncorrelated catastrophe exposures. For instance, a Japanese insurer with excess Japanese earthquake risk may swap a portion of its risk with a US insurer that has excess exposure to North Atlantic hurricanes. Because the two risks are uncorrelated the insurers receive diversification benefits through the swap.

The temperature derivative, an example of a non-catastrophic weather contract (*see Weather Derivatives*), can be used by a company whose revenues or expenses are exposed to changes in temperature. Consider a local gas distribution company delivering gas during the winter season to a base of residential customers. The company favors very cold winters (i.e., those with a high number of heating degree days (HDDs), or days when the average temperature exceeds 65°F), as demand for gas and gas selling prices will both rise. The reverse scenario presents a risk: warm winters (i.e., those with a small number of HDDs) mean less demand and lower prices and, thus, lower revenue. The company can protect its downside risk by selling an exchange-traded HDD future (or over-the-counter swap) based on a proximate reference city. To do so, it quantifies its exposure, determining that its revenues vary by \$1 million for each 100 change in HDDs. A warm season that generates 4700 HDDs *versus* a seasonal budget

of 5000 HDDs leads to a \$3 million loss in revenues, a cool season with 5200 HDDs leads to a \$2 million increase in revenues, and so forth. Assume the company sells 100 futures contracts on the HDD index at a level of 5000. If the season becomes very cold HDDs might rise to 5300, meaning the company loses \$3 million on its futures position; however, it earns an incremental \$3 million in revenues as a result of stronger demand for fuel and higher fuel prices. If the winter is warm, HDDs might only amount to 4700, meaning a \$3 million revenue loss. This, however, will be offset by a \$3 million futures hedge gain.

Captives

A captive is a closely held insurer/reinsurer that a company can establish to self-manage retention/transfer activities. The sponsoring company, as owner, provides up-front capital, receiving periodic interest and/or dividends in return. The captive then insures the owner (or third-party users) by accepting a transfer of risk in exchange for premium. Because insurance premiums normally cover the present value of expected losses, along with insurance acquisition costs, overhead expenses, and profit loading, the non-claims portion of the premium can be as high as 30–40% of the total. A company with highly predictable risks can create a captive in order to save on this nonclaims portion.

The captive, which can be created as a single owner/user structure (e.g., the pure captive) or a multiple owner/multiple user structure (e.g., group captive, agency captive, protected cell company), has proven popular because it provides: appropriate and flexible risk cover, particularly for exposures that might otherwise be hard to insure; lower costs, primarily by avoiding agent and broker commissions and insurance overhead/profit loadings; possible tax

advantages related to investment income, premiums and/or incurred losses; incentives to implement loss control measures; decreased earnings volatility as a result of increased cost predictability; and, incremental profit potential if third-party business is written.

Assume, for instance, that a company faces a predictable level of expected losses in its workers' compensation program and is comfortable retaining a certain amount of risk. Historically, the company has paid \$2 million per year in premiums for a standard insurance policy that transfers its workers' compensation exposure, but has estimated that it can save \$250 000 per year by retaining the risk and reinsuring through a captive. It can establish a pure captive as a licensed reinsurer for a one-time fee of \$200 000 and annual captive management fees of \$50 000. Because the captive is established as a reinsurer, the company must use a fronting insurer, which will require the payment of \$75 000 of annual fronting fees.

The captive is not expected to write third-party business, so the company will obtain no premium tax deductibility benefits. Furthermore, the company does not intend to alter its investment policy on retained funds and is expected to face a 5% cost of capital and a 34% tax rate during the 3-year planning horizon. Given these assumptions, the net present value of the decision on whether to establish and use a captive can be determined through the cash flows shown in Table 2.

The annual cash flows, discounted at a 5% cost of capital, yield a net present value of:

$$\begin{aligned} \$175\,166 = & -\$49\,500 + (\$82\,500/1.05^1) \\ & + (\$82\,500/10.05^2) + (\$82\,500/1.05^3) \end{aligned} \quad (1)$$

Since the figure is positive, the company can increase its enterprise value by retaining its workers'

Table 2 Captive cash flows^(a)

	Start	Year 1	Year 2	Year 3
Captive start-up costs	−\$200 000	—	—	—
Captive administration fee	−\$50 000	−\$50 000	−\$50 000	−\$50 000
Fronting fee	−\$75 000	−\$75 000	−75 000	−\$75 000
Insurance savings	+\$250 000	+\$250 000	+250 000	+\$250 000
Pretax cash flow	−\$75 000	+\$125 000	+\$125 000	+\$125 000
Taxes (34%)	+\$25 500	−\$42 500	−\$42 500	−\$42 500
After-tax cash flow	−\$49 500	+\$82 500	+\$82 500	+\$82 500

^(a)Reproduced from [1] with permission from John Wiley & Sons, Inc., © 2004

compensation exposures and insuring them *via* the captive.

Enterprise Risk Management Programs

Enterprise risk management (*see* **Enterprise Risk Management (ERM)**) – a risk-management process that combines disparate financial and operating risks and time horizons into a single, multiyear program of action – can provide a company with an efficient and cost-effective risk solution that may be superior to transactional or “incremental” approaches. A successful program relies on proper identification of all of a company’s risk exposures, and how they interact with one another. Consider a company facing a

series of insurance and financial exposures (Figure 3). Each individual exposure is considered and managed separately, and the end result is a series of individual insurance policies, financial derivatives, and other loss financing techniques that are designed to provide protection. This, however, may be an inefficient way of managing corporate resources, leading to excess costs, overinsurance/overhedging, and capital mismanagement—all of which can reduce enterprise value. Uniting the individual vertical covers under a comprehensive enterprise risk-management program (Figure 4) eliminates coverage gaps, lowers costs, and improves capital and administrative efficiencies.

The platform does not require that all exposures be channeled through a “master” insurance policy.

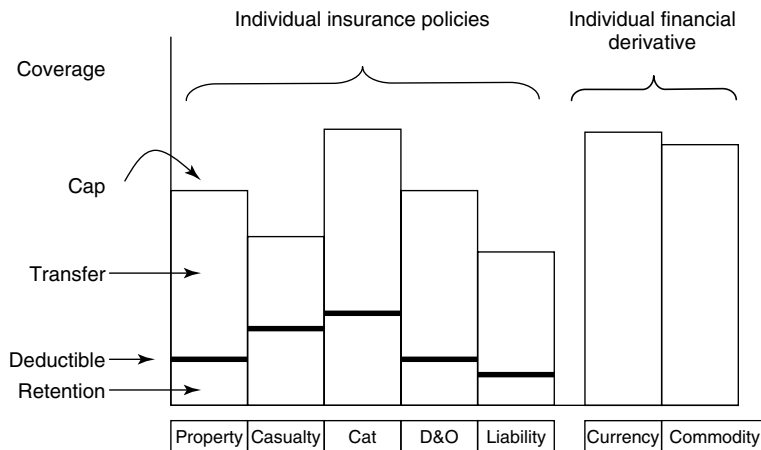


Figure 3 Individual risk exposures [Reproduced from [1] with permission from John Wiley & Sons, Inc., © 2004.]

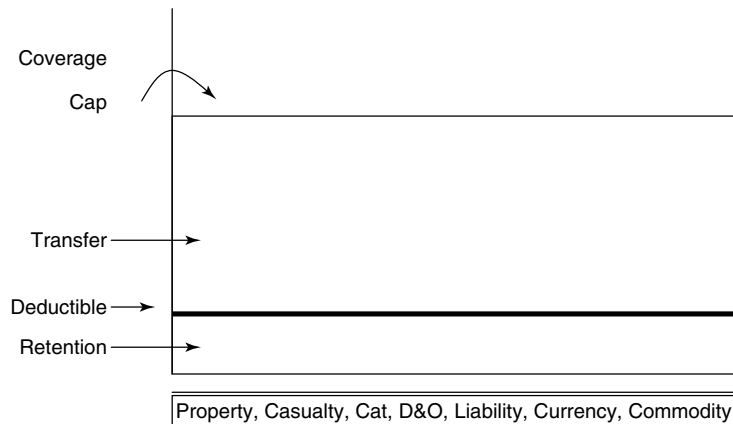


Figure 4 Consolidated risk exposures [Reproduced from [1] with permission from John Wiley & Sons, Inc., © 2004.]

For instance, the risk retained under a program can be channeled through a group captive or funded *via* liquid resources, catastrophe coverage can be based on a dual trigger contract and allow for the issuance of incremental equity, returns on an investment portfolio might be floored through the use of equity options, and so on. In fact, such flexibility is a key characteristic and advantage of the enterprise risk-management process.

Reference

- [1] Banks, E. (2004). *Alternative Risk Transfer*, John Wiley & Sons, Chichester.

Further Reading

- Doherty, N. (1985). *Corporate Risk Management: A Financial Exposition*, McGraw-Hill, New York.
- Shimpi, P. (2001). *Integrating Corporate Risk Management*, Texere, New York.

ERIK BANKS

Arsenic

Arsenic is a naturally occurring element, found throughout the environment in many different forms [1]. Inorganic arsenic compounds, mainly arsenate and arsenite, are commonly detected in soil and some groundwaters. Organic arsenic compounds (i.e., containing carbon) are also found in the environment. The most abundant organic arsenic compound is arsenobetaine, which is commonly detected in seafood and considered relatively nontoxic. Arsenic compounds have also been used as medicines, pesticides, wood preservatives, and animal feed additives. This chapter focuses on the toxicity and risk assessment of inorganic arsenic.

For most chemicals, scientists primarily rely on studies of laboratory animals to evaluate chemical toxicity (*see Cancer Risk Evaluation from Animal Studies*). Arsenic is unusual in that many human populations throughout the world have been or continue to be exposed to high levels of arsenic. Workers in certain fields, such as smelting and certain types of chemical manufacturing, have been exposed to elevated levels of arsenic in air, particularly under historical workplace conditions. Naturally occurring arsenic contamination in groundwater continues to be a concern in certain areas of the world, including parts of Taiwan, India, Mongolia, and Bangladesh. As a result, arsenic has been widely studied in humans and an extensive database of human studies is available. Scientists have not identified a representative animal model for arsenic carcinogenicity in a standard two-year bioassay [2].

Sources and Pathways of Arsenic Exposure

Inorganic arsenic occurs naturally in groundwater, with widely varying concentrations (*see Water Pollution Risk*). Although the median arsenic groundwater concentration is $\leq 1 \mu\text{g l}^{-1}$ in the US., mean well-water concentrations in certain US counties range as high as $190 \mu\text{g l}^{-1}$ [3–5]. Around the world (Taiwan, Bangladesh, China, India, Nepal, and Chile), groundwater concentrations of inorganic arsenic can average several hundred micrograms per liter [6–12]. Such levels result in significant human exposures *via* drinking water, and studies of these populations provide important information about the adverse health effects from chronic arsenic ingestion.

Inorganic arsenic is also naturally present in soil; the natural content of arsenic in soils globally ranges from 0.01 to over 600 mg kg^{-1} , with an average of about $2\text{--}20 \text{ mg kg}^{-1}$ [13]. Soils near smelting sites, or in areas where arsenical pesticides have been used, can contain elevated levels of arsenic. Human exposure to arsenic in soil can occur through incidental ingestion (e.g., while eating or during normal hand-to-mouth contact, particularly for children less than 6 years because of their increased hand-to-mouth contact), absorption from soil through the skin, and inhalation of wind-blown soil particles. As arsenic is not readily absorbed through the skin from soil [14], and the amount of exposure to naturally occurring concentrations of arsenic in soil by inhalation is quite small [15], the ingestion pathway is typically the dominant pathway of exposure for arsenic in soil (*see Soil Contamination Risk*). The bioavailability of arsenic in soil, or the amount of arsenic absorbed into the body from the gut after ingestion, ranges from about 3 to 50%, depending on soil type and test method [16–19], whereas the bioavailability of arsenic in drinking water is close to 100% [1].

Arsenic can enter the air *via* multiple sources. In addition to resuspension of arsenic in soil as airborne dust, arsenic can be also emitted through volcanic eruptions, or released into the air through smelting, municipal waste incineration, or certain other activities [1]. Airborne arsenic is generally not the most significant pathway of arsenic exposure, except in certain occupational settings.

For the general population, food is typically the largest source of arsenic exposure, with dietary exposure accounting for about 70% of the daily intake of inorganic arsenic [20]. The average daily intake of inorganic arsenic in food is about $3 \mu\text{g day}^{-1}$, with a range of about $1\text{--}20 \mu\text{g day}^{-1}$, primarily from grains, fruits, rice, and vegetables [21, 22]. Total arsenic ingestion, including organic arsenobetaines in seafood, is higher, averaging about $2\text{--}92 \mu\text{g day}^{-1}$ [23].

Human Health Effects of Inorganic Arsenic

Acute Health Effects

Arsenic has been used as a poison throughout history; very large oral doses of arsenic can be fatal. Nonetheless, present-day reports of arsenic toxicity

from acute exposure (i.e., high-dose, short-term) are rare [1, 24]. A few reported cases have involved short-term exposure to very high levels of arsenic in well water ($1200\text{--}10\,000\ \mu\text{g l}^{-1}$) [24–26]. Most other information regarding acute arsenic toxicity in humans derives from its medicinal use or from incidents in which individuals were poisoned with arsenic. Acute high doses can lead to effects such as encephalopathy, nausea, anorexia, vomiting, abdominal pain, cardiovascular effects, hepatotoxicity, hematologic abnormalities, vascular lesions, peripheral neuropathy, and death [1]. For example, in a case of contaminated soy-sauce in Japan, doses of about $0.05\ \text{mg kg}^{-1}\ \text{day}^{-1}$ for 2–3 weeks resulted in facial edema and mild gastrointestinal effects [27]. More severe effects, such as damage to the peripheral nervous system, occur only when acute doses are about $2\ \text{mg kg}^{-1}\ \text{day}^{-1}$ or higher (e.g., [26, 28, 29]).

Chronic Health Effects – Noncancer

Chronic exposures to high doses of arsenic are associated with various noncancer health effects in humans, the most sensitive of which are skin and possibly vascular effects [30–32]. There is no convincing association between arsenic exposure in humans and reproductive/developmental effects [30, 33, 34]. Overall, it is not until chronic arsenic doses exceed about $0.01\ \text{mg kg}^{-1}\ \text{day}^{-1}$ that there is an established association between arsenic exposure and noncancer health effects in humans [30].

Skin effects are generally recognized to be the most sensitive and characteristic noncancer effects resulting from long-term exposure to arsenic [1, 30–32, 35, 36]. Arsenic-induced pigmentation changes are often described as a “raindrop” pattern, and arsenic-induced hyperkeratosis is more frequently observed on the palms and soles [37]. A key arsenic epidemiologic study, initially published in 1968 by Tseng *et al.*, investigated over 40 000 Taiwanese people exposed to high levels of arsenic in drinking water [38, 39]. Hyperpigmentation and hyperkeratosis were increased in all exposed groups; the lowest dose at which effects were observed was $170\ \mu\text{g l}^{-1}$. In subsequent studies around the world, similar effect levels for arsenic-induced skin changes have been reported [40–44]. Skin effects have also been observed in workers exposed to arsenic in air ranging as high as $1\text{--}100\ \text{mg m}^{-3}$ [1, 45–48].

Consumption of high levels of arsenic in drinking water in non-US populations has also been associated with vascular effects, including peripheral vascular disease, high blood pressure, cardiovascular and cerebrovascular disease, circulatory problems, ischemic heart disease, and Raynaud’s disease [49–58]. In contrast, there is no convincing evidence for cardiovascular effects in US studies [5, 30, 59]. “Blackfoot disease”, a vascular disease characterized by loss of circulation in the hands and feet, is endemic in certain areas of Taiwan with $170\text{--}800\ \mu\text{g l}^{-1}$ arsenic in drinking water, but has not been reported outside of Taiwan [1, 30, 39]. Furthermore, dietary deficits and other drinking water contaminants may play a causative role in Blackfoot disease in Taiwan [60]. The evidence for an association between inhaled arsenic and cardiovascular effects is more limited; despite an increase in cardiovascular disease, some of the studies failed to show a clear dose–response relationship (e.g., [61, 62]).

Long-term exposure to high doses of arsenic can lead to peripheral neuropathy, but reports of neurological effects at lower arsenic doses are inconsistent [1, 63–65]. For example, chronic exposures to arsenic in drinking water in China (mean $158.3\ \mu\text{g l}^{-1}$) were not associated with changes in nerve conduction velocity [66]. In contrast, in a recent Bangladesh study (mean arsenic, $115\ \mu\text{g l}^{-1}$), the authors reported a significant association between toe vibration threshold (a “subclinical” indicator of sensory neuropathy) and some, but not all, measures of exposure.

A limited number of publications have reported an association between elevated arsenic exposure (milligrams of arsenic ingested per day) and diabetes mellitus [30, 67–71]. Because diabetes mellitus occurs only at high arsenic exposures, and often in conjunction with skin effects, it is not considered a sensitive arsenic-induced health endpoint [35]. Gastrointestinal symptoms have also been reported with chronic exposures to arsenic, but generally not until exposures of $0.04\ \text{mg kg}^{-1}\ \text{day}^{-1}$ or more, and usually in conjunction with skin effects [1, 30, 40, 41, 43, 72].

Chronic Health Effects – Cancer

On the basis of studies of human populations outside the United States, arsenic has been classified as a “human carcinogen” by the United States Environmental Protection Agency (US EPA) [73] and

as “carcinogenic to humans” by the International Agency for Research on Cancer (IARC), part of the World Health Organization (WHO) [74]. Exposure to high levels of arsenic in drinking water, usually greater than $150\text{ }\mu\text{g l}^{-1}$, is associated with cancer of the lung, bladder, and skin, in multiple studies conducted in countries such as Taiwan [39, 75–78], Chile [11], Bangladesh [6], and Inner Mongolia [79]. A number of other cancers, such as kidney and liver cancer, have been associated with exposure to high levels of arsenic in drinking water, but the associations are not as strong and have not been consistently observed [35]. Reports linking arsenic exposure to other cancers (e.g., prostate, colon, stomach, nasal, laryngeal, and leukemia) are isolated, inconsistent, and lack sufficient statistical substantiation [5, 35, 58, 80, 81].

The most comprehensive study on skin cancer was performed by Tseng *et al.* [39]; they compared skin cancer incidence in an arsenic-exposed population of over 40 000 people and identified a dose-dependent relationship between arsenic and skin cancer. In the 1980s, a series of epidemiologic studies were published reporting the relationship between arsenic exposure and other cancer sites in the same Southwestern Taiwanese population studied by Tseng *et al.* [38, 39]. Chen *et al.* found increased cancer deaths from bladder, kidney, skin, lung, liver, and colon cancers in individuals exposed to artesian well water and made some general statements about the relationship between arsenic exposure and increased cancer incidence [82]. In a follow-up study of over 32 000 people from 42 different villages with arsenic concentrations ranging from 10 to $935\text{ }\mu\text{g l}^{-1}$, Wu *et al.* found that mortality increased with increasing arsenic concentrations in water for cancers of the bladder, kidney, skin, lung, liver, prostate, and leukemia [58]. US EPA has relied solely on these Taiwanese studies to develop the cancer toxicity factors used in risk assessments, and, thus, arsenic regulatory standards and criteria are based on the Taiwanese data. However, while the Taiwanese studies are useful because of their large size and quantitative information regarding high doses of arsenic and cancer, they offer limited information about the effects of arsenic at low doses.

With respect to bladder cancer, results from international epidemiologic studies have shown effects only at relatively high ($>150\text{ }\mu\text{g l}^{-1}$) arsenic exposure levels, if at all. Bates *et al.* found no association between arsenic ingestion and bladder cancer

in Argentina, even at concentrations $>200\text{ }\mu\text{g l}^{-1}$, contradicting a finding of increased bladder cancer at lower, but less well defined, arsenic exposures in an earlier Argentinean study [83, 84]. In a Taiwanese study, Chiou *et al.* found no increase in bladder cancer incidence in Taiwan [78], even at arsenic levels of $710\text{ }\mu\text{g l}^{-1}$ in drinking water [30]. Guo *et al.* found a statistically significant increase in bladder cancer only at arsenic concentrations of $640\text{ }\mu\text{g l}^{-1}$ or more in drinking water [76]. Lamm *et al.* reanalyzed data from Taiwan and concluded that arsenic-induced bladder cancer occurs only at levels $>400\text{ }\mu\text{g l}^{-1}$, and that certain artesian wells in Taiwan may have contained contaminants (e.g., humic acids, fluorescent substances, and fungi) that increase the potency of arsenic [85]. Lamm *et al.* subsequently reanalyzed the extensive Taiwanese dataset, and demonstrated that geographically related risk factors (by township) for bladder and lung cancer may have distorted previous analyses of these data [86]. Separating out these factors by excluding townships where arsenic well water was not the most important determinant of cancer, Lamm revealed that the dose–response for arsenic-related bladder and lung cancer had an apparent threshold, increasing only at concentrations above $150\text{ }\mu\text{g l}^{-1}$ in drinking water.

Lung cancer has been associated with both arsenic ingestion (*via* drinking water) and inhalation. Guo reported elevated lung cancer mortality in Taiwan at concentrations $>640\text{ }\mu\text{g l}^{-1}$ in drinking water [77], although in a different Taiwanese population, Chiou *et al.* did not find an increase in lung cancer with exposures to arsenic levels of $710\text{ }\mu\text{g l}^{-1}$ and higher [30, 78]. Workers with high-dose, long-term occupational exposure to airborne arsenic have also shown an increase in lung cancer [1, 87–89].

Well-designed US epidemiologic studies show no evidence of increased cancer risk at lower arsenic doses. In a large study in Utah [5], US EPA found no convincing evidence of carcinogenic or noncancer effects at average arsenic concentrations in drinking water up to almost $200\text{ }\mu\text{g l}^{-1}$. In a *case–control* study (*see Case–Control Studies*) using estimates of arsenic exposure based on individual intakes, Steinmaus *et al.* found no clear association between arsenic and bladder cancer even in individuals exposed to arsenic in excess of $80\text{ }\mu\text{g day}^{-1}$ for 20 years [4].

Discussion

Overall, consistent noncancer and cancer health effects have been observed only in populations outside the United States with relatively high concentrations of arsenic in drinking water, often in populations suffering from nutritional deficiencies that may increase susceptibility to arsenic [72, 90, 91]. Extrapolation of health risks from arsenic drinking water studies conducted outside the United States may lead to an overestimate of risks for populations where arsenic exposures are significantly lower and nutritional status is better.

Interestingly, there are no reliable studies demonstrating that ingestion of arsenic-contaminated soil or dust has led to toxicity. Several studies have shown that chronic exposure to arsenic in soil at concentrations $>100 \text{ mg kg}^{-1}$ is not associated with an increase in body burden or adverse effects [92–96].

Arsenic Measurement in Humans

Urine arsenic measurements are considered a useful biomarker for recent arsenic exposures. Most inorganic arsenic is excreted from the body in urine within 1–2 days, and typical urine arsenic levels in the United States are about $50 \mu\text{g l}^{-1}$ or less [30, 97]. Since levels of arsenic in urine can vary on the basis of dietary habits (seafood consumption can raise levels up to $2000 \mu\text{g l}^{-1}$), it can be important for laboratories to distinguish arsenobetaine (the relatively nontoxic form of arsenic in seafood) from inorganic arsenic and its metabolites.

Arsenic in hair and fingernail can be a qualitative way to assess long-term exposure to arsenic, confirming past exposures up to 6–12 months. Normal arsenic levels in hair and nails are 1 mg kg^{-1} or less [1]. Measurements of nail and hair arsenic may be misleading due to the presence of arsenic adsorbed to the external surfaces of hair and fingernail clippings [1, 97]. Arsenic clears the bloodstream within a few hours of exposure [98], and is therefore a poor means to quantify arsenic exposure.

Arsenic Risk Assessment

Arsenic Toxicity Criteria

Assessing arsenic health risks involves consideration of the magnitude, duration, and route of arsenic

exposure, together with toxicity information from epidemiologic studies. US EPA uses noncancer and cancer toxicity information to develop chemical-specific toxicity factors; these values are published on the Integrated Risk Information System (IRIS). The IRIS database serves as an important resource because it allows scientists to standardize the risk assessment process, using a common set of toxicity criteria.

Noncancer. The US EPA oral reference dose (RfD) for arsenic, a chronic dose that is believed to be without significant risk of causing adverse noncancer effects in even susceptible humans, is $0.0003 \text{ mg kg}^{-1} \text{ day}^{-1}$, based on skin and vascular lesions in the Taiwanese study by Tseng *et al.* [39]. The Agency for Toxic Substances and Disease Registry (ATSDR) independently develops chemical-specific toxicity criteria based on noncancer health effects. The ATSDR minimal risk level (MRL) for arsenic, an “estimate of daily human exposure to a substance that is likely without an appreciable risk of adverse effects (noncarcinogenic) over a specified duration of exposure”, is also $0.0003 \text{ mg kg}^{-1} \text{ day}^{-1}$ for chronic oral exposures, based on the Tseng *et al.* study [1, 39]. ATSDR has also developed a provisional MRL of $0.005 \text{ mg kg}^{-1} \text{ day}^{-1}$ for acute-duration (≤ 14 days) oral exposures [1]. Neither EPA nor ATSDR has developed a toxicity factor for noncancer effects of inhaled arsenic.

Cancer – Oral Arsenic Exposures. For compounds that are considered known or likely carcinogens, the US EPA develops cancer slope factors (CSFs) using mathematical models to extrapolate risks observed at high doses (either in humans or animals) to lower doses reflective of more typical human exposure levels. Use of a CSF to quantify cancer risks often involves the default assumption that the dose–response relationship is linear at low doses. That is, even a very small dose of arsenic confers some excess cancer risk, and as the dose increases linearly, the risk increases linearly. For many compounds, including arsenic, it is likely that this assumption is not correct; the dose–response relationship is likely to be sublinear or have a threshold [99]. From a toxicological perspective, this means that low doses of arsenic would be relatively less effective than higher doses, and may, in fact, be associated with zero risk.

The oral cancer potency of inorganic arsenic as well as the assumption of linearity has been a source of substantial scientific debate, giving rise to several different evaluations of the carcinogenic potency of arsenic (Table 1). Analyses by US government agencies have suggested CSFs for arsenic ranging from 0.4 to 23 kg day mg⁻¹, an almost 60-fold range. Differences are mainly dependent on the type of cancer used in the evaluation, assumptions used to relate the high-dose Taiwanese data to low-dose arsenic exposure in the US, and choice of control populations. The US EPA CSF for arsenic of 1.5 kg day mg⁻¹ that is currently published on IRIS [73] is based on skin cancer in the Tseng Taiwan study [39]. Although there are several more recent CSF evaluations for arsenic, EPA has not yet updated the arsenic CSF value published on IRIS, and thus, 1.5 kg day mg⁻¹ is used most commonly in arsenic risk assessments.

Although the various evaluations of the arsenic CSF use different assumptions to arrive at different quantitative estimates of the potency of arsenic, all the analyses contain certain conservative elements and are thus likely to overestimate arsenic cancer risk for United States and other Western populations with generally low to modest levels of arsenic exposure. There are considerable scientific concerns about the exposure estimates in the Taiwanese study. Individual exposures were not characterized, and

exposures were based on average arsenic concentrations of groundwater in wells in each village [105]. What is particularly relevant is the recent evidence of a confounding factor in well water from certain townships that was associated with cancer [86]. Genetic factors, dietary patterns, and other lifestyle factors may alter arsenic metabolism and detoxification [106–108]. Nutritional deficiencies contribute to arsenic susceptibility [72, 90, 91]. Therefore, the use of a CSF derived based on relatively high levels of arsenic in water in a population with nutritional deficiencies (e.g., Taiwan) may overestimate cancer risks for populations where arsenic exposures are significantly lower and nutritional status is better.

Finally, there is convincing human and mechanistic data supporting a nonlinear dose–response relationship for arsenic. As discussed above, US studies do not indicate an increased cancer risk, even with levels as high as 200 µg l⁻¹ in drinking water, and in other parts of the world, arsenic does not cause bladder, lung, or skin cancer until levels in drinking water are greater than 150 µg l⁻¹ (e.g. [85, 86]). The indication from human studies for a nonlinear or threshold dose–response for arsenic carcinogenicity is further supported by a mechanistic understanding of how arsenic interacts with DNA to produce carcinogenic changes. Specifically, arsenic does not interact directly with DNA to produce point

Table 1 Summary of US cancer slope factors for arsenic

Slope factor (mg/kg-day) ⁻¹	Source	Agency	Comments
1.5	Integrated Risk Information System (IRIS)	US EPA [73]	Currently listed in IRIS; based on skin cancer incidence in Taiwan
0.4–3.67	Final rule for arsenic maximum contaminant level (MCL) for arsenic in drinking water	US EPA [100]	Based on bladder and lung as opposed to skin cancer; also based on Taiwanese water intake and arsenic in food
As high as 23	National Research Council (NRC) arsenic in drinking water report	National Academy of Sciences (NAS), review panel for US EPA [35]	Based on lung and bladder cancer risk in Taiwan, using more conservative modeling assumptions
3.67	Draft Chromated Copper Arsenate (CCA) reregistration, CCA risk assessment, and organic arsenic herbicide reregistration	US EPA [101–103]	Based on upper range established in MCL rule
0.41–23	Petition to ban CCA wood	Consumer Product Safety Commission (CPSC)	Based on EPA and NRC assessments
5.7	Proposed IRIS revision	US EPA [104]	Based on bladder and lung cancer; incorporates many of NRC's recommendations

mutations, but instead may modify DNA function through one or more indirect mechanisms, including chromosome alterations, changes in methylation status of DNA, and alterations in gene transcription [99, 109]. This indicates that arsenic carcinogenicity likely has a sublinear dose–response. Thus, assuming a linear dose–response relationship, as the US EPA and National Academy of Sciences (NAS) have done in their assessments, likely overestimates arsenic risk at low levels of exposure.

Cancer – Inhalation Arsenic Exposures. The US EPA inhalation unit risk for arsenic is $4.3 \times 10^{-3} \text{ m}^3 \mu\text{g}^{-1}$, based on linear extrapolation of lung cancer risks from several studies of smelter workers with occupational exposures to airborne arsenic [45, 47, 73, 110–112].

Regulatory Standards and Criteria for Arsenic in Environmental Media

Regulatory standards and criteria for environmental media are derived using toxicity criteria (RfDs and CSFs), human exposure assumptions, and other information. In January 2001, the US EPA lowered the maximum contaminant level (MCL) for arsenic in drinking water from 50 to $10 \mu\text{g l}^{-1}$ [100], based on a reevaluation of the carcinogenic potency of arsenic with a focus on bladder and lung cancer (as opposed to skin cancer). WHO has also set a guideline value of $10 \mu\text{g l}^{-1}$ for arsenic in drinking water [113]. The US EPA soil screening level (SSL) for arsenic in residential settings is 0.4 mg kg^{-1} [15]. As naturally occurring levels of arsenic are often elevated above standard regulatory risk thresholds, it is important for risk managers to put arsenic exposures in perspective, and consider background when setting clean-up levels.

References

- [1] Agency for Toxic Substances and Disease Registry (ATSDR) (2005). *Toxicological Profile for Arsenic (Update) (Draft for Public Comment)*, at <http://www.atsdr.cdc.gov/toxprofiles/tp2.pdf> (accessed Aug 2005), p. 533.
- [2] Huff, J., Chan, P. & Nyska, A. (2000). Is the human carcinogen arsenic carcinogenic to laboratory animals? *Toxicological Sciences* **55**(1), 17–23.
- [3] Focazio, M.J., Welch, A.H., Watkins, S.A., Helsel, D.R., Horn, M.A. & US Geological Survey (1999). A retrospective analysis on the occurrence of arsenic in ground-water resources of the United States and limitations in drinking-water-supply characterizations, USGS Water-Resources Investigation Report 99–4279, at <http://co.water.usgs.gov/trace/pubs/wrir-99-4279> (accessed May 2000), p. 27.
- [4] Steinmaus, C., Yuan, Y., Bates, M.N. & Smith, A.H. (2003). Case–control study of bladder cancer and drinking water arsenic in the Western United States, *American Journal of Epidemiology* **158**, 1193–2001.
- [5] Lewis, D.R., Southwick, J.W., Ouellet-Hellstrom, R., Rench, J. & Calderon, R.L. (1999). Drinking water arsenic in Utah: a cohort mortality study, *Environmental Health Perspectives* **107**(5), 359–365.
- [6] Rahman, M.M., Chowdhury, U.K., Mukherjee, S.C., Mondal, B.K., Paul, K., Lodh, D., Biswas, B.K., Chanda, C.R., Basu, G.K., Saha, K.C., Roy, S., Das, R., Palit, S.K., Quamruzzaman, Q. & Chakraborti, D. (2001). Chronic arsenic toxicity in Bangladesh and West Bengal, India: a review and commentary, *Clinical Toxicology* **39**(7), 683–700.
- [7] Lianfang, W. & Jianzhong, H. (1994). Chronic arsenism from drinking water in some areas of Xinjiang, China, in *Arsenic in the Environment, Part II: Human Health and Ecosystem Effects*, J.O. Nriagu, ed, John Wiley & Sons, New York, pp. 159–172.
- [8] Chakraborti, D., Sengupta, M.K., Rahman, M.M., Chowdhury, U.K., Lodh, D.D., Ahamed, S., Hossain, M.A., Basu, G.K., Mukherjee, S.C. & Saha, K.C. (2003). Groundwater arsenic exposure in India, in *Arsenic Exposure and Health Effects V*, W.R. Chappell, C.O. Abernathy, R.L. Calderon & D.J. Thomas, eds, Elsevier Science, Amsterdam, pp. 3–24.
- [9] Shrestha, R.R., Shrestha, M.P., Upadhyay, N.P., Pradhan, R., Khadka, R., Maskey, A., Tuladhar, S., Dahal, B.M., Shrestha, S. & Shrestha, K. (2003). Groundwater arsenic contamination in Nepal: a new challenge for water supply sector, in *Arsenic Exposure and Health Effects V*, W.R. Chappell, C.O. Abernathy, R.L. Calderon & D.J. Thomas, eds, Elsevier Science, Amsterdam, pp. 25–37.
- [10] Chen, S.L., Dzen, S.R., Yang, M.-H., Chiu, K.-H., Shieh, G.-M. & Wai, C.M. (1994). Arsenic species in groundwaters of the blackfoot disease area, Taiwan, *Environmental Science and Technology* **28**(5), 877–881.
- [11] Ferreccio, C., Gonzalez, C., Milosavljevic, V., Marshall, G., Sancha, A.M. & Smith, A.H. (2000). Lung cancer and arsenic concentrations in drinking water in Chile, *Epidemiology* **11**(6), 673–679.
- [12] Nordstrom, D.K. (2002). Worldwide occurrences of arsenic in ground water, *Science* **296**, 2143–2145.
- [13] Yan-Chu, H. (1994). Arsenic distribution in soils, in *Arsenic in the Environment, Part I: Cycling and Characterization*, J.O. Nriagu, ed, John Wiley & Sons, New York, pp. 99–118.
- [14] US EPA (2001). *Risk Assessment Guidance for Superfund Volume I: Human Health Evaluation Manual (Part*

- E, Supplemental Guidance for Dermal Risk Assessment Interim Review Draft for Public Comment*, Office of Emergency and Remedial Response, Washington, DC.
- [15] US EPA. (1996). *Soil Screening Guidance: Technical Background Document*, Office of Solid Waste and Emergency Response, NTIS PB96-963502, EPA-540/R-95/128, OSWER Publication 9355.4-17A, at <http://www.epa.gov/oerrpage/superfund/resources/soil/toc.htm> (accessed on Jan 2001).
 - [16] Roberts, S.M., Munson, J.W., Lowney, Y.W. & Ruby, M.V. (2007). Relative oral bioavailability of arsenic from contaminated soils measured in the cynomolgus monkey, *Toxicological Sciences* **95**(1), 281–288.
 - [17] Roberts, S.M., Weimar, W.R., Vinson, J.R., Munson, J.W. & Bergeron, R.J. (2002). Measurement of arsenic bioavailability from soils using a primate model, *Toxicological Sciences* **67**, 303–310.
 - [18] Rodriguez, R.R., Basta, N.T., Casteel, S.W. & Pace, L.W. (1999). An *in vitro* gastrointestinal method to estimate bioavailable arsenic in contaminated soils and solid media, *Environmental Science and Technology* **33**(4), 642–649.
 - [19] Casteel, S.W., Brown, L.D., Dunsmore, M.E., Weis, C.P., Henningsen, G.M., Hoffman, E., Brattin, W.J. & Hammon, T.L. (1997). *Relative Bioavailability of Arsenic in Mining Wastes. EPA Region VIII Report*.
 - [20] Borum, D.R. & Abernathy, C.O. (1994). Human oral exposure to inorganic arsenic, in *Arsenic, Exposure and Health*, W.R. Chappell, C.O. Abernathy & C.R. Cothorn, eds, Science and Technology Letters, Northwood, pp. 21–29.
 - [21] Schoof, R.A., Eickhoff, J., Yost, L.J., Crecelius, E.A., Cragin, D.W., Meacher, D.M. & Menzel, D.B. (1999). Dietary exposure to inorganic arsenic, in *Arsenic Exposure and Health Effects*, W.R. Chappell, C.O. Abernathy & R.L. Calderon, eds, Elsevier Science, pp. 81–88.
 - [22] Yost, L.J., Tao, S.H., Egan, S.K., Barraj, L.M., Smith, K.M., Tsuji, J.S., Lowney, Y.W., Schoof, R.A. & Rachman, N.J. (2004). Estimation of dietary intake of inorganic arsenic in U.S. children, *Human and Ecological Risk Assessment* **10**, 473–483.
 - [23] Tao, S.S. & Bolger, P.M. (1999). Dietary arsenic intakes in the United States: FDA total diet study, September 1991–December 1996, *Food Additives and Contaminants* **16**, 465–472.
 - [24] Franzblau, A. & Lilis, R. (1989). Acute arsenic intoxication from environmental arsenic exposure, *Archives of Environment Health* **44**(6), 385–390.
 - [25] Wagner, S.L., Maliner, J.S., Morton, W.E. & Braman, R.S. (1979). Skin cancer and arsenical intoxication from well water, *Archives of Dermatology* **115**(10), 1205–1207.
 - [26] Armstrong, C.W., Stroube, R.B., Rubio, T., Siudyla, E.A. & Miller, G.B. (1984). Outbreak of fatal arsenic poisoning caused by contaminated drinking water, *Archives of Environment Health* **39**(4), 276–279.
 - [27] Mizuta, N., Mizuta, M., Ito, F., Ito, T., Uchida, H., Watanabe, Y., Akama, H., Murakami, T., Hayashi, F., Nakamura, K., Yamaguchi, T., Mizuizawa, W., Oishi, S. & Matsumura, H. (1956). An outbreak of acute arsenic poisoning caused by arsenic contaminated soy-sauce (shoyu): a clinical report of 220 cases, *The Bulletin of the Yamaguchi Medical School* **4**(2/3), 131–149.
 - [28] Cullen, N.M., Wolf, L.R. & St. Clair, D. (1995). Pediatric arsenic ingestion, *The American Journal of Emergency Medicine* **13**, 432–435.
 - [29] Quatrehomme, G., Ricq, O., Lapalus, P., Jacomet, Y. & Ollier, A. (1992). Acute arsenic intoxication: forensic and toxicologic aspects (an observation), *Journal of Forensic Sciences* **37**(4), 1163–1171.
 - [30] National Research Council (NRC), Subcommittee on Arsenic in Drinking Water (1999). *Arsenic in Drinking Water*, National Academy Press, Washington, DC, p. 310.
 - [31] Tsuda, T., Babazono, A., Yamamoto, E., Kurumatani, N., Mino, Y., Ogawa, T., Kishi, Y. & Aoyama, H. (1995). Ingested arsenic and internal cancer: a historical cohort study followed for 33 years, *American Journal of Epidemiology* **141**, 198–209.
 - [32] Crump, K., Clewell, H. & Yager, J. (2000). Noncancer risk assessment for arsenic based on its vascular effects, *Toxicologist* **54**(1), 74.
 - [33] DeSesso, J.M., Jacobson, C.F., Scialli, A.R., Farr, C.H. & Holson, J.F. (1998). An assessment of the developmental toxicity of inorganic arsenic, *Reproductive Toxicology* **12**, 385–433.
 - [34] DeSesso, J.M. (2001). Teratogen update: inorganic arsenic, *Teratology* **64**(3), 170–173.
 - [35] National Research Council (NRC), Subcommittee on Arsenic in Drinking Water (2001). *Arsenic in Drinking Water: 2001 Update*, National Academy Press, Washington, DC, p. 189.
 - [36] Brown, K.G., Kuo, T.L., Guo, H.R., Ryan, L.M. & Abernathy, C.O. (2000). Sensitivity analysis of U.S. EPA's estimates of skin cancer risk from inorganic arsenic in drinking water, *Human and Ecological Risk Assessment* **6**, 1055–1074.
 - [37] Yu, H.S., Sheu, H.M., Ko, S.S., Chiang, L.C., Chien, C.H., Lin, S.M., Tserng, B.R. & Chen, C.S. (1984). Studies on blackfoot disease and chronic arsenism in Southern Taiwan, with special reference to skin lesions and fluorescent substances, *Journal of Dermatology* **11**, 361–370.
 - [38] Tseng, W.P. (1977). Effects and dose-response relationships of skin cancer and blackfoot disease with arsenic, *Environmental Health Perspectives* **19**, 109–119.
 - [39] Tseng, W.P., Chu, H.M., How, S.W., Fong, J.M., Lin, C.S. & Yeh, S. (1968). Prevalence of skin cancer in an endemic area of chronic arsenicism in Taiwan, *Journal of the National Cancer Institute* **4**(3), 453–463.
 - [40] Borgono, J.M. & Greiber, R. (1972). Epidemiological study of arsenicism in the city of Antofagasta, *Trace Substances in Environmental Health – V*, University of Missouri, Columbia, Missouri, pp. 13–24.

- [41] Zaldivar, R. (1974). Arsenic contamination of drinking water and foodstuffs causing endemic chronic poisoning, *Beitrage Zur Pathologie Bd* **151**, 384–400.
- [42] Borgono, J.M., Venturino, H. & Vicent, P. (1980). Clinical and epidemiologic study of arsenicism in Northern Chile, *Revista Medica De Chile* **108**, 1039–1048.
- [43] Cebrian, M.E., Albores, A., Aguilar, M. & Blakely, E. (1983). Chronic arsenic poisoning in the north of Mexico, *Human Toxicology* **2**, 121–133.
- [44] Haque, R., Guha Mazumder, D.N., Samanta, S., Ghosh, N., Kalman, D., Smith, M.M., Mitra, S., Santra, A., Lahiri, S., Das, S., De, B.K. & Smith, A.H. (2003). Arsenic in drinking water and skin lesions: dose-response data from West Bengal, India, *Epidemiology* **14**(2), 174–182.
- [45] Enterline, P.E. & Marsh, G.M. (1982). Cancer among workers exposed to arsenic and other substances in a copper smelter, *American Journal of Epidemiology* **116**(6), 895–911.
- [46] Jarup, L., Pershagen, G. & Wall, S. (1989). Cumulative arsenic exposure and lung cancer in smelter workers: a dose-response study, *American Journal of Industrial Medicine* **15**, 31–41.
- [47] Lee-Feldstein, A. (1986). Cumulative exposure to arsenic and its relationship to respiratory cancer among copper smelter employees, *Journal of Occupational Medicine* **28**(4), 296–302.
- [48] Perry, K., Bowler, R.G., Buckell, H.M., Druett, H.A. & Schilling, R.S.F. (1948). Studies in the incidence of cancer in a factory handling inorganic compounds of arsenic II: clinical and environmental investigations, *British Journal of Industrial Medicine* **5**, 6–15.
- [49] Chen, C.J., Hsueh, Y.M., Lai, M.S., Shyu, M.P., Chen, S.Y., Wu, M.M., Kuo, T.L. & Tai, T.Y. (1995). Increased prevalence of hypertension and long-term arsenic exposure, *Hypertension* **25**, 53–60.
- [50] Chen, C.J., Chiou, H.Y., Chiang, M.H., Lin, L.J. & Tai, T.Y. (1996). Dose-response relationship between ischemic heart disease mortality and long-term arsenic exposure, *Arteriosclerosis Thrombosis and Vascular Biology* **16**, 504–510.
- [51] Chiou, H.Y., Huang, W.I., Su, C.L., Chang, S.F., Hsu, Y.H. & Chen, C.J. (1997). Dose-response relationship between prevalence of cerebrovascular disease and ingested inorganic arsenic, *Stroke* **28**, 1717–1723.
- [52] Hsueh, Y.M., Wu, W.L., Huang, Y.L., Chiou, H.Y., Tseng, C.H. & Chen, C.J. (1998). Low serum carotene level and increased risk of ischemic heart disease related to long-term arsenic exposure, *Atherosclerosis* **141**(2), 249–257.
- [53] Navas-Acien, A., Sharrett, A.R., Silbergeld, E.K., Schwartz, B.S., Nachman, K.E., Burke, T.A. & Guallar, E. (2005). Arsenic exposure and cardiovascular disease: a systematic review of the epidemiologic evidence, *American Journal of Epidemiology* **162**(11), 1037–1049.
- [54] Rahman, M., Tondel, M., Ahmad, S.A., Chowdhury, I.A., Faruquee, M.H. & Axelson, O. (1999). Hypertension and arsenic exposure in Bangladesh, *Hypertension* **33**, 74–78.
- [55] Tseng, C.H., Chong, C.K., Chen, C.J. & Tay, T.Y. (1997). Lipid profile and peripheral vascular disease in arseniasis-hyperendemic villages in Taiwan, *Angiology* **28**, 321–335.
- [56] Tseng, C.H., Chong, C.K., Tseng, C.P., Hsueh, Y.M., Chiou, H.Y., Tseng, C.C. & Chen, C.J. (2003). Long-term arsenic exposure and ischemic heart disease in arseniasis-hyperendemic villages in Taiwan, *Toxicology Letters* **137**(1–2), 15–21.
- [57] Yu, H.S., Lee, C.H. & Chen, G.S. (2002). Peripheral vascular diseases resulting from chronic arsenical poisoning, *Journal of Dermatology* **29**, 123–130.
- [58] Wu, M.M., Kuo, T.L., Hwang, Y.H. & Chen, C.J. (1989). Dose-response relationship between arsenic concentration in well water and mortality from cancers and vascular diseases, *American Journal of Epidemiology* **130**(6), 1123–1132.
- [59] Engel, R.R. & Smith, A.H. (1994). Arsenic in drinking water and mortality from vascular disease: an ecologic analysis in 30 counties in the United States, *Archives of Environment Health* **49**(5), 418–427.
- [60] Lu, F.J. (1990). Blackfoot disease: arsenic or humic acid? *Lancet* **336**, 115–116.
- [61] Enterline, P.E., Day, R. & Marsh, G.M. (1995). Cancers related to exposure to arsenic at a copper smelter, *Occupational and Environmental Medicine* **52**, 28–32.
- [62] Welch, K., Higgins, I., Oh, M. & Burchfiel, C. (1982). Arsenic exposure, smoking, and respiratory cancer in copper smelter workers, *Archives of Environment Health* **37**(6), 325–335.
- [63] Southwick, J.W., Western, A.E., Beck, M.M., Whitley, T., Isaacs, R., Petajan, J. & Hansen, C.D. (1983). An epidemiological study of arsenic in drinking water in Millard County, Utah, *Arsenic: Industrial, Biomedical, and Environmental Perspectives. Proceedings of Arsenic Symposium, 1983*, Van Nostrand Reinhold, New York.
- [64] Kreiss, K., Zack, M.M., Landrigan, P.J., Feldman, R.G., Niles, C.A., Chirico-Post, J., Sax, D.S., Boyd, M.H. & Cox, D.H. (1983). Neurologic evaluation of a population exposed to arsenic in Alaskan well water, *Archives of Environment Health* **38**(2), 116–121.
- [65] Hindmarsh, J.T., McLetchie, O.R., Heffernan, L.P.M., Hayne, O.A., Ellenberger, H.A., McCurdy, R.F. & Thiebaut, H.J. (1977). Electromyographic abnormalities in chronic environmental arsenicalism, *Journal of Analytical Toxicology* **1**, 270–276.
- [66] Japan Inner Mongolia Arsenic Pollution Study Group (JIMAPSG) (2006). Arsenic in drinking water and peripheral nerve conduction velocity among residents of a chronically arsenic-affected area in Inner Mongolia, *Journal of Epidemiology* **16**(5), 207–213.
- [67] Lai, M.S., Hsueh, Y.M., Chen, C.J., Shyu, M.P., Chen, S.Y., Kuo, T.L., Wu, M.M. & Tai, T.Y. (1994).

- Ingested inorganic arsenic and prevalence of diabetes mellitus, *American Journal of Epidemiology* **139**, 484–492.
- [68] Tseng, C.H., Tseng, S.P., Chiou, H.Y., Hsueh, Y.M., Chong, C.K. & Chen, C.J. (2002). Epidemiologic evidence of diabetogenic effect of arsenic, *Toxicology Letters* **133**, 69–76.
- [69] Rahman, M., Tondel, M., Ahmad, S.A. & Axelson, O. (1998). Diabetes mellitus associated with arsenic exposure in Bangladesh, *American Journal of Epidemiology* **148**(2), 198–203.
- [70] Tsai, S.M., Wang, T.N. & Ko, Y.C. (1999). Mortality for certain diseases in areas with high levels of arsenic in drinking water, *Archives of Environment Health* **54**, 186–193.
- [71] Wang, S.L., Chiou, J.M., Chen, C.J., Tseng, C.H., Chou, W.L., Wang, C.C., Wu, T.N. & Chang, L.W. (2003). Prevalence of non-insulin-dependent diabetes mellitus and related vascular diseases in southwestern arseniasis-endemic and nonendemic areas in Taiwan, *Environmental Health Perspectives* **111**, 155–159.
- [72] Guha Mazumder, D., Haque, R., Ghosh, N., De, B.K., Santra, A., Chakraborty, D. & Smith, A.H. (1998). Arsenic levels in drinking water and the prevalence of skin lesions in West Bengal, India, *International Journal of Epidemiology* **27**, 871–877.
- [73] US EPA (2006). *Inorganic Arsenic (CASRN 7440-38-2)*, Integrated Risk Information System (IRIS), <http://www.epa.gov/iris/subst/0278.htm>.
- [74] International Agency for Research on Cancer (IARC) (2004). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Some Drinking-water Disinfectants and Contaminants*, World Health Organization (WHO), Including Arsenic, Vol. 84.
- [75] Chen, C.W., Chen, C.J., Wu, M.M. & Kuo, T.L. (1992). Cancer potential in liver, lung, bladder and kidney due to ingested inorganic arsenic in drinking water, *British Journal of Cancer* **66**, 888–892.
- [76] Guo, H.R., Yu, H.S., Hu, H. & Monson, R.R. (2001). Arsenic in drinking water and skin cancers: cell-type specificity (Taiwan, R.O.C.), *Cancer Causes Control* **12**(10), 909–916.
- [77] Guo, H.R. (2004). Arsenic level in drinking water and mortality of lung cancer (Taiwan), *Cancer Causes Control* **15**, 171–177.
- [78] Chiou, H.Y., Hsueh, Y.M., Liaw, K.F., Horng, S.F., Chiang, M.H., Pu, Y.S., Lin, J.S., Huang, C.H. & Chen, C.J. (1995). Incidence of internal cancers and ingested inorganic arsenic: a seven-year follow-up study in Taiwan, *Cancer Research* **55**, 1296–1300.
- [79] Tucker, S.P., Lamm, S.H., Li, F.X., Wilson, R., Li, F., Byrd, D.M., Lai, S., Tong, Y., Loo, L., Zhao, H.X., Zhendong, L. & Polkanov, M. (2001). *Relationship between Consumption of Arsenic-contaminated Well Water and Skin Disorders in Huhhot, Inner Mongolia*, <http://phys4.harvard.edu/~wilson/arsenic/references/imcap/IMCAP-report.html> (accessed Jul 2001).
- [80] Moore, L.E., Lu, M. & Smith, A.H. (2002). Childhood cancer incidence and arsenic exposure in drinking water in Nevada, *Archives of Environment Health* **57**, 201–206.
- [81] Chen, C.J., Wu, M.M., Lee, S.S., Wang, J.D., Cheng, S.H. & Wu, H.Y. (1988). Atherogenicity and carcinogenicity of high-arsenic artesian well water, *Arteriosclerosis* **8**(5), 452–460.
- [82] Chen, C.J., Chuang, Y.C., Lin, T.M. & Wu, H.Y. (1985). Malignant neoplasms among residents of a blackfoot disease-endemic area in Taiwan: high-arsenic artesian well water and cancers, *Cancer Research* **45**, 5895–5899.
- [83] Bates, M.N., Rey, O.A., Biggs, M.L., Hopenhayn, C., Moore, L.E., Kalman, D., Steinmaus, C. & Smith, A.H. (2004). Case–control study of bladder cancer and exposure to arsenic in Argentina, *American Journal of Epidemiology* **159**(4), 381–389.
- [84] Hopenhayn-Rich, C., Biggs, M.L., Fuchs, A., Bergoglio, R., Tello, E.E., Nicolli, H. & Smith, A.H. (1996). Bladder cancer mortality associated with arsenic in drinking water in Argentina, *Epidemiology* **7**, 117–124.
- [85] Lamm, S.H., Byrd, D.M., Kruse, M.B., Feinleib, M. & Lai, S. (2003). Bladder cancer and arsenic exposure: differences in the two populations enrolled in a study in Southwest Taiwan, *Biomedical and Environmental Sciences* **16**, 355–368.
- [86] Lamm, S.H., Engel, A., Penn, C.A., Chen, R. & Feinleib, M. (2006). Arsenic cancer risk confounder in SW Taiwan dataset, *Environmental Health Perspectives* **114**, 1077–1082.
- [87] Wall, S. (1980). Survival and mortality pattern among Swedish smelter workers, *International Journal of Epidemiology* **9**(1), 73–87.
- [88] Enterline, P.E., Marsh, G.M., Esmen, N.A., Henderson, V.L., Callahan, C. & Paik, M. (1987). Some effects of cigarette smoking, arsenic, and SO₂ on mortality among US copper smelter workers, *Journal of Occupational Medicine* **29**(10), 831–838.
- [89] Ott, M.G., Holder, B.B. & Gordon, H.L. (1974). Respiratory cancer and occupational exposure to arsenicals, *Archives of Environment Health* **29**, 250–255.
- [90] Hsueh, Y.M., Chiou, H.Y., Huang, Y.L., Wu, W.L., Huang, C.C., Yang, M.H., Lue, L.C., Chen, G.S. & Chen, C.J. (1997). Serum B-carotene level, arsenic methylation capability, and incidence of skin cancer, *Cancer Epidemiology Biomarkers and Prevention* **6**(8), 589–596.
- [91] Mitra, S.R., Guha Mazumder, D.N., Basu, A., Block, G., Haque, R., Samanta, S., Ghosh, N., Smith, M.M.H., von Ehrenstein, O.S. & Smith, A.H. (2004). Nutritional factors and susceptibility to arsenic-caused skin lesions in West Bengal, India, *Environmental Health Perspectives* **112**(10), 1104–1109.
- [92] Wong, O., Whorton, M.D., Foliart, D.E. & Lowengart, R. (1992). An ecologic study of skin cancer and environmental arsenic exposure, *International*

- Archives of Occupational and Environmental Health* **64**, 235–241.
- [93] Valberg, P.A., Beck, B.D., Bowers, T.S., Keating, J.L., Bergstrom, P.D. & Boardman, P.D. (1997). Issues in setting health-based cleanup levels for arsenic in soil, *Regulatory Toxicology and Pharmacology* **26**, 219–229.
- [94] Gebel, T.W., Suchenwirth, R.H.R., Boltén, C. & Dunkelberg, H.H. (1998). Human biomonitoring of arsenic and antimony in case of an elevated geogenic exposure, *Environmental Health Perspectives* **106**(1), 33–39.
- [95] Tollestrup, K., Frost, F.J., Harter, L.C. & McMillan, G.P. (2003). Mortality among children residing near the American Smelting and Refining Company (ASARCO) copper smelter in Ruston, Washington, *Archives of Environment Health* **58**(11), 683–691.
- [96] Hewitt, D.J., Millner, G.C., Nye, A.C. & Simmons, H.F. (1995). Investigation of arsenic exposure from soil at a superfund site, *Environmental Research* **68**, 73–81.
- [97] Hughes, M.F. (2006). Biomarkers of exposure: a case study with inorganic arsenic, *Environmental Health Perspectives* **114**, 1790–1796.
- [98] Vahter, M. (1983). Metabolism of arsenic, in *Biological and Environmental Effects of Arsenic*, B.A. Fowler, ed, Elsevier Science, pp. 171–197.
- [99] Schoen, A., Beck, B., Sharma, R. & Dubé, E. (2004). Arsenic toxicity at low doses: epidemiological and mode of action considerations, *Toxicology and Applied Pharmacology* **198**, 253–267.
- [100] US EPA (2001). National primary drinking water regulations; arsenic and clarifications to compliance and new source contaminants monitoring (Final rule), *Federal Register* **66**, 6975–7066.
- [101] US EPA (2003). *Preliminary Risk Assessment for the Reregistration Eligibility Decision on CCA*, Office of Prevention, Pesticides and Toxic Substances.
- [102] US EPA (2003). *A Probabilistic Risk Assessment for Children who Contact CCA-Treated Playsets and Decks (Draft)*.
- [103] US EPA (2006). *Reregistration Eligibility Decision for MSMA, DMSA, CAMA, and Cacodylic Acid*, List B. Case Nos. 2395, 2080, Office of Prevention, Pesticides and Toxic Substances, EPA-738-R-06-021.
- [104] US EPA (2005). *Toxicological Review of Ingested Inorganic Arsenic*, p. 61.
- [105] Brown, K.G. & Ross, G.L. (2002). Arsenic, drinking water, and health: a position paper of the American Council on science and health, *Regulatory Toxicology and Pharmacology* **36**(2), 162–174.
- [106] Steinmaus, C., Yuan, Y., Kalman, D., Attallah, R. & Smith, A.H. (2005). Intraindividual variability in arsenic methylation in a U.S. population, *Cancer Epidemiology Biomarkers and Prevention* **14**(4), 919–924.
- [107] Meza, M.M., Yu, L., Rodriguez, Y.Y., Guild, M., Thompson, D., Gandolfi, A.J. & Klimecki, W.T. (2005). Developmentally restricted genetic determinants of human arsenic metabolism: association between urinary methylated arsenic and CYT19 polymorphisms in children, *Environmental Health Perspectives* **113**(6), 775–781.
- [108] Hsueh, Y.M., Huang, Y.L., Wu, W.L., Huang, C.C., Yang, M.H., Chen, G.S. & Chen, C.J. (1995). Serum B-carotene level, arsenic methylation capability and risk of skin cancer, *SEGH Second International Conference on Arsenic Exposure and Health Effects, Book of Posters*, San Diego, June 12–14, 1995.
- [109] Rossman, T.G. (2003). Mechanism of arsenic carcinogenesis: an integrated approach, *Mutation Research* **533**, 37–65.
- [110] Brown, C.C. & Chu, K.C. (1983). A new method for the analysis of cohort studies: implications of the multistage theory of carcinogenesis applied to occupational arsenic exposure, *Environmental Health Perspectives* **50**(16), 293.
- [111] Brown, C.C. & Chu, K.C. (1983). Implications of the multistage theory of carcinogenesis applied to occupational arsenic exposure, *Journal of the National Cancer Institute* **70**(3), 455–463.
- [112] Brown, C.C. & Chu, K.C. (1983). Approaches to epidemiologic analysis of prospective and retrospective studies: example of lung cancer and exposure to arsenic, in *Environmental Epidemiology: Risk Assessment, Proceedings of the SIMS Conference on Environmental Epidemiology*, June 28–July 2, 1982, Alta, R.L. Prentice & A.S. Whittemore, eds, SIAM Publications, Philadelphia, pp. 94–106.
- [113] World Health Organization (WHO) (2006). *Guidelines for Drinking-water Quality*, First Addendum to Third Edition, Document available online at http://www.who.int/water_sanitation_health/dwq/gdwq0506.pdf, Geneva.

Related Articles

Environmental Hazard

Environmental Health Risk

Environmental Risk Assessment of Water Pollution

What are Hazardous Materials?

TRACEY M. SLAYTON, ARI S. LEWIS
AND BARBARA D. BECK

As Low as Reasonably Practicable/As Low as Reasonably Achievable

The *ALARP principle* refers to the notion that risks should be “as low as reasonably practicable”. ALARA, “as low as reasonably achievable”, is generally taken to mean the same, but is used more frequently in the context of radiation protection (*see Risk from Ionizing Radiation*). Certain pieces of legislation, such as the UK Health and Safety at Work Act refer to risks being reduced “so far as is reasonably practicable” (SFAIRP).

The ALARP principle is embedded in UK law, and is applied mainly in the United Kingdom, but has influenced risk acceptance criteria and risk regulation throughout the world. Other important principles include the precautionary principle used in various contexts, and the emerging concept of risk-informed regulation used by the US Nuclear Regulatory Commission.

Much of the legal framework surrounding the ALARP principle is based on the Court of Appeal judgment in the case of *Edwards v the National Coal Board* from 1948:

... in every case, it is the risk that has to be weighed against the measures necessary to eliminate the risk. The greater the risk, no doubt, the less will be the weight to be given to the factor of cost.

and

Reasonably practicable’ is a narrower term than ‘physically possible’ and seems to me to imply that a computation must be made by the owner in which the quantum of risk is placed on one scale and the sacrifice involved in the measures necessary for averting the risk (whether in money, time or trouble) is placed in the other, and that, if it be shown that there is a gross disproportion between them—the risk being insignificant in relation to the sacrifice—the defendants discharge the onus on them.

This brings out the essential elements of ALARP. On the one hand there is a risk, and against that there is a sacrifice required to avert that risk. Only when there is a great disproportion of sacrifice against risk, the duty holder (normally the management of a company carried out activities bearing a degree of risk) does not have to take any steps to reduce that risk.

Modeling an ALARP problem involves, however, making decisions about the bounds of the model: the population at risk, the items to be costed, etc. Studies usually consider hypothetical individuals and different kinds of populations, to ensure that there is no particular subpopulation at more risk than others. Costing is no less difficult, but guidance exists from regulators about what can and cannot be taken into account.

The government regulator (in the United Kingdom, normally the Health and Safety Executive, HSE) is responsible for testing that risk is ALARP, but final responsibility for a particular installation lies with the duty holder, and whether the law has been broken is ultimately a matter for courts rather than the regulator to decide.

While the ALARP principle has been adopted into law, there is little case law to provide guidance and therefore the regulatory authorities have put a great deal of effort into explaining how they interpret the elements of ALARP, that is, how risk is measured, how the sacrifice is to be measured, and what is meant by gross disproportion. These are only the views of the regulator, largely untested in court.

The ALARP principle is applied in the United Kingdom in many different areas. The original applications are in factories, mines, etc., but modern applications are also in transportation (trains, vehicles, tunnels, and airplanes), environment (flood, fire, etc.), medicine, and even the military. The principle has been incorporated into national and international standards, so that it is being applied even in areas where national law does not explicitly use it.

Measuring Risk

In order to balance the changes in risk with the costs of achieving those changes, we should be able to measure risks and costs, and then make them comparable. A later section will discuss the use of cost–benefit analysis, but for now we concentrate on measuring risks.

An extended discussion on risk measurement is given in [1]. However, for the purposes of this paper we can restrict the discussion broadly to simple individual risk measures and societal risk measures.

Individual risk is commonly used to refer to the probability that a specific individual will die as a consequence of the activity under discussion, at some

point in a given 1-year period. Many assumptions have to be made in order to model this risk, one of the most important being the exposure time, the proximity to the risk, and the physical condition of the person involved. For example, calculations for members of the public might conservatively assume that the individual was always exposed, while calculations for workers might assume that exposure only takes place during working hours.

While *individual risk* refers to the risk for a specific individual, *societal risk* refers to the frequency with which events occur leading to multiple fatalities. A common way to illustrate this is with an FN (Frequency-Number) curve, which shows the frequency of events (per year) involving the death of at least N individuals, ($N = 1, 2, 3, \dots$).

Various other risk measures are used in different sectors. For example, in the nuclear industry it is also common to discuss the risk of exposure to a given level of radiation. As these may be difficult to assess, core damage frequencies are also used as risk measures.

Often risk measures are applied to different populations. While it is common to distinguish between fatality risks for the public and for the workers, the UK Railway Safety and Standards Board goes further, by distinguishing workers, passengers, and third-party members of the public. It further models minor injuries, major injuries, and fatalities giving nine categories in total. However, for the purpose of making ALARP judgments, these injury categories are considered as “equivalent fatalities” by using weighting factors for each category.

Finally, in some industries it is appropriate to use risk measures that are normalized by exposure time. An example of this is the fatal accident rate (FAR) used in the maritime and offshore industries. This measures accidents per hour (or million hours) of working hours.

Tolerability of Risk

As a result of the Sizewell B Enquiry in the United Kingdom, it was recommended that HSE develops guidance on “tolerable” levels of risk to members of the public and to workers from nuclear power plants. The use of the word “tolerable” was quite new, as the common usage then was to talk about “acceptability” of risk. The subtle distinction is used to indicate that

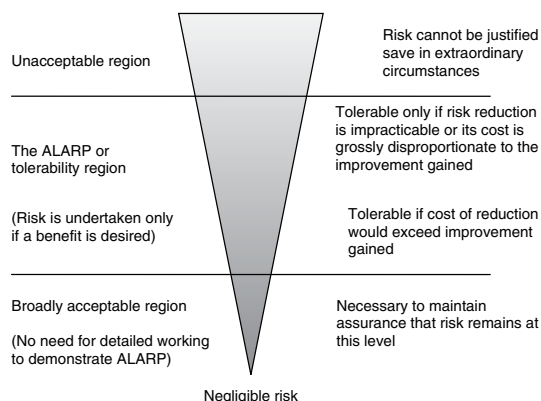


Figure 1 Tolerability of risk triangle [From [2], 1988. Health and Safety Executive. © Crown copyright material is reproduced with the permission of the Controller of HMSO and Que Printer for Scotland.]

we never accept a level of risk, but tolerate it in return for other benefits that we obtain through the activity that generates the risk.

The tolerability of risk (TOR) guidance document [2] produced by HSE in 1988 (revised in 1992) introduced the TOR triangle shown in Figure 1.

Activities that generate a high level of risk are considered intolerable under all circumstances. At the lower end, there is a level below which risk is considered to be negligible, and therefore so low that it is not worth the cost of actively managing it. There is a region between these two limits, called the *ALARP* or the *tolerability region*, in which it is necessary to make a trade-off between risk and the cost of further risk reduction.

In keeping with the Court of Appeal judgment in the *Edwards v The National Coal Board* case, the issue of whether or not a particular risk is tolerable in the ALARP region depends on whether or not the cost of reducing it *further* would be disproportionate to the risk. Furthermore, the degree of disproportionality depends on how big the risk is.

The TOR guidance suggested an upper limit of individual risk for workers as 1 in 1000 per year, and 1 in 10 000 for the public for nonnuclear industrial risks, and 1 in 100 000 for the public for nuclear risks.

In safety assessment principles (SAPs), the upper and lower risk limits are called *basic safety limit* (BSL) and *basic safety objective* (BSO). A number of different measurement scales are used, so that BSLs and BSOs are defined for diverse quantities such as annual radiation exposure of a member of the public,

worker radiation dose, plant damage frequency, and critical event frequency. Hence, in principle at least, there are a number of different measurement scales on which the ALARP principle would be applied.

In order to determine the level of risk, a probabilistic safety analysis (PSA) is used (in nonnuclear industries the term *quantitative risk analysis* (QRA) is used more frequently). It should be noted however that, in general, good engineering practice is more of a driver of actual design and operating safety decisions than PSA.

The HSE's *Reducing Risks, Protecting People* [3] describes the ALARP region as follows:

The zone between the unacceptable and broadly acceptable regions is the tolerable region.

Risks in that region are typical of the risks from activities that people are prepared to tolerate in order to secure benefits, in the expectation that:

- the nature and level of the risks are properly assessed and the results used properly to determine control measures. The assessment of the risks needs to be based on the best available scientific evidence and, where evidence is lacking, on the best available scientific advice;
- the residual risks are not unduly high and kept as low as reasonably practicable (the ALARP principle. . .); and
- the risks are periodically reviewed to ensure that they still meet the ALARP criteria, for example, by ascertaining whether further or new control measures need to be introduced to take into account changes over time, such as new knowledge about the risk or the availability of new techniques for reducing or eliminating risks.

While the ALARP region is generally taken to apply to individual risk measures, something very similar is used for societal risk measures. Indeed, several governments have included FN criteria in legislation or risk regulation, among others being The Netherlands, which had upper and lower limits similar to the BSO and BSL used by the SAPs for individual risk [4].

Measuring “The Sacrifice” Required to Reduce Risk

As we have seen, the original court judgment introducing the notion that we now call ALARP discussed the need to balance risk with the sacrifice required to reduce that risk.

Current practice is to measure “sacrifice” in monetary terms, using cost–benefit analysis. There are three main difficulties in this regard:

1. defining the scope of the analysis, in terms of time, geography, scale, etc;
2. quantifying monetary values for items that are not normally traded commodities, such as human lives, plant and animal species, and so on;
3. dealing costs and benefits accrued at different points of time.

Analysis scope is very difficult to determine because of the wide range of possible knock-on effects that might result from a change to a system. In some cases the scope is simply defined by a regulation or law, as in the Dutch definition of societal risk that includes only short-term fatalities. In other cases, the scope is not always clear and might affect the outcome of the analysis – for example, if one considers the idea of introducing seat belts in trains then this restricts capacity and therefore increases road-use and associated environmental costs.

The quantification of monetary values for items that are not traded is the subject of a large literature within the area of cost–benefit analysis (*see Cost-Effectiveness Analysis*). There has been much criticism of this, in particular; see [5]. In the context of this short article, we shall concentrate on the concept of the value of a statistical life, often also called the *value of preventing a fatality* (VPF).

While the layperson's view of VPF is often that it represents an actual valuation of a human life, with all the unpleasant moral associations that it brings, this is not the view taken by most professionals. Indeed, the UK HSE [3] states as follows:

VPF is often misunderstood to mean that a value is being placed on a life. This is not the case. It is simply another way of saying what people are prepared to pay to secure a certain averaged risk reduction. A VPF of £1 000 000 corresponds to a reduction in risk of one in a hundred thousand being worth about £10 to an average individual. VPF therefore, is not to be confused with the value society, or the courts, might put on the life of a real person or the compensation appropriate to its loss.

This interpretation is of interest because it relates the definition of VPF to its use in risk decision making. In practice, one is not trying to decide whether or not a particular person (a “statistical person” does

not, of course, exist) should live or die, but whether a particular person (in fact the particular group of people exposed to risk) should have a change to that risk. For this society has to commit resources, and has to determine an appropriate amount.

The HSE view [3] is that the VPF should be the same for each type of risk, except for cancer, where the VPF is doubled. The VPF used in the United Kingdom is of the order of £ 1 million, and was originally developed by the UK Department of Transport. While consistency of the VPF is a reasonable administrative objective, it seems reasonable to suggest that in determining priorities for risk reduction (for example, through budget setting), some types of risk are prioritized over others – for example, multiple fatality accidents as opposed to suicides.

The VPF is typically determined through stated preference techniques, particularly contingent valuation and choice modeling. In the former, respondents have to state how much they would be willing to pay for a reduction in risk and this is converted to a VPF. Choice modeling is a more general class of techniques in which a VPF can be inferred from rankings of different choices offered to the participant. For a longer discussion of issues related to stated preference modeling, see [6, 7].

Within cost–benefit analysis, the standard way of comparing items from different time periods is to use discounting. Clearly, the way costs and benefits are discounted over time plays an important role in determining the relative overall balance. Reference [3] suggests using a standard real interest rate of 6% for costs and benefits, but then states the following:

“The value that individuals place on safety benefits tends to increase as living standards improve, so the future values applied to such benefits should be uprated to allow for the impact on well-being of expected growth in average real income. On the basis of past trends and Treasury guidance, HSE regards an uprating factor of 4% a year as appropriate on the benefits side of the comparison.”

Discounting has been criticized for use in studies with very long timescales; see, for example [8].

Criticisms of ALARP

ALARP is widely used in the United Kingdom, but has not been adopted consistently elsewhere, though many countries have regulatory frameworks that, to

some extent, resemble ALARP; see, for example, the discussion of Norwegian FAR regulations in [9].

There are a number of problems with the cost–benefit analysis element to ALARP. French *et al.* [10], following Adams [5], explore the problems with the difference between willingness to pay (WTP) and willingness to accept (WTA) evaluations. Essentially, the difference in the way a contingent valuation question is posed (WTP or WTA) substantially affects the values obtained. French *et al.* [10] argue that the usual starting point is that people have a right not to be exposed to risks, and that therefore a WTA formulation is appropriate, while in practice WTP is used. They also make a comparison to multiattribute decision making, and discuss problems with the interpretation of “gross disproportionality”.

Kletz [11] points out that in practice the amount of money used to save a human life varies by orders of magnitude between the health service, road safety, chemical industry, and the nuclear industry. He also points to problems arising from the bounding of ALARP analysis. For example, he says that the rebuilt Flixborough site (after the 1974 explosion) was designed to use a safer chemical ingredient than before the disaster. However, this had to be manufactured elsewhere using a process that was at least as risky as the process that had taken place at Flixborough previously. The risk had been exported, out of Flixborough and out of the ALARP calculation.

Examples of ALARP

An example of ALARP case is described in [12]. The authors describe work carried out during the design of the BP Andrew oil platform whose objective was to assess and reduce the risk of impairing the temporary refuge owing to gas explosion and/or ingress of gas or smoke. These had been assessed as the main drivers of risk to the temporary refuge—a unit built away from the main accommodation module on the platform to act as an emergency area. The case illustrated the way in which design changes can be found, which both reduce cost and reduce risk. This case was different from many other applications that take place during the upgrading of existing installations, but shows how beneficial it is to include ALARP considerations in the design phase. In this particular case, the team was able to find new design solutions that lowered risk and saved money.

Ingress of gas and smoke was modeled using computational fluid dynamics models and wind tunnel tests. This enabled the designers to find a position for the air intake pipe for the temporary refuge that was least susceptible to gas and smoke intake, and to estimate the probability of unacceptable ingress in the worst-case scenario. Gas explosions were modeled using a Flame Acceleration Simulator – a computer code that had been previously developed by an industry consortium to model the effect of equipment layout and wall arrangements on combustion and turbulence. Again, a worst-case scenario was computed. Hence, the derived quantitative risks are against those that are considered worst-case scenarios rather than averaged over all scenarios. This is common practice and gives some risk margin, which has enabled some industries to argue later that even when equipment is degraded in later life they are still within ALARP bounds.

Conclusions

The ALARP principle has been the guiding principle for UK risk regulation for several decades and has provided a good basis for consistency of decision making across different sectors. Consistency is achieved on the basis of a number of slightly arbitrary choices in risk limits, VPF, and so on. In practice, this consistency is sometimes undermined by higher level choices, such as budget setting, on the part of decision makers. Finally, the application of ALARP requires the setting of boundaries around the problem area, and the outcome of ALARP decisions may depend on these boundaries. Having said this, ALARP has proved itself in practice to be a reasonably robust approach capable of application across a wide range of sectors.

References

- [1] Bedford, T. & Cooke, R. (2001). *Probabilistic Risk Analysis: Foundations and Methods*, Cambridge University Press, Cambridge.

- [2] HSE (1988). *The Tolerability of Risk from Nuclear Power Stations*, HSE: HMSO, London.
- [3] HSE (2001). *Reducing Risks, Protecting People*, Ha, S., ed, HMSO.
- [4] Ball, D.J. & Floyd, P.J. (1998). *Societal Risks*, Health and Safety Executive.
- [5] Adams, J. (1995). *Risk*, UCL Press, London.
- [6] Beattie, J., Covey, J., Dolan, P., Hopkins, L., Jones-Lee, M., Loomes, G., Pidgeon, N., Robinson, A. & Spencer, A. (1998). On the contingent valuation of safety and the safety of contingent valuation: part I – caveat investigator, *Journal of Risk and Uncertainty* 17(1), 5–25.
- [7] Carthy, T., Chilton, S., Covey, D., Hopkins, L., Jones-Lee, M., Loomes, G., Pidgeon, N. & Spencer, A. (1998). On the contingent valuation of safety and the safety of contingent valuation: part 2 – the CV/SG “chained” approach, *Journal of Risk and Uncertainty* 17(3), 187–213.
- [8] Atherton, E. & French, S. (1998). Valuing the future: a MADA example involving nuclear waste storage, *Journal of Multi-Criteria Decision Analysis* 7(6), 304–321.
- [9] Aven, T. & Vinnem, J.E. (2005). On the use of risk acceptance criteria in the offshore oil and gas industry, *Reliability Engineering and System Safety* 90(1), 15–24.
- [10] French, S., Bedford, T. & Atherton, E. (2005). Supporting ALARP decision making by cost benefit analysis and multiattribute utility theory, *Journal of Risk Research* 8(3), 207–223.
- [11] Kletz, T.A. (2005). Looking beyond ALARP – Overcoming its limitations, *Process Safety and Environmental Protection* 83(B2), 81–84.
- [12] Tam, V., Moros, T., Webb, S., Allinson, J., Lee, R. & Bilimoria, E. (1996). Application of ALARP to the design of the BP Andrew platform against smoke and gas ingress and gas explosion, *Journal of Loss Prevention in the Process Industries* 9(5), 317–322.

Related Articles

Chernobyl Nuclear Disaster

Risk–Benefit Analysis for Environmental Applications

TIM BEDFORD

Asbestos

“Asbestos”, derived from the Greek word meaning “unquenchable” or “indestructible” is a general term applied to a family of fibrous hydrated magnesium silicates [1]. Use of asbestos dating back to ancient times has been documented by the Finnish (2500 B.C. in pottery) and the Romans (400–500 B.C. in lamp wicks and cremation cloths). It was not until the late 1800s, however, that asbestos was mined commercially. Since this time, owing to its strength, flexibility, and heat resistance, asbestos has been incorporated into over 3000 commonly used products, such as oven mitts, driveway sealer, and automotive brakes and clutches [2–7].

Mineralogic Features of Asbestos

Asbestos is composed of two distinct mineralogic groups, the amphiboles and the serpentines. The amphibole group consists of five minerals, including, crocidolite, amosite, anthophyllite, tremolite, and actinolite. Of these, only crocidolite and amosite, also known as blue and brown asbestos, respectively, were mined and used commercially. The remaining amphiboles are primarily considered contaminants of other minerals, such as chrysotile, vermiculite, and talc [8, 9]. Amphibole fibers are rhombic in shape, and consist of chain structures, with nine structural sites that accommodate cations. The principal cations are magnesium, iron, calcium, and sodium, and their ratios can be used to distinguish between the mineral species [10]. Amphibole crystals are characterized as double chain inosilicates, in which the two side-by-side chains share one oxygen molecule per pair of silica tetrahedral. Members of this family have very similar crystal structures, and therefore cannot be distinguished by electron diffraction [9].

The most well-known amphibole deposits, predominantly of crocidolite, are located in South Africa and Australia. The use of amphiboles was banned in most countries between the early 1970s and the mid-1980s, prior to which amphibole asbestos was widely used in asbestos-cement and various insulation products. In fact, from the 1940s to the 1960s, the US Navy often required the use of amosite asbestos pipe and boiler insulation on naval ships. Amosite was particularly suitable for naval use, owing to its resistance

to degradation by salt water. In general, during the years of war, amphiboles were the fiber type of choice for the products used by the navy due to their thermal conductivity, light weight, high tensile strength, and refractoriness [11].

The serpentine group, which is structurally and biologically different from that of the amphiboles, consists solely of chrysotile, or “white asbestos”. Unlike the amphiboles, chrysotile exists as a phyllosilicate, which wraps around itself in a spiral, forming a hollow tubular macrostructure [10]. Chrysotile typically exists in fiber bundles, and consists of varying quantities of individual fibrils. As the name “serpentine” suggests, this class of minerals are known for their snake-like, curly or wavy appearance. This form of asbestos has accounted for roughly 96% of the asbestos production and consumption worldwide from 1900 to 2003, and of the estimated 2.15 million tons of asbestos produced in 2003, it is likely that all but a few thousand tons were chrysotile [12, 13]. Mining of this mineral has mainly occurred in Canada, Russia, Zimbabwe, and the United States, with Canada dominating the industry for most of the twentieth century. Current worldwide chrysotile uses include gaskets, roofing and friction products, corrugated and flat sheets, fittings, and pipes [12].

Pathways of Exposure

By far, inhalation of asbestos fibers is the most important human health hazard. However, humans can be exposed to asbestos fibers *via* ingestion since they have also been detected in potable water supplies [14–19], predominantly at concentrations below 1 million fibers/l [20]. Several sources of the asbestos have been proposed, including the erosion of natural deposits or waste piles, rain water that has run off of asbestos-cement shingles, and leaching from asbestos-cement pipe [16, 21]. Despite the presence of fibers in drinking water, the weight of the scientific evidence indicates that asbestos ingestion does not cause any significant noncancer or carcinogenic health effects [21, 22].

Toxicology of Inhaled Asbestos

There are several factors that are believed to determine the toxicity of inhaled asbestos: fiber dimensions, durability, and chemistry/surface activity.

Fiber Dimensions

The aerodynamic diameter is defined as the diameter of a spherical particle with unit density and the same settling velocity as the actual particle. The aerodynamic diameter of a particle or fiber helps to characterize and define respirability and location of pulmonary deposition. It is believed that particles with aerodynamic diameters between 5 and 30 μm deposit in the nasopharyngeal region, while smaller particles ($\sim 1 \mu\text{m}$) can penetrate more deeply in to the alveolar region of the lungs. In contrast to short, thick fibers, long, thin fibers have small aerodynamic diameters, which allow them to move in parallel with the air stream and deposit deeper into the respiratory tract. Asbestos fibers that are deposited and retained in the alveolar region are most likely to induce a biologically relevant response.

Durability

The durability of an asbestos fiber will influence the rate of fiber clearance. Fibers are generally removed from the lungs *via* the following physiological processes: (a) mucociliary transport (stops at terminal bronchioles), (b) alveolar macrophage translocation, (c) uptake by epithelial cells lining the airways, and (d) lymphatic clearance [23]. In general, clearance from the upper respiratory tract occurs at a faster rate than clearance in the deep lung. For asbestos fibers that reach the alveolar region, the main determinant of the speed of clearance is the rate of phagocytosis. Fiber length is thought to be particularly influential in the pathogenesis of asbestos-related disease, as macrophage-mediated clearance is less effective for long fibers. Further, it has been suggested that fibers greater than the diameter of alveolar macrophages (10–15 μm in rodents, 14–21 μm in humans) are unlikely to be completely phagocytosed [24]. It is believed that this ineffective phagocytosis results in the release of reactive oxygen species and other inflammatory agents. These processes may be one of the causes for the carcinogenic response observed in some animal and human studies.

Numerous studies have supported the assertion that long fibers, regardless of fiber type are more potent than short fibers [25–29]. Recently, the US Environmental Protection Agency (EPA) convened a workshop to discuss a proposed protocol to assess asbestos-related risk [30]. The peer consultation panel concluded that there was considerably greater risk for

developing cancer with asbestos fibers longer than 10 μm and for fibers less than 5 μm the risk was very low and may even be zero [31]. The Agency for Toxic Substances and Disease Registry (ATSDR) also recently sponsored an expert panel to study the influence of fiber length on asbestos health effects [32]. The expert panel also agreed that fibers less than 5 μm pose a negligible risk of cancer [33].

In addition, it has been shown that chrysotile fibers, unlike the amphiboles, undergo longitudinal (splitting) and transverse breaking [23, 34–39]. The effects of the splitting or breaking of chrysotile fibers is unclear. Although some argue that this process renders the chrysotile fibers more apt to macrophage ingestion, others indicate that this may aid in the translocation to the pleural or peritoneal cavities, and the bronchial-associated lymphoid tissue [21, 23].

Chemistry/Surface Activity

Asbestos fibers that are not effectively cleared by physicochemical processes may be removed by other processes, such as dissolution. The physicochemical properties of the different fiber types determine the rate at which they undergo dissolution. Typically, experimentally derived dissolution rates for amphiboles are between 10^{-12} and $10^{-10} \text{ mol m}^{-2} \text{ s}^{-1}$. In contrast, at 37 °C and under conditions similar to the physiological environment in the lung, a dissolution rate of $5.9 \times 10^{-10} \text{ mol m}^{-2} \text{ s}^{-1}$ has been observed for chrysotile asbestos [40]. The practical implications of such dissolution are that an amphibole fiber will not dissolve in a human lung over the course of a lifetime, while a chrysotile fiber, even as thick as 1 μm , would be expected to dissolve in the human lung in less than 1 year [24].

The differing rates of dissolution are the result of the orientation of the magnesium molecules within the fiber types. While magnesium is present to varying degrees in all types of asbestos (e.g., 33% by weight in chrysotile and 6–25% by weight in amphiboles), in chrysotile, the magnesium is located on the outside of the macrostructure, rendering it accessible to lung fluids. On the contrary, the magnesium molecules in the amphiboles are locked within the I-beam type structure, which consists of corner-linked $(\text{SiO}_4)_4$ tetrahedra linked together in a double tetrahedral chain that sandwiches a layer with the $\text{Ca}_2 \text{Mg}_5$ [23]. Numerous studies have demonstrated that, upon exposure to mildly acidic solutions, or even water, the

magnesia completely dissociates from the chrysotile fiber [41–44]. Upon dissociation of the magnesia, the dimensional stability of the chrysotile fibril is lost [45].

The variability in iron composition has also been hypothesized to account for the observed difference in the cancer potency of the fiber types. The amphiboles, such as amosite and crocidolite, are composed of between 25 and 36% iron by weight [46]. This is in contrast to 1–5% iron by weight in chrysotile [47]. In general, it is supposed that fibers with more iron associated with them are better free radical generators, and thus are more biologically active [48, 49]. The reactive oxygen species are believed to be the result of a Fenton-type (Haber–Weiss) redox cycle, which yields hydroxyl radicals. Although the exact mechanism of fibrosis and tumorigenesis has not been elucidated, it has been proposed that the iron content of fibers has a positive correlation with the activator protein-1 (AP-1) induction [50]. In the presence of iron, AP-1, a redox sensitive transcription factor, induces uncontrolled cell proliferation.

Studies of Health Effects

During the first 30 years of the twentieth century, there were only a few case reports that suggested that exposure to asbestos might be associated with adverse health effects. However, these early case reports provided little, if any, information regarding the specific activities that resulted in exposure, the concentration of airborne particles, or details about the disease in these workers [51]. In 1930, the first epidemiology study confirmed an increased incidence of asbestosis in employees working in dusty factory settings [52]. In the subsequent two decades, with few exceptions, the focus of the more than 30 occupational health studies conducted was on workers performing activities in the dustiest of manufacturing environments [11, 53–83].

The first study conducted on end users of asbestos-containing products was not published until 1946, and reported three cases of asbestosis in 1074 pipe coverers involved in the following operations: (a) laying out and cutting of insulating felt, (b) band sawing insulation blocks and boards, (c) manufacturing of boots and jackets to cover valves and joints, (d) mixing asbestos-cement, (e) molding of block insulation, (f) grinding of scrap material, and (g) application of insulation [11]. Owing to the existing US

Naval specifications, amosite was the major ingredient in insulation material, and comprised up to 94% of the asbestos used in pipe covering. Consequently, owing to the low incidence of asbestosis observed, and the fact that all cases had been employed in shipyards for 20 or more years, the authors reported “it may be concluded that such pipe covering is not a dangerous occupation” [11, p. 16]. In 1949, Dr. Canepa conducted a study on 52 “insulation installers” at the Port of Genoa, and reported 5 cases of “clear and obvious” asbestosis and 10 cases of early asbestosis. This was the first study on asbestos-related health effects focused on end users that suggested an increased risk of asbestosis [70].

In 1951, Gloyne examined 1247 lung sections for the presence of pulmonary diseases attributable to the inhalation of dust, and reported lung cancer in 19.6 and 9.7% of men and women with asbestosis, respectively [84]. This raised the question of the possible link between occupational exposures to asbestos and lung cancer. Soon after, [74] published the first epidemiological study providing solid evidence that occupation and smoking were factors in the development of lung cancer [74]. The authors examined incidence of lung cancer in regard to occupation, and reported that lung cancer patients were 10 times more likely to have worked for at least 5 years in occupations involving the use asbestos than the controls. However, the authors concluded that “the group of steam fitters, boilermakers, and asbestos workers lies on the borderline of statistical significance when the effect of cigarette smoking is controlled” [74]. Thus, at this point, it was unclear whether asbestos exposure alone could cause lung cancer.

In the early 1960s, significant progress was made in regard to health outcomes associated with asbestos exposures. Two pivotal studies were published that highlighted the potential for asbestos-related disease among individuals with relatively low exposures [85, 86]. Wagner *et al.* reported 33 cases of mesothelioma identified among South Africans occupationally exposed to asbestos, as well as those residing in proximity to a crocidolite mine. Mesothelioma was also observed in South Africans with no known occupational exposure to asbestos [86]. Soon after, Dr. Irving Selikoff reported an increased incidence of lung and pleural cancer following a cohort mortality study of 632 insulation workers [87, 88]. This was

the first epidemiology study of persons using a “finished product”, which showed a significant association between asbestos exposure and lung cancer and mesothelioma. Because asbestos was a component of many commercial products, such as automotive friction products (i.e., brakes and clutches), cement pipe, and even oven mitts, questions were soon raised regarding the health hazards of asbestos exposures from end products.

Following the shift in attention from dusty factory workers to end users, numerous occupational cohorts (see **Occupational Cohort Studies**) (e.g., asbestos sprayers, cigarette filter workers, construction workers, shipyard workers, electrochemical plant workers, garage mechanics, locomotive drivers, railroad maintenance workers, and rock salt workers) were the subjects of epidemiology studies [89–97]. As the cumulative number of publications on asbestos rose from about 150 in 1940 to over 10 000 in 2000, increased attention was paid to quantifying historical exposure levels. In addition, the specific characteristics of asbestos fibers in asbestos-exposed

workers were examined in several studies [98, 99]. The results of these studies revealed that many of the workers exposed to serpentine fibers had a significant number of amphibole fibers retained in the lungs, while the chrysotile fibers were cleared from the lungs [51]. In addition, chrysotile miners were found to have developed fewer lung cancers and mesotheliomas when compared to miners exposed to other types of asbestos [100]. Therefore, it was postulated that amphiboles were a more potent asbestos type than chrysotile [101, 102].

A number of researchers have quantified the differences in risk of asbestos-related disease between comparable levels of exposure to airborne chrysotile and amphiboles. For example, Hodgson and Darn-ton [103] concluded that “the exposure specific risk of mesothelioma from the three principal commercial asbestos types is broadly in the ratio 1 : 100 : 500 for chrysotile, amosite and crocidolite, respectively”. This magnitude is similar to that agreed upon by a recently convened panel of experts who issued a report on asbestos-related risk for the EPA. This

Table 1 US occupational asbestos guidelines and regulations

Decade	Year	Agency	Duration	Occupational guideline or standard for asbestos ^(a)
1940s	1946	ACGIH	MAC ^(b) /TLV ^(c) 8-h TWA	5 mppcf
1970s	1971	OSHA	PEL ^(d) 8-h TWA	12 fibers/cc
			PEL: 8-h TWA	5 fibers/cc
			STEL ^(e) 15 min	10 fibers/cc
	1974	ACGIH	TLV: 8-h TWA ^(f)	5 fibers/cc
	1976	OSHA	PEL: 8-h TWA	2 fibers/cc
1980s	1980	ACGIH	STEL: 15 min	10 fibers/cc
			TLV: 8-h TWA	Amosite 0.5 fibers/cc
				Chrysotile 2 fibers/cc
				Crocidolite 0.2 fibers/cc
				Tremolite 0.5 fibers/cc
	1986	OSHA	PEL: 8-h TWA	Other forms 2 fibers/cc
				0.2 fibers/cc
				Short-term exposure limit removed
	1988	OSHA	PEL: 8-h TWA	0.2 fibers/cc
				STEL: 30 min
1990s	1994	OSHA	PEL: 8-h TWA	1.0 fibers/cc
			STEL: 30 min	0.1 fibers/cc
	1998	ACGIH	TLV: 8-h TWA	1.0 fibers/cc
				0.1 fibers/cc (all fiber types)

^(a) Beginning in 1971, regulations were specific to fibers >5 µm in length with an aspect ratio of 3 : 1

^(b) MAC: Maximum Allowable Concentration

^(c) TLV: Threshold Limit Value

^(d) PEL: Permissible Exposure Limit

^(e) STEL: Short Term Exposure Limit

^(f) TWA: Time Weighted Average

report states, “The panelists unanimously agreed that the available epidemiology studies provide compelling evidence that the carcinogenic potency of amphibole fibers is two orders of magnitude greater than that for chrysotile fibers” [104]. Likewise, it has been noted that the risk differential between chrysotile and the two amphibole fibers for lung cancer is between 1 : 10 and 1 : 50 [103].

Regulations

The first guidance level for asbestos in workroom air was proposed by Dreessen *et al.* [62] following a study of asbestos workers in manufacturing settings, in which asbestosis was not observed in individuals with exposures below 5 million particles per cubic foot (mppcf) [62]. In 1946, the American Conference of Governmental Industrial Hygienists (ACGIH) adopted this recommendation as the

Threshold Limit Value (TLV) for occupational exposure to asbestos [105]. In the early 1960s, as the result of the Walsh–Healy Act and the Longshoreman’s Act, 5 mppcf became an enforceable regulatory limit for specific industries. Later that decade the ACGIH issued a Notice of Intended Change, lowering the asbestos guideline from 5 mppcf to 12 fibers/ml as an 8-h time weighted average (TWA) [106]. With the conception of the Occupational Safety and Health Administration (OSHA), the first legally enforceable health standard governing exposure to asbestos in all industries was issued on May 29, 1971. A depiction of the occupational exposure limits set forth by OSHA and the ACGIH is presented in Table 1 [107–115]. Current regulations and guidelines governing exposure to asbestos in the United States and in several other countries are presented in Table 2. For policy rather than scientific reasons, most government agencies have chosen to treat all asbestos

Table 2 Current international asbestos regulations

Country	Agency	Long term occupational exposure limit ^(a) (fibers/cc)	Fiber type
Australia	National Occupational Health and Safety Commission	0.1	All types
Canada (Quebec)	National Public Health Institute of Quebec	1	Actinolite Anthophyllite Chrysotile Tremolite
		0.2	Amosite Crocidolite
Denmark	Danish Working Environment Authority	0.1	All types
France	Ministry of the Employment, Social Cohesion and Lodging	0.1	All types
Germany	German Committee on Hazardous Substances	0.15	All types
Hungary	Ministry of Social Affairs and Health	0.1	All types
Japan	Japan Society for Occupational Health	0.15	Chrysotile
		0.03	All other types
South Africa	Department of Labour/Department of Minerals and Energy	0.2	All forms
Spain	National Institute for Occupational Safety and Health	0.1	All types
Switzerland	Swiss Commission of Occupational Exposure Limit Values	0.01	All types
United Kingdom	Health and Safety Commission	0.01	All types
United States	Occupational Safety and Health Administration	0.01	All types
European Union	European Commission/Scientific Committee for Occupational Exposure Limits to Chemical Agents	0.1	All types

^(a) All long term exposure limits are based on an 8-h TWA, excluding France (1-h reference period) and the United Kingdom (4-h reference period)

fiber types equally, despite evidence of large differences in the potential toxicity between amphibole and chrysotile asbestos.

References

- [1] Lee, D.H.K. & Selikoff, I.J. (1979). Historical background to the asbestos problem, *Environmental Research* **18**, 300–314.
- [2] U.S. Environmental Protection Agency (EPA) (1985). *Asbestos Waste Management Guidance: Generation, Transport, Disposal*, EPA/530-SW-85-007, United States Environmental Protection Agency (USEPA), Office of Solid Waste, Washington, DC.
- [3] U.S. Environmental Protection Agency (EPA) (1990). *Managing Asbestos In Place: A Building Owner's Guide To Operations And Maintenance Programs For Asbestos-Containing Materials*, Office of Pesticides and Toxic Substances, 20T-2003, Washington, DC.
- [4] Summers, A.L. (1919). *Asbestos And The Asbestos Industry*, Sir Isaac Pitman & Sons, London.
- [5] Asbestos Institute (AI) (2000). *Chrysotile Products: Regulation*, www.chrysotile.com/en/products.htm (accessed Jun 29, 2004).
- [6] Michaels, L. & Chissick, S.S. (1979). *Asbestos: Properties, Applications, and Hazards*, John Wiley & Sons, Chichester, Vol. 1, pp. 74–75.
- [7] Craighead, J.E. & Mossman, B.T. (1982). The pathogenesis of asbestos-associated diseases, *New England Journal of Medicine* **306**, 1446–1455.
- [8] Addison, J. & Davies, L.S.T. (1990). Analysis of amphibole asbestos in chrysotile and other minerals, *Annals of Occupational Hygiene* **34**, 159–175.
- [9] Roggli, V.L. & Coin, P. (2004). Chapter 1: Mineralogy of asbestos, in *Pathology of Asbestos-Associated Diseases*, 2nd Edition, V.L. Roggli, T.D. Oury & T.A. Sporn, eds, Springer-Verlag, New York, pp. 1–16.
- [10] National Toxicology Program (NTP) (2005). *Asbestos Report on Carcinogens*, 11th Edition, U.S. Department of Health and Human Services, Public Health Service, National Toxicology Program, Washington, DC.
- [11] Fleischer, W.E., Viles, F.J., Gade, R.L. & Drinker, P. (1946). A health survey of pipe covering operations in constructing naval vessels, *The Journal of Industrial Hygiene and Toxicology* **28**, 9–16.
- [12] Virta, R.L. (2005). *Mineral Commodity Profiles – Asbestos*, US Geological Survey (USGS) Circular 1255-KK.
- [13] Moore, P. (2004). Chrysotile in crisis, *Industrial Minerals* **439**, 56–61.
- [14] Cook, P.M., Glass, G.E. & Tucker, J.H. (1974). Asbestiform amphibole minerals: detection and measurement of high concentrations in municipal water supplies, *Science* **185**, 853–855.
- [15] Nicholson, W.J. (1974). Analysis of amphibole asbestiform fibers in municipal water supplies, *Environmental Health Perspectives* **9**, 165–172.
- [16] Millette, J.R., Clark, P.J., Pansing, M.F. & Twyman, J.D. (1980). Concentration and size of asbestos in water supplies, *Environmental Health Perspectives* **34**, 13–25.
- [17] Meigs, J.W., Walter, S., Heston, J.F., Millette, J.R., Craun, G.F. & Flannery, J.T. (1980). Asbestos cement pipe and cancer in Connecticut 1955–1974, *Environmental Research* **42**, 187–197.
- [18] Toft, P., Wigle, D., Meranger, J.C. & Mao, Y. (1981). Asbestos and drinking water in Canada, *The Science of the Total Environment* **18**, 77–89.
- [19] McGuire, M.J., Bowers, A.E. & Bowers, D.A. (1982). Asbestos analysis case history: surface water supplies in Southern California, *Journal of American Water Works Association* **74**, 470–477.
- [20] DHHS Committee to Coordinate Environmental and Related Programs (1987). Report on cancer risks associated with the ingestion of asbestos, *Environmental Health Perspectives* **72**, 253–265.
- [21] Agency for Toxic Substances and Disease Registry (ATSDR) (2001). *Toxicological Profile for Asbestos*, US Department of Health and Human Services (DHHS), Public Health Service, Agency for Toxic Substances and Disease Registry (ATSDR), Washington, DC.
- [22] Condie, L.W. (1983). Review of published studies of orally administered asbestos, *Environmental Health Perspectives* **53**, 3–9.
- [23] Bernstein, D.M., Rogers, R. & Smith, P. (2005). The biopersistence of Canadian chrysotile asbestos following inhalation: final results through 1 year after cessation of exposure, *Inhalation Toxicology* **17**, 1–14.
- [24] Institute of Medicine of the National Academies (IOM) (2006). *Asbestos: Selected Cancers*, Committee on Asbestos: Selected Health Effects Board on Population health and Public Health Practices, The National Academies Press, Washington, DC.
- [25] Davis, J.M.G. & Jones, A.D. (1988). Comparisons of the pathogenicity of long and short fibres of chrysotile asbestos in rats, *British Journal of Experimental Pathology* **69**, 717–737.
- [26] Ye, J., Zeidler, P., Young, S., Martinez, A., Robinson, V., Jones, W., Baron, P., Shi, X. & Castronova, V. (2001). Activation of mitogen-activated protein kinase p38 and extracellular signal-regulated kinase is involved in glass fiber-induced tumor necrosis factor- α production in macrophages, *The Journal of Biological Chemistry* **276**, 5360–5367.
- [27] Adamson, I.Y.R. & Bowden, D.H. (1987). Response of mouse lung to crocidolite asbestos 1. Minimal fibrotic reaction to short fibers, *The Journal of Pathology* **152**, 99–107.
- [28] Adamson, I.Y.R. & Bowden, D.H. (1987). Response of mouse lung to crocidolite asbestos 2. Pulmonary fibrosis after long fibres, *The Journal of Pathology* **152**, 109–117.
- [29] Hesterberg, T.W., Miiller, W.C., Musselman, R.P., Kamstrup, O., Hamilton, R.D. & Thevenaz, P. (1996).

- Biopersistence of man-made vitreous fibers and crocidolite asbestos in the rat lung following inhalation, *Fundamental and Applied Toxicology* **29**, 267–279.
- [30] U.S. Environmental Protection Agency (2003). *Workshop to Discuss A Proposed Protocol to Assess Asbestos-Related Risk*, San Francisco, CA, February 25–27, Washington, DC.
- [31] Eastern Research Group (ERG) (2003a). *Report on the Peer Consultation Workshop to Discuss a Proposed Protocol to Assess Asbestos-Related Risk*, Prepared for the U.S. Environmental Protection Agency (EPA), Office of Solid Waste and Emergency Response, EPA contract no. 68-C-98-148.
- [32] Agency for Toxic Substances and Disease Registry (ATSDR) (2002). *Expert Panel On Health Effects Of Asbestos And Synthetic Vitreous Fibers (SVF): The Influence Of Fiber Length. Premeeting Comments*, Agency for Toxic Substances and Disease Registry (ATSDR), Division of Health Assessment and Consultation.
- [33] Eastern Research Group, Inc (ERG) (2003b). *Report on the Expert Panel on Health Effects off Asbestos and Synthetic Vitreous Fibers: The Influence off Fiber Length*, Prepared for the Agency for Toxic Substances and Disease Registry (ATSDR), Division of Health Assessment and Consultation.
- [34] Roggli, V.L., George, M.H. & Brody, A.R. (1987). Clearance and dimensional changes of crocidolite asbestos fibers isolated from lungs of rats following short-term exposure, *Environmental Research* **42**, 94–105.
- [35] Churg, A., Wright, J.L., Gilks, B. & Depaoli, L. (1989). Rapid short-term clearance of chrysotile compared with amosite asbestos in the guinea pig, *The American Review of Respiratory Disease* **139**, 885–890.
- [36] Bellmann, B., Muhle, H., Pott, F., Konig, H., Kloppel, H. & Spurny, K. (1987). Persistence of man-made mineral fibres (MMMF) and asbestos in rat lungs, *British Occupational Hygiene Society* **31**, 693–709.
- [37] Coin, P.G., Roggli, V.L. & Brody, A.R. (1992). Deposition, clearance, and translocation of chrysotile asbestos from peripheral and central regions of the rat lung, *Environmental Research* **58**, 97–116.
- [38] Musselman, R.P., Miiller, W.C., Eastes, W., Hadley, J.G., Kamstrup, O., Thevenaz, P. & Hesterberg, T.W. (1994). Biopersistences of man-made vitreous fibers and crocidolite fibers in rat lungs following short-term exposures, *Environmental Health Perspectives* **102**, 139–143.
- [39] Kimizuka, G., Wang, N.S. & Hayashi, Y. (1987). Physical and microchemical alterations of chrysotile and amosite asbestos in the hamster lung, *Journal of Toxicology and Environmental Health* **21**, 251–264.
- [40] Hume, L.A. & Rimstidt, J.D. (1992). The bi durability of chrysotile asbestos, *The American Mineralogist* **77**, 1125–1128.
- [41] Hargreaves, T.W. & Taylor, W.H. (1946). An X-ray examination of decomposition products of chrysotile (asbestos) and serpentine, *Miner Magazine* **27**, 204–216.
- [42] Morgan, A. (1997). Acid leaching studies of chrysotile asbestos from mines in the Coalinga region of California and from Quebec and British Colombia, *British Occupational Hygiene Society* **41**, 249–268.
- [43] Bernstein, D.M. (2005). Understanding chrysotile asbestos: a new perspective based upon current data, *Presented at: International Occupational Hygiene Association's (IOHA) Sixth Annual Conference*, 19–23 September 2005, Pilanesberg National Park.
- [44] Bernstein, D.M. & Hoskins, J.A. (2006). The health effects of chrysotile: current perspectives based upon recent data, *Regulatory Toxicology Pharmacology* **45**, 252–264.
- [45] Wypych, F., Adad, L.B., Mattoso, N., Marangon, A.A. & Schreiner, W.H. (2005). Synthesis and characterization of disordered layered silica obtained by selective leaching of octahedral sheets from chrysotile and phlogopite structures, *Journal of Colloid Interface Science* **283**, 107–112.
- [46] Hodgson, A.A. (1979). Chemistry and physics of asbestos, in *Asbestos-Properties, Applications, And Hazards*, L. Michaels & S.S. Chissick, eds, John Wiley & Sons, New York, pp. 67–114.
- [47] Skinner, H.C.W., Ross, M. & Frondel, C. (1988). *Asbestos and Other Fibrous Materials: Mineralogy, Crystal Chemistry, and Health Effects*, Oxford University Press, New York.
- [48] Governa, M., Amati, M., Fontana, S., Visona, I., Botta, G.C., Mollo, F., Bellis, D. & Bo, P. (1999). Role of iron in asbestos-body-induced oxidant radical generation, *Journal of Toxicology and Environmental Health A* **58**, 279–287.
- [49] Ghio, A.J., LeFurgey, A. & Roggli, V.L. (1997). In vivo accumulation of iron on crocidolite is associated with decrements in oxidant generation by the fiber, *Journal of Toxicology and Environmental Health* **50**, 125–142.
- [50] Mossman, B.T. (2003). Introduction to serial reviews on the role of reactive oxygen and nitrogen species (ROS/RNS) in lung injury and diseases, *Free Radical Biology & Medicine* **34**, 1115–1116.
- [51] Paustenbach, D.J., Finley, B.L., Lu, E.T., Brorby, G.P. & Sheehan, P.J. (2004). Environmental and occupational health hazards associated with the presence of asbestos in brake linings and pads (1900 to present): a “state-of-the-art” review, *Journal of Toxicology and Environmental Health B* **7**, 33–110.
- [52] Merewether, E.R.A. & Price, C.W. (1930). *Report on Effects of Asbestos Dust on the Lungs and Dust Suppression in the Asbestos Industry*, Part I and II, His Majesty's Stationery Office, London, pp. 1–34.
- [53] Osborn, S.H. (1934). *Asbestos Dust Hazards*, Fortyninth Report (57th year) of the State Department of Health, State of Connecticut, Public Document No. 25, Hartford, Connecticut, 507–511.

- [54] Wood, W.B. & Gloyne, S.R. (1934). Pulmonary asbestosis: a review of one hundred cases, *Lancet* **227**, 1383–1385.
- [55] Fulton, W.B., Dooley, A., Matthews, J.L. & Houtz, R.L. (1935). *Asbestosis: Part II: The Nature and Amount of Dust Encountered in Asbestos Fabricating Plants, Part III: The Effects of Exposure to Dust Encountered in Asbestos Fabricating Plants on the Health of A Group of Workers*, Industrial Hygiene Section, Bureau of Industrial Standards, Commonwealth of Pennsylvania Department of Labor and Industry, Special Bulletin No. 42, pp. 1–35.
- [56] Home Office (1935). *Memorandum on the Industrial Diseases of Silicosis and Asbestosis*, His Majesty's Stationery Office, London.
- [57] Lanza, A.J., McConnell, W.J. & Fehnel, J.W. (1935). Effects of the inhalation of asbestos dust on the lungs of asbestos workers, *Public Health Reports* **50**, 1–12.
- [58] Page, R.C. (1935). A study of the sputum in pulmonary asbestosis, *American Journal of the Medical Sciences* **189**, 44–55.
- [59] Donnelly, J. (1936). Pulmonary asbestosis: incidence and prognosis, *The Journal of Industrial Hygiene and Toxicology* **18**, 222–228.
- [60] McPheeters, S.B. (1936). A survey of a group of employees exposed to asbestos dust, *The Journal of Industrial Hygiene and Toxicology* **18**, 229–239.
- [61] Teleky, L. (1937). Review of Windel's "Asbestosis and its prevention" in *Arbeitsschutz* 19(5):9–16, 1937, *The Journal of Industrial Hygiene and Toxicology* **19**, 112.
- [62] Dreessen, W.C., Dallavalle, J.M., Edwards, T.I., Miller, J.W. & Sayers, R.R. (1938). *A Study of Asbestosis in the Asbestos Textile Industry*, US Treasury Department, Public Health Service, National Institute of Health, Division of Industrial Hygiene, Washington, DC. Public health bulletin no. 241.
- [63] George, A.W. & Leonard, R.D. (1939). An X-ray study of the lungs of workmen in the asbestos industry, covering a period of ten years, *Radiology* **33**, 196–202.
- [64] Brachmann, D. (1940). Asbestose bei Bremsbandschleifern und Bohrern [Asbestosis in brake-belt grinders and drillers], *Arbeitsschutz* **3**, 172–174.
- [65] Stone, M.J. (1940). Clinical studies in asbestosis, *American Review of Tuberculosis* **41**, 12–21.
- [66] Teleky, L. (1941). Review of "Studio Sull Asbestosi Nelle Manipature Di Amianto" (Study of asbestosis in the manufacture of asbestos) by Enrico C. Vigliani, 1941, *The Journal of Industrial Hygiene and Toxicology* **23**, 90.
- [67] Wegelius, C. (1947). Changes in the lungs in 126 cases of asbestosis observed in Finland, *Acta Radiologica* **28**, 139–152.
- [68] Castrop, V.J. (1948). Recognition and control of fume and dust exposure, *National Safety News* **57**, 20-21, 52, 73-80.
- [69] Lynch, K.M. & Cannon, W.M. (1948). Asbestos VI: analysis of forty necropsied cases, *Asbestosis* **16**, 874–884.
- [70] Canepa, G. (1949). Asbestos in port workers [L'asbestosi nei lavoratori portuali], *Journal of Legal Medicine and Insurance* **12**, 188–205.
- [71] Barnett, G.P. (1949). *Annual Report of the Chief Inspector of Factories for the Year 1947*, His Majesty's Stationery Office, London.
- [72] Cartier, P. (1952). Abstract of discussion, *Archives of Industrial Hygiene Occupational Medicine* **5**, 262–263.
- [73] Lynch, K.M. (1953). Discussion – Potential occupational factors in lung cancer: Asbestos, in *Proceedings of the Scientific Session, Cancer of the Lung: An Evaluation of the Problem, Annual Meeting*, November 3–4, American Cancer Society, pp. 115–118.
- [74] Breslow, L., Hoaglin, L., Rasmussen, G. & Abrams, H.K. (1954). Occupations and cigarette smoking as factors in lung cancer, *American Journal of Public Health* **44**, 171–181.
- [75] Knox, J.F. & Beattie, J. (1954). Mineral content of the lungs after exposure to asbestos dust, *American Medical Association Archives of Industrial Hygiene* **10**, 23–29.
- [76] Bonser, G.M., Stewart, M.J. & Faulds, J.S. (1955). Occupational cancer of the urinary bladder in dyestuffs operatives and the lung in asbestos textile workers and iron-ore miners, *American Journal of Clinical Pathology* **25**, 126–134.
- [77] Cartier, P. (1955). Some clinical observations of asbestosis in mine and mill workers, *American Medical Association Archives of Industrial Health* **11**, 204–207.
- [78] Doll, R. (1955). Mortality from lung cancer in asbestos workers, *British Journal of Industrial Medicine* **12**, 81–86.
- [79] Frost, J., Georg, J. & Møller, P.F. (1956). Asbestosis with pleural calcification among insulation workers, *Danish Medical Bulletin* **3**, 202–204.
- [80] Williams, W.J. (1956). Alveolar metaplasia: its relationship to pulmonary fibrosis in industry and the development of lung cancer, *British Journal of Cancer* **11**, 30–42.
- [81] Thomas, D.L.G. (1957). Pneumonokoniosis in Victorian industry, *Medical Journal of Australia* **1**, 75–77.
- [82] Braun, C.D. & Truan, T.D. (1958). An epidemiological study of lung cancer in asbestos miners, *American Medical Association Archives of Industrial Health* **17**, 634–654.
- [83] Horai, Z., Tsujimoto, T., Ueshima, M. & Sano, H. (1958). Studies on asbestosis. III: A survey of the asbestosis in an asbestos factory in 1956, *Journal of Nara Medical Association* **9**, 48–56.
- [84] Gloyne, S.R. (1951). Pneumoconiosis: a histological survey of necropsy material in 1205 cases, *Lancet* **260**, 810–814.
- [85] Wagner, J.C., Sleggs, C.A. & Marchand, P. (1960). Diffuse pleural mesothelioma and asbestos exposure in the north western Cape Province, *British Journal of Industrial Medicine* **17**, 260–271.

- [86] Thomson, J.G., Kaschula, R.O.C. & MacDonald, R.R. (1963). Asbestos as a modern urban hazard, *South African Medical Journal* **37**, 77–81.
- [87] Selikoff, I.J., Churg, J. & Hammond, E.C. (1964). Asbestos exposure and neoplasia, *Journal of the American Medical Association* **188**, 22–26.
- [88] Selikoff, I.J., Churg, J. & Hammond, E.C. (1965). Relation between exposure to asbestos and mesothelioma, *New England Journal of Medicine* **272**, 560–565.
- [89] Talcott, J.A., Thurber, W.A., Kantor, A.F., Gaensler, E.A., Danahy, J.F., Antman, K.H. & Li, F.P. (1989). Asbestos-associated diseases in a cohort of cigarette-filter workers, *New England Journal of Medicine* **321**, 1220–1223.
- [90] Hilt, B., Andersen, A., Rosenberg, J. & Langard, S. (1991). Cancer incidence among asbestos-exposed chemical industry workers: an extended observation period, *American Journal of Industrial Medicine* **20**, 261–264.
- [91] Tarchi, M., Orsi, D., Comba, P., De Santis, M., Pirastu, R., Battista, G. & Valiani, M. (1994). Cohort mortality study of rock salt workers in Italy, *American Journal of Industrial Medicine* **25**, 251–256.
- [92] Oksa, P., Pukkala, E., Karjalainen, A., Ojajarvi, A. & Huuskonen, M.S. (1997). Cancer incidence and mortality among Finnish asbestos sprayers and in asbestosis and silicosis patients, *American Journal of Industrial Medicine* **31**, 693–698.
- [93] Fletcher, A.C., Engholm, G. & Englund, A. (1993). The risk of lung cancer from asbestos among Swedish construction workers: Self-reported exposure and a job exposure matrix compared, *International Journal of Epidemiology* **22**, S29–S35.
- [94] Ohlson, C.G., Klaesson, B. & Hogstedt, C. (1984). Mortality among asbestos-exposed workers in a railroad workshop, *Scandinavian Journal of Work, Environment, and Health* **10**, 283–291.
- [95] Nokso-Koivisto, P. & Pukkala, E. (1994). Past exposure to asbestos and combustion products and incidence of cancer among Finnish locomotive drivers, *Occupational Environmental Medicine* **51**, 330–334.
- [96] Sanden, A., Jarvholm, B., Larsson, S. & Thiringer, G. (1992). The risk of lung cancer and mesothelioma after cessation of asbestos exposure: a prospective cohort study of shipyard workers, *The European Respiratory Journal* **5**, 281–285.
- [97] Gustavsson, P., Plato, N., Lidstrom, E.B. & Hogstedt, C. (1990). Lung cancer and exposure to diesel exhaust among bus garage workers, *Scandinavian Journal of Work, Environment, and Health* **16**, 348–354.
- [98] Wagner, J.C., Berry, G. & Pooley, F.D. (1982). Mesotheliomas and asbestos type in asbestos textile workers: a study of lung contents, *British Medical Journal* **285**, 603–605.
- [99] Davis, J.M.G. (1989). Mineral fibre carcinogenesis: experimental data relating to the importance of fibre type, size, deposition, dissolution and migration, in *Non-occupational Exposure to Mineral Fibres*, IARC Scientific Publications No. 90, J. Bignon, J. Peto & R. Saracci, eds, International Agency for Research on Cancer, Lyon.
- [100] Churg, A. (1988). Chrysotile, tremolite and malignant mesothelioma in man, *Chest* **93**, 621–628.
- [101] Wagner, J.C. (1991). The discovery of the association between blue asbestos and mesotheliomas and the aftermath, *British Journal of Industrial Medicine* **48**, 399–403.
- [102] Wagner, J.C. (1997). Asbestos-related cancer and the amphibole hypothesis: the first documentation of the association, *American Journal of Public Health* **87**, 687–688.
- [103] Hodgson, J.T. & Darnton, A. (2000). The quantitative risks of mesothelioma and lung cancer in relation to asbestos exposure, *The Annals of Occupational Hygiene* **44**, 565–601.
- [104] Berman, D.W. & Crump, K.S. (2003). *Final Draft: Technical Support Document For A Protocol To Assess Asbestos-Related Risk*, EPA# 9345.4-06, US Environmental Protection Agency (EPA), Office of Solid Waste and Emergency Response, Washington, DC.
- [105] American Conference of Governmental Industrial Hygienists (ACGIH) (1946). Report of the sub-committee on threshold limits, *Annals of the American Conference of Industrial Hygiene* **9**, 343–480.
- [106] LaNier, M.E. (ed) (1984). Threshold Limit Values – Discussion and Thirty-Five Year Index with Recommendations, *American Conference of Governmental Industrial Hygienists (ACGIH)*, Cincinnati.
- [107] American Conference of Governmental Industrial Hygienists (ACGIH) (1974). *TLVs Threshold Limit Values for Chemical Substances in Workroom Air Adopted By Acih For 1974*, pp. 33–34.
- [108] American Conference of Governmental Industrial Hygienists (ACGIH) (1980). Asbestos', *American Conference of Governmental Industrial Hygienists (ACGIH) Documentation of the Threshold Limit Values: For Chemical Substances in the Workroom Environment*, 4th Edition, ACGIH Signature Publications, pp. 27–30.
- [109] American Conference of Governmental Industrial Hygienists (ACGIH) (1998). Asbestos, *All Forms. Supplements to the Sixth Edition: Documentation of the Threshold Limit Values and Biological Exposure Indices*, ACGIH Signature Publications, pp. 1–13.
- [110] Occupational Safety and Health Administration (OSHA) (1971). National consensus standards and established federal standards, *Federal Register* **36**, 10466–10506.
- [111] Occupational Safety and Health Administration (OSHA) (1972). Standard for exposure to asbestos dust, *Federal Register* **37**, 11318–11322.
- [112] Occupational Safety and Health Administration (OSHA) (1975). Occupational exposure to asbestos – notice of proposed rulemaking, *Federal Register* **40**, 47652–47665.
- [113] Occupational Safety and Health Administration (OSHA) (1986). Occupational exposure to asbestos,

tremolite, anthophyllite, and actinolite; final rules, *Federal Register* **51**, 22612.

- [114] Occupational Safety and Health Administration (OSHA) (1988). Occupational exposure to asbestos, tremolite, anthophyllite and actinolite, *Federal Register* **53**, 35610–35629.

- [115] Occupational Safety and Health Administration (OSHA) (1994). Occupational exposure to asbestos; final rule, *Federal Register* **59**, 40964–40966.

Related Articles

Environmental Hazard

What are Hazardous Materials?

JENNIFER S. PIERCE AND DENNIS J.
PAUSTENBACH

Assessment of Risk Association with Beryllium Exposure

Beryllium is the lightest of all solid, chemically stable compounds. It is the metal with an atomic number of 4 and a molecular weight of 9.04. The two commercially important ore forms are bertrandite extracted as beryllium hydroxide and beryl extracted as beryllium oxide. In addition to beryllium mined in the United States, three to four times as much is imported, mainly from Brazil. There are two gem forms of beryl: emerald and aquamarine.

Beryllium has a high strength to weight ratio, but the cost of the metal limits its use to the computer, airline, and space industries. In alloy forms, it is widely used in home, commercial, and industrial tools primarily as a hardening agent, and in ceramics. Beryllium exposures are mainly limited to industrial settings of mining, processing, and machining. As exposure to tiny amounts ($1\text{--}2\text{ }\mu\text{g m}^{-3}$) can cause significant health effects, emissions from these industries are carefully scrutinized to protect the surrounding communities.

Hazard Identification

A number of more comprehensive reviews on beryllium toxicity are available [1–9].

Human Toxicity

As with many metals mined as ores, beryllium toxicity was first documented in workers in mining, processing, and machining facilities. Beryllium fluoride was first reported to produce serious health effects in 1933 based on reported cases of bronchitis, broncholitis, and confluent lung densities in beryllium ore extraction workers in Europe [10]. Other reports from Europe and Asia [11] also noted serious health effects in extraction workers, due to beryllium compounds. The first report of health effects due to a beryllium compound in the United States was a 1943 report of three cases of chemical pneumonitis in workers from an extraction plant in Ohio [12]. As the compound was beryllium fluoride, the disease was thought to be due to fluoride rather than

beryllium. As more health concerns were reported from the beryllium industry, the effects seen with soluble forms were significantly different from the effects with insoluble forms. This led to a clinical picture that was very confusing and it did not seem plausible that all of the effects were due to beryllium [13]. The first clear description of a clinical beryllium disease was reported in 1946. Hardy and Tabershaw [14] reported on the case histories of 3 men and 14 women who had been employed for only 17 months in the beryllium industry. The syndrome was described as a delayed chemical pneumonitis accompanied with granulomas in the lung, lymph nodes, and liver. As more cases were reported, it became clear that beryllium was the cause of specific health effects as reported at the Sixth Saranac Symposium in 1947 [15, 16].

In following the cohort in the Hardy and Tabershaw [14] report, six of the women died of what became known as *chronic beryllium disease* (CBD) or *berylliosis*. Autopsies revealed pulmonary granulomas, lung fibrosis, and enlarged heart [14, 17, 18]. CBD is an interstitial lung disease characterized by interstitial cellular infiltration, fulminating edema, and fibrosis, with or without granulomas and does not always develop during exposure, but may suddenly appear 10–15 years after termination of beryllium exposure [3]. Unless treated, the latent form of the disease is usually fatal [19]. Beryllium workers often report symptoms of shortness of breath, fatigue, and weight loss and it should be noted that CBD is a systemic disease causing anemia, anorexia, and granulomas in bone, liver, kidney, and spleen, as well as in the lung [20].

Subsequent study of beryllium workers identified a distinct acute form of beryllium disease [12, 21] as well as a chronic form of the disease [14, 22]. The acute form is seen in some workers regularly exposed to more than $100\text{ }\mu\text{g}$ of Be/m^3 of soluble beryllium compounds and in almost all workers exposed to $1000\text{ }\mu\text{g}$ of Be/m^3 , even for a short time [23]. The disease is characterized by acute pneumonitis and cases involving massive pulmonary edema are often fatal. In workers who survive the pulmonary edema, the disease usually resolves within 1 year after removal from exposure and does not progress to CBD [24]. A total of 10 fatalities among 93 cases of acute beryllium disease were noted in workers from two beryllium processing plants in the United States [25]. These workers had

begun working in the plants prior to 1949 and were exposed to massive levels of beryllium. No new cases of acute beryllium disease have been reported since 1950 except for cases involving an industrial accident [2]. Some of the patients who recovered from the acute disease eventually developed lung cancer [26].

The findings of Van Ordstrand *et al.* [12, 23] and Hardy and Tabershaw [14], along with the conclusions of the Saranac Symposium [15, 16] prompted the Atomic Energy Commission in 1949 to implement a hygienic standard of $2\text{ }\mu\text{g}$ of Be/m^3 as a quarterly daily weighted average (DWA) with a ceiling of $25\text{ }\mu\text{g}$ of Be/m^3 . The DWA is based on area samples for the entire shift rather than a time-weighted average over a specific time period within the workshift. This rather drastic reduction in exposure concentrations sharply reduced the number of new cases of acute beryllium disease. In order to collect pertinent medical data for study and evaluation, the US Beryllium Case Registry (BCR) was initiated in 1952 [27].

Further study of CBD noted that the workers were generally exposed to the metal or other insoluble forms of beryllium over a long period of time [2]. Once the disease was diagnosed, only mild forms of the disease resolved after removal from exposure [22]. Workers may develop symptoms of CBD but often a latent form suddenly develops several years after the last exposure [17] and is usually fatal unless treated. The etiology of CBD was confusing to clinicians and toxicologists as some beryllium plants had high concentrations, but the workers were relatively free of CBD [3]. In other plants, the workers were exposed to significantly lower concentrations, but often experienced a much higher incidence of CBD. Some of the confusing aspects of CBD were resolved when Hall *et al.* [28] noted that the surface area of insoluble beryllium particles determined the toxic response, i.e., the greater the surface area, the greater the toxicity. This did not help resolve the enigma of a latent form of the disease. The latent form of the disease was not resolved until clinicians and toxicologists began to consider an immune mechanism for CBD.

The very first beryllium epidemiology study did not deal with beryllium workers, but instead involved 10 000 individuals residing within 2 miles of beryllium plants [29, 30]. Eleven neighborhood cases

of CBD were discovered *via* X-ray analysis. The ambient 24-h concentration was estimated to range between 0.05 and $0.2\text{ }\mu\text{g}$ of Be/m^3 . In a follow-up study, three additional cases within 0.75 miles of the plant were identified [17]. Beyond 0.75 miles, the average ambient air concentration was between 0.01 and $0.1\text{ }\mu\text{g}$ of Be/m^3 and no cases of CBD were found between 0.75 and 2 miles from the plant. On this basis, $0.1\text{ }\mu\text{g}$ of Be/m^3 was considered the no-observed-adverse-effect level (NOAEL) for the development of CBD. A total of 60 neighborhood cases of CBD were reported within 0.75 miles of a beryllium plant by 1966 [31]. Similar results were noted in residents near plants in Pennsylvania [32, 33]. These nonoccupational, low-exposure level cases along with a number of cases among secretaries at plants or others with light exposure suggested that a small percentage of the population can be sensitized to beryllium [34].

Sensitization has been confirmed in animal studies [35, 36] and in a number of epidemiology studies [37–41]. A major step forward in the diagnosis of CBD was the development of a beryllium lymphocyte proliferation test (LPT) [22, 42]. In the Kriess *et al.* [41] study, LPT was performed by two independent laboratories on samples collected from 136 employees. Five workers had positive LPT results from both laboratories and were diagnosed as having CBD on detecting granulomas in lung biopsy samples. Two other workers had abnormal LPT but no granulomas were found in biopsy samples. At least one of these two employees developed clinical signs of CBD within 2 years. Another worker developed a skin granuloma and had abnormal LPT. This worker developed lung granulomas within 2 years. Only one of these cases of CBD had abnormal chest X rays. A total of 11 out of 573 former workers had confirmed cases of CBD bringing the total incidence to 19/709. After considerable analysis of exposure levels, the lowest level producing CBD was estimated to be $0.55\text{ }\mu\text{g}$ of Be/m^3 . Based on ambient air concentration data, the NOAEL for sensitization was estimated to be between 0.01 and $0.1\text{ }\mu\text{g}$ of Be/m^3 [29].

In the studies that followed, it gradually became clear that CBD is an immunological disease or hypersensitivity reaction. Beryllium-sensitized cells accumulate to form granulomas in the lung [43–45]. All CBD cases that have progressed to the granulomas stage have a cell-mediated immune response

to beryllium [43, 44], and treatment that controls the immune response controls CBD [46]. Specific antibodies to beryllium have been identified [47]. Although variable results were seen with the blood LPT, the findings of the lung LPT (lymphocytes obtained *via* lung lavage) has been consistently positive in CBD patients [43, 44]. Improvements in the blood LPT, i.e., the use of tritiated thymidine, led to a much more reliable test [43, 48]. While the LPT has been extremely useful, a recent evaluation of the LPT in a long-term surveillance program [49], cautions against the reliability of a single LPT result. Multiple tests over weeks or months are considerably more reliable.

It is interesting to note that the incidence among workers exposed to 2–25 μg of Be/m^3 is about the same as past exposures to much higher levels ($>25 \mu\text{g}$ of Be/m^3) [17]. The TLV of 2 μg of Be/m^3 was first proposed in 1955 and adopted in 1959 [50, 51]. Occupational Safety and Health Administration (OSHA) adopted the level of 2 μg of Be/m^3 in 1971 [52]. This level was expected to eliminate the development of CBD in the workforce and while the number of new cases was reduced, CBD was not eliminated [24]. None-the-less, the BCR was eliminated in the 1980s. Of the 846 registered cases, 224 were of acute beryllium disease that resulted in 93 deaths [34]. Of the 622 CBD cases, 557 were due to occupational exposure, 42 due to neighborhood exposure, and 23 due to household exposure *via* handling contaminated clothing. The fact that beryllium is a sensitizer led the American Conference of Governmental Industrial Hygienists ACGIH [53] to recommend reducing the TLV from 2 to 0.2 μg of Be/m^3 in 2004, adding a sensitizer and a skin notation. The recommended level was lowered to 0.05 μg of Be/m^3 in 2006, with a short term exposure limit (STEL) of 0.02 μg of Be/m^3 [9, 54].

The last aspect of beryllium toxicity to surface was lung cancer. Early studies of the beryllium cohort were negative for cancer [31, 55, 56]. However, in workers with short-term exposure to beryllium, six out of eight lung cancer patients had recovered from a previous episode of acute beryllium disease [26]. More recent epidemiology studies have reported increased cancer rates for beryllium workers beginning in 1979 [57–59] with a follow-up in 1992 study [60]. In 1980, Infante *et al.* [61] studied the entrants to the BCR with a follow-up

study in 1991 [62]. As a result, the International Agency for Research on Cancer [63] stated that there is “limited evidence” that beryllium causes cancer in humans, but this was later changed to “sufficient evidence” that beryllium causes cancer [64]. Beryllium was first listed as “reasonably anticipated to be a human carcinogen” in the 2nd Annual Report on Carcinogens [65] and was later listed as “a known human carcinogen” in the 10th Report on Carcinogens [66]. In 1998, EPA identified beryllium as a “probable” human carcinogen in IRIS [4].

Animal Toxicity

Animal bioassays of beryllium were first conducted in 1946 as part of the Manhattan Project and only after several studies of health effects in humans had been reported [3]. Since that time, animal studies have confirmed many of the effects seen in worker populations.

In acute studies, soluble beryllium compounds are quite toxic *via* inhalation with 4-h LC_{50} 's ranging from 0.15 mg of Be/m^3 in the rat to 4.2 mg of Be/m^3 in the guinea pig [67]. Inhalation of insoluble forms is less acutely toxic than soluble forms.

Exposure of Fischer rats to 800 mg m^{-3} (800 000 $\mu\text{g m}^{-3}$) for 50 min produced effects in 14 days that resembled acute beryllium disease as seen in humans, including necrotizing, hemorrhagic exudative pneumonitis and alveolar fibrosis [68]. Eventually, foreign body granulomas were formed, but not similar to the granulomas as seen in the immune mechanism of CBD. Exposure to 0.013 mg m^{-3} (13 $\mu\text{g m}^{-3}$) [69] induced hyperplasia and infiltration of macrophages. Infiltrates of lymphocytes were not present [70] and the response resolved within 3 weeks postexposure.

Sensitization leading to granuloma formation was not seen in inhalation studies using soluble salts of beryllium in rats (Hart *et al.* [71]) or in mice [72]. Some aspects of CBD were seen in guinea pigs [73], beagle dogs [36, 74], and cynomolgus monkeys [74]. Evidence of a lymphocyte accumulation in the lungs of rats exposed to 1–100 mg m^{-3} (1000–100 000 $\mu\text{g m}^{-3}$) for 30–180 min [75]. Unfortunately, most animal inhalation studies involve liquid aerosols of soluble salts, e.g., Schepers *et al.* [76] rather than dust aerosols of insoluble forms

Table 1 Dose–response data based on beryllium epidemiology studies

Exposure level (μg of Be/m^3)	Results	References
0.01–0.1	No effect level for human exposure to airborne beryllium.	[17, 29]
0.05–0.2	Sensitization in susceptible individuals; CBD may develop in sensitized individuals upon continued exposure to $0.5\text{--}2.0\mu\text{g m}^{-3}$.	[17, 85]
2.0–28.4	No effect in nonsensitized workers; 6/146 sensitized workers exposed to $2\mu\text{g}$ of Be/m^3 developed CBD.	[24, 34]
5–100	Many early cases of CBD resulted from exposure to high levels of insoluble forms. Particles with large surface area are more potent than smaller particles.	[1, 2, 28]
>25 (peaks to 1310)	31/214 had chest X-ray abnormalities which resolved within 3 years after levels were reduced below $25\mu\text{g m}^{-3}$. Some evidence that mild early stages of CBD may be reversible.	[86, 87]
25–100	2/19 workers exposed to soluble forms developed acute beryllium disease. Both recovered and did not develop CBD.	[24]
100–1000	Most workers exposed to soluble beryllium salts develop acute beryllium disease, sometimes after a single exposure to high levels ($1000\mu\text{g}$ of Be/m^3). The disease resolves except in workers with massive pulmonary edema.	[12, 17]

such as the metal or beryllium oxide, e.g., Wagner *et al.* [77].

Lung cancer was seen in chronic inhalation studies in rats [78–81] and in monkeys [77, 78, 82]. No increase in tumor incidence was noted in hamsters [77].

Dose–Response

EPA has reviewed the animal toxicity and human epidemiology literature for beryllium [4, 83]. In making risk calculations, EPA selected the Wagoner *et al.* [59] study as the principal study for determining cancer risk; the Kreiss *et al.* [41] and Eisenbud *et al.* [29] as the principal studies for determining the reference concentration (RfC); and the Morgareidge *et al.* [84] as the principal study for calculating the reference dose (RfD). Of these, only the Morgareidge *et al.* study [84] had suitable dose–response data (*see Dose–Response Analysis*).

ATSDR also reviewed the animal toxicity and human epidemiology literature for beryllium [1, 2]. No inhalation minimum risk levels (MRLs) were calculated because the acute beryllium disease and CBD seen in animals did not totally mimic the disease

states seen in humans. For the chronic inhalation MRL, ATSDR did not accept the Eisenbud study [29] as a NOAEL for CBD.

While epidemiology data are often weakened because long-term, individual exposure levels in all job classifications are not readily available, the exposure data for beryllium has been reviewed rather closely and the dose–response data (Table 1) are presented as follows:

Exposure Assessment

The current environmental and hygienic standards for beryllium are as follows:

$$RfD = 0.002\text{ mg kg}^{-1}\text{ day}^{-1} \text{ [4].}$$

$$RfC = 0.02\mu\text{g m}^{-3} \text{ [4]}$$

$$TLV = 2\mu\text{g m}^{-3} \text{ with a notice to change to } 0.05\mu\text{g m}^{-3} \text{ [54]}$$

The RfD is considered a safe daily intake for the most sensitive members of the population. Similarly, the RfC is considered to be a safe concentration in the ambient air, i.e., safe even for the most sensitive in the population to breathe

continuously everyday. Ideally, the TLV is a safe time-weighted average concentration for the most sensitive members of the workforce to breathe for up to 8 h day⁻¹, 40 h wk⁻¹ but, by definition, it is meant to protect “nearly all of the exposed worker population”.

Almost all exposure will be limited to occupational exposure once the recommended TLV takes effect. In most regions of the United States, the ambient air beryllium concentration is below the level of detection of 0.03 ng m⁻³ [2]. Currently, urban areas have higher levels primarily due to the burning of coal and fuel oil [5]. Between 1982 and 1992, the annual ambient air concentration of beryllium was between 0.02 and 0.2 ng m⁻³ in Detroit (ATSDR [2]). In areas near beryllium processing plants, the ambient air concentration has been as high as 0.2 µg m⁻³ [29]. When the recommended TLV of 0.05 µg m⁻³ becomes effective, one could expect the neighborhood ambient air concentration to drop below 0.001–0.005 µg m⁻³ (1–5 ng m⁻³) within 0.75 miles of the plant and below 1 ng m⁻³ beyond 0.75 miles. Since sensitization is the critical effect and a NOAEL was reported as 0.01–0.1 µg m⁻³ (10–100 ng m⁻³), levels of 1–5 ng m⁻³ are not expected to affect the most sensitive members of the population.

A possible exception to this could be a worker, e.g., a jeweler working alone, who is not aware of the beryllium concentrations generated in his shop. Some hobbies, e.g., maintenance of the brakes on a private plane, could also result in inhaling beryllium dust. Beyond these types of possible exposure scenarios, the members of the general population will be exposed at levels well below the RfC.

Risk Characterization

EPA has calculated an RfD, an RfC, and a cancer unit risk for beryllium [4]. These levels are as follows:

$$RfD = 0.002 \text{ mg kg}^{-1} \text{ day}^{-1}.$$

$$RfC = 0.02 \text{ } \mu\text{g m}^{-3}.$$

$$\text{Cancer unit risk (q1*)} = 2.43 \text{ } \mu\text{g}^{-1} \text{ m}^{-3}.$$

ATSDR did not calculate any inhalation MRLs because their reviewers were: (a) not convinced that 0.1 µg m⁻³ was a NOAEL for sensitization because the latest detection techniques for sensitization were

not available in 1949; and (b) the acute and chronic forms of beryllium disease seen in animals does not totally mimic the diseases seen in humans. ATSDR did calculate a chronic oral MRL as 0.002 mg⁻¹ kg⁻¹ day⁻¹ [2].

References

- [1] ATSDR (1993). *Toxicological Profile for Beryllium*, US Department of Health and Human Services, Atlanta.
- [2] ATSDR (2003). *Toxicological Profile for Beryllium*, US Department of Health and Human Services, Atlanta.
- [3] Beliles, R.P. (1994). In *Patty's Industrial Hygiene and Toxicology*, 4th Edition, G.D. Clayton & F.E. Clayton, eds, John Wiley & Sons, New York, Vol. IIC, pp. 1930–1948.
- [4] Integrated Risk Information System (IRIS) (2002). *Beryllium and Compounds*, at <http://www.epa.gov/iris/subst/0014.htm> (accessed Mar 2002).
- [5] National Toxicology Program (NTP) (2005). *Eleventh Report on Carcinogens*, U.S. Department of Health and Human Services, Research Triangle Park, at <http://ntp-server.niehs.nih.gov> (accessed Jan 2005).
- [6] Mroz, M.M. (2005). In *Patty's Industrial Hygiene and Toxicology*, 5th Edition, E. Bingham, B. Cohnssen & C.H. Powell, eds, John Wiley & Sons, New York, Vol. IIC.
- [7] Hazardous Substances Data Bank (HSDB) (2006). *Beryllium*, at <http://toxnet.nlm.nih.gov/cgi-bin> (accessed Nov 2007).
- [8] Borak, J. (2006). The beryllium occupational exposure limit: historical origin and current inadequacy, *Journal of Occupational and Environmental Medicine* **48**(2), 109–116.
- [9] American Conference of Governmental Industrial Hygienists (ACGIH) (2007). *Documentation of TLVs and BEIs with Other Worldwide Occupational Exposure Values*, Cincinnati.
- [10] Weber, H.H. & Engelhardt, W.E. (1933). Lentralbe, *Gewerbehyg. Unfallverhuet* **10**, 41.
- [11] Gelman, I. (1936). Poisoning by vapors of beryllium oxyfluoride, *The Journal of Industrial Hygiene and Toxicology* **18**, 371–379.
- [12] Van Ordstrand, H.S., Hughes, R. & Carmody, M.G. (1943). Chemical pneumonia in workers extracting beryllium oxide, *Cleveland Clinic Quarterly* **10**, 10–18.
- [13] Hyslop, F., Palmes, E.D., Alford, W.C., Monaco, A.R. & Fairhall, L.T. (1943). *The Toxicology of Beryllium (US Public Health Service Bulletin 181)*, US Public Health Service, Washington, DC.
- [14] Hardy, H.L. & Tabershaw, I.R. (1946). Delayed chemical pneumonitis occurring in workers exposed to beryllium compounds, *The Journal of Industrial Hygiene and Toxicology* **28**, 197–211.
- [15] Hardy, H.L. (1950). Clinical and epidemiologic aspects, in *Pneumoconiosis Sixth Saranac Symposium*, Saranac

- Lake, 1947, A.J. Vorwald, ed, Paul B. Hoeber, New York, pp. 133–151.
- [16] Hoeber, P. (1950). Pneumoconiosis: beryllium, bauxite fumes, and compensation, *Sixth Saranac Symposium*, Saranac.
- [17] Sterner, J.H. & Eisenbud, M. (1951). Epidemiology of beryllium intoxication, *Archives of Industrial Hygiene and Occupational Medicine* **4**, 123–151.
- [18] Aronchik, J.M. (1992). Chronic beryllium disease occupational lung disease, *Radiologic Clinics of North America* **30**(6), 1209–1217.
- [19] Stokinger, H.E. (1966). In *Patty's Industrial Hygiene and Toxicology*, 3rd Edition, G. Clayton & D. Clayton, eds, John Wiley & Sons, New York, Vol. IIA, pp. 1537–1558.
- [20] Tepper, L.B., Hardy, H.L. & Chamberlin, R.I. (1961). *Toxicity of Beryllium Compounds*, Elsevier Science, New York.
- [21] Sprince, N.L. & Kazemi, H. (1980). US beryllium case registry through 1977, *Environmental Research* **21**, 44–47.
- [22] Newman, L.S., Kreiss, K., King Jr, T.E., Seay, S. & Campbell, P.A. (1989). Pathologic and immunological alterations in early stages of beryllium disease, *The American Review of Respiratory Disease* **139**, 1479–1486.
- [23] Van Ordstrand, H.S., Hughes, R., DeNardi, J.M. & Carmody, G.M. (1945). Beryllium poisoning, *Journal of the American Medical Association* **129**, 1084–1090.
- [24] Cotes, J.E., Gilson, J.C., McKerrow, C.B. & Oldham, P.D. (1983). A long-term follow-up of workers exposed to beryllium, *British Journal of Industrial Medicine* **40**, 13–21.
- [25] American College of Chest Physicians (1965). Beryllium disease: report of the section on nature and prevalence, *Diseases of the Chest* **48**, 550–558.
- [26] Mancuso, T.F. (1970). Relation of duration of employment and prior respiratory illness to respiratory cancer among beryllium workers, *Environmental Research* **3**, 251–275.
- [27] Hardy, H.L. (1952). *American Review of Tuberculosis* **72**, 129.
- [28] Hall, R.H., Scott, J.K., Laskin, S., Stroud, C.A. & Stokinger, H.E. (1950). Acute toxicity of inhaled beryllium, *Archives of Industrial Hygiene and Occupational Medicine* **12**, 25–48.
- [29] Eisenbud, M., Wanta, R.C., Dustan, C., Steadman, L.T., Harris, W.B. & Wolf, B.S. (1949). Non-occupational berylliosis, *The Journal of Industrial Hygiene and Toxicology* **31**, 282–294.
- [30] DeNardi, J.M. (1959). Long term experience with beryllium, *AMA Archives of Industrial Health* **19**, 110–116.
- [31] Stoeckle, J.D., Hardy, H.L. & Weber, A.L. (1969). Chronic beryllium disease, *American Journal of Medicine* **46**, 545–561.
- [32] Leiben, J. & Metzner, F. (1959). Epidemiological findings associated with beryllium extraction, *American Industrial Hygiene Association Journal* **20**, 494–499.
- [33] Sussman, V.H., Lieben, J. & Cleland, J.G. (1959). An airpollution study of a community surrounding a beryllium plant, *American Industrial Hygiene Association Journal* **20**, 504–508.
- [34] Eisenbud, M. & Lisson, J. (1983). Epidemiological aspects of beryllium-induced nonmalignant lung disease: a 30-year update, *Journal of Occupational Medicine* **25**, 196–202.
- [35] Stiefel, T., Schulze, K., Zorn, H. & Tolg, G. (1980). Toxicokinetic and toxicodynamic studies of beryllium, *Archives of Toxicology* **45**, 81–92.
- [36] Haley, P.J., Finch, G.L., Mewhinney, J.A., Harmsen, A.G., Hahn, F.F., Hoover, M.D., Muggenburg, B.A. & Bice, D.E. (1989). A canine model of beryllium-induced granulomatous lung disease, *Laboratory Investigation* **61**, 219–227.
- [37] Rom, W.N., Lockney, J.E., Bang, K.M., Dewitt, C. & Jones Jr, R.E. (1983). Reversible beryllium sensitization in a prospective study of beryllium workers, *Archives of Environmental Health* **38**(5), 302–307.
- [38] Cullen, M.R., Kominsky, J.R., Rossman, M.D., Cherniack, M.G., Rankin, J.A., Balmes, J.R., Kern, J.A., Daniele, R.P., Palmer, L. & Naegel, G.P. (1987). Chronic beryllium disease in a precious metal refinery, *The American Review of Respiratory Disease* **135**, 201–208.
- [39] Kreiss, K., Newman, L.S., Mroz, M.M. & Campbell, P.A. (1989). Screening blood test identifies subclinical beryllium disease, *Journal of Occupational Medicine* **31**, 603–608.
- [40] Kreiss, K., Wasserman, S., Mroz, M.M. & Newman, L.S. (1993). Beryllium disease screening in the ceramics industry. Blood lymphocyte test performance and exposure-disease relations, *Journal of Occupational Medicine* **35**(3), 267–274.
- [41] Kreiss, K., Mroz, M.M., Newman, L.S., Martyny, J. & Zhen, B. (1996). Machining risk of beryllium disease and sensitization with median exposures below $\mu\text{g}/\text{m}^3$, *American Journal of Industrial Medicine* **30**(1), 16–25.
- [42] Newman, L.S. (1996). Significance of the blood beryllium lymphocyte proliferation test, *Environmental Health Perspectives*, **104**(Suppl. 5), 953–956.
- [43] Rossman, M.D., Kern, J.A., Elias, J.A., Cullen, M.R., Epstein, P.E., Pruess, O.P., Markham, T.N. & Daniele, R.P. (1988). Proliferative response of bronchoalveolar lymphocytes to beryllium, *Annals of Internal Medicine* **108**, 687–693.
- [44] Saltini, C., Winestock, K., Kirby, M., Pinkston, P. & Crystal, R.G. (1989). Maintenance of alveolitis in patients with chronic beryllium disease by beryllium-specific helper T cells, *New England Journal of Medicine* **320**, 1103–1109.
- [45] Saltini, C., Kirby, M., Trapnell, B.C., Tamura, N. & Crystal, R.G. (1990). Biased accumulation of T lymphocytes with “memory”-type CD45 leukocyte common antigen gene expression on the epithelial surface of the human lung, *Journal of Experimental Medicine* **171**, 1123–1140.

- [46] Aronchick, J.M., Rossman, M.D. & Miller, W.T. (1987). Chronic beryllium disease: diagnosis, radiographic findings, and correlation with pulmonary function tests, *Radiology* **163**, 677–682.
- [47] Clarke, S.M. (1991). A novel enzyme-linked immuno sorbent assay (ELISA) for the detection of beryllium antibodies, *Journal of Immunological Methods* **137**, 65–72.
- [48] Mroz, M.M., Kreiss, K., Lezotte, D.C., Campbell, P.A. & Newman, L.S. (1991). Reexamination of the blood lymphocyte transformation test in the diagnosis of chronic beryllium disease, *Journal of Allergy and Clinical Immunology* **88**, 54–60.
- [49] Donovan, E. P., Kolan, M. E., Galbraith, D. A., Chapman, P.S. & Paustenbach, D.J. (2007). Performance of the beryllium blood lymphocyte proliferation test based on a long-term occupational program, *International Archives of Occupational and Environmental Health* **81**, 165–178.
- [50] American Conference of Governmental Industrial Hygienists (ACGIH) (1955). *Threshold Limit Values for Chemical Substances and Physical Agents for 1955-1956*, Cincinnati.
- [51] American Conference of Governmental Industrial Hygienists (ACGIH) (1959). *Threshold Limit Values for Chemical Substances and Physical Agents for 1959-1960*, Cincinnati.
- [52] Occupational Safety and Health Administration (OSHA) (2001). *Air Contaminants*, 29 CFR 1910.1000, Code of Federal Regulations. Table Z-1 and Z-2.
- [53] American Conference of Governmental Industrial Hygienists (ACGIH) (2004). *Threshold Limit Values for Chemical Substances and Physical Agents and Biological Exposure Indices for 2004*, Cincinnati.
- [54] American Conference of Governmental Industrial Hygienists (ACGIH) (2006). *Threshold Limit Values for Chemical Substances and Physical Agents and Biological Exposure Indices for 2006*, Cincinnati.
- [55] Bayliss, D.L., Lainhart, W.S., Crally, L.J., Ligo, R., Ayer, H. & Hunter, F. (1971). Mortality patterns in a group of former beryllium workers, *Proceedings of the American Conference of Governmental Industrial Hygienists 33rd Annual Meeting*, Toronto, pp. 94–107.
- [56] Mancuso, T.F. & El-Attar, A.A. (1969). Epidemiological study of the beryllium industry. Cohort methodology and mortality studies, *Journal of Occupational Medicine* **11**, 422–434.
- [57] Mancuso, T.F. (1979). Occupational lung cancer among beryllium workers: dust and disease, in *Proceedings of the Conference on Occupational Exposure to Fibrous and Particulate Dust and Their Extension into the Environment*, R. Lemen & J. Dement, eds, Pathotox Publishers, Park Forest South, pp. 463–482.
- [58] Mancuso, T.F. (1980). Mortality study of beryllium industry workers' occupational lung cancer, *Environmental Research* **21**, 48–55.
- [59] Wagoner, J.K., Infante, P.F. & Bayliss, D.L. (1980). Beryllium: an etiologic agent in the induction of lung cancer, nonneoplastic respiratory disease, and heart disease among industrially exposed workers, *Environmental Research* **21**, 15–34.
- [60] Ward, E., Okun, A., Ruder, A., Fingerhut, M. & Steadman, K. (1992). A mortality study of workers at seven beryllium processing plants, *American Journal of Industrial Medicine* **22**(6), 885–904.
- [61] Infante, P.F., Wagoner, J.K. & Sprince, N.L. (1980). Mortality patterns from lung cancer and nonneoplastic respiratory disease among white males in the beryllium case registry, *Environmental Research* **21**, 35–43.
- [62] Steenland, K. & Ward, E. (1992). Lung cancer incidence among patients with beryllium disease: a cohort mortality study, *Journal of the National Cancer Institute* **83**, 1380–1385.
- [63] IARC (1993). *IARC Monographs of the Evaluation of the Carcinogenic Risk of Chemicals to Humans*, World Health Organization, Lyon, Vol. 58.
- [64] International Agency for the Research on Cancer (IARC) (2001). *Overall Evaluations of Carcinogenicity to Humans*, at <http://monographs.iarc.fr/ENG/Classification/crthall.php> (accessed Nov 2007).
- [65] National Toxicology Program (NTP) (1981). *Second Report on Carcinogens*, U.S. Department of Health and Human Services, Research Triangle Park.
- [66] National Toxicology Program (NTP) (2002). *Tenth Report on Carcinogens*, U.S. Department of Health and Human Services, Research Triangle Park, at <http://ntp-server.niehs.nih.gov> (accessed Jul 2002).
- [67] Stokinger, H.E., Spiegel, C.J., Root, R.E., Hall, R.H., Steadman, L.T., Stroud, C.A., Scott, J.K., Smith F.A. & Gardner, D.E. (1953). Acute inhalation toxicity of beryllium IV, *Beryllium fluoride at exposure concentrations of one and ten milligrams per cubic meter*. *AMA Archives of Industrial Hygiene and Occupational Medicine* **8**, 493–506.
- [68] Haley, P.J., Finch, G.L., Hoover, M.D. & Cuddihy, R.G. (1990). The acute toxicity of inhaled beryllium metal in rats, *Fundamental and Applied Toxicology* **15**, 767–778.
- [69] Haley, P.J. (1991). Mechanisms of granulomatous lung disease from inhaled beryllium: the role of antigenicity in granuloma formation, *Toxicologic Pathology* **19**, 514–525.
- [70] Sendelbach, L.E., Witschi, H.P. & Tryka, A.F. (1986). Acute pulmonary toxicity of beryllium sulfate inhalation in rats and mice: cell kinetics and histopathology, *Toxicology and Applied Pharmacology* **85**(2), 248–256.
- [71] Hart, B.A., Harmsen, A.G., Low, R.B. & Emerson, R. (1984). Biochemical, cytological and histological alterations in rat lung following acute beryllium aerosol exposure, *Toxicology and Applied Pharmacology* **75**, 454–465.
- [72] Huang, H., Meyer, K.C., Kubai, L. & Auerbach, R. (1992). An immune model of beryllium-induced pulmonary granulomata in mice. Histopathology, immune

- reactivity, and flow-cytometric analysis of bronchoalveolar lavage-derived cells, *Laboratory Investigation* **67**(1), 138–146.
- [73] Barna, B.P., Deodhar, S.D., Chiang, T., Gautam S. & Edinger M. (1984). Experimental beryllium-induced lung disease. I. Differences in immunologic response to beryllium compounds in strains 2 and 13 guinea pigs, *International Archives of Allergy and Applied Immunology* **73**, 42–48.
- [74] Haley, P., Pavia, K.F., Swafford, D.S., Davila, D.R., Hoover, M.D. & Finch, G.L., (1994). The comparative pulmonary toxicity of beryllium metal and beryllium oxide in cynomolgus monkeys, *Immunopharmacology and Immunotoxicology* **16**(4), 627–644.
- [75] Sanders, C.L., Cannon, W.C., Powers, G.J., Adee, R.R. & Meier, D.M. (1975). Toxicology of high-fired beryllium oxide inhaled by rodents. I. Metabolism and early effects, *Archives of Environmental Health* **30**(11), 546–551.
- [76] Schepers, G.W.H., Durkan, T.M., Delahant, A.B., Smart, R.H. & Smith, C.R. (1957). The biological action of inhaled beryllium sulfate: a preliminary chronic toxicity study on rats, *Archives of Industrial Health* **15**, 32–38.
- [77] Wagner, W.D., Groth, D.H., Holtz, J.L., Madden, G.E. & Stokinger, H.E. (1969). Comparative chronic inhalation toxicity of beryllium ores, bertrandite and beryl, with production of pulmonary tumors by beryl, *Toxicology and Applied Pharmacology* **15**, 10–129.
- [78] Reeves, A.L., Deitch, D. & Vorwald, A.J. (1967). Beryllium carcinogenesis. I. Inhalation exposure of rats to beryllium sulfate aerosol, *Cancer Research* **27**, 439–445.
- [79] Vorwald, A.J. & Reeves, A.L. (1959). Pathologic changes induced by beryllium compounds, *Archives of Industrial Health* **19**, 190–199.
- [80] Reeves, A.L. & Deitch, D. (1969). Influence of age on the carcinogenic response to beryllium inhalation, in *Proceedings of the 16th International Congress on Occupational Health*, S. Harishima, ed, Japan Industrial Safety Association, Tokyo, pp. 651–652.
- [81] Nickell-Brady, C., Hahn, F.F., Finch, G.L. & Belinsky, S.A. (1994). Analysis of K-ras, p53 and c-raf-1 mutations in beryllium-induced rat lung tumors, *Carcinogenesis* **15**, 257–262.
- [82] Vorwald, A.J. (1968). Biologic manifestations of toxic inhalants in monkeys, in *Use of Nonhuman Primates in Drug Evaluation: A symposium. Southwest Foundation for Research and Education*, H. Vagrborg, ed, University of Texas Press, Austin, pp. 222–228.
- [83] EPA (1998). *Toxicological Review of Beryllium and Compounds*, U.S. Environmental Protection Agency, Washington, DC.
- [84] Morgareidge, K., Cox, G.E. & Gallo, M.A. (1976). *Chronic Feeding Studies with Beryllium in Dogs*, Food and Drug Research Laboratories, Submitted to the Aluminum Company of America, Alcan Research and Development, Kawecki-Berylco Industries, and Brush-Wellman.
- [85] Madl, A.K., Unice, K., Brown, J.L., Kolan, M.E. & Kent, M.S. (2007). Exposure-response analysis for beryllium sensitization and chronic beryllium disease among workers in a beryllium metal machining plant, *Journal of Occupational and Environmental Hygiene* **4**(6), 448–466.
- [86] Kanarek, D.J., Wainer, R.A., Chamberlin, R.I., Weber, A.L. & Kazemi, H. (1973). Respiratory illness in a population exposed to beryllium, *The American Review of Respiratory Disease* **108**, 1295–1302.
- [87] Sprince, N.L., Kanarek, D.J., Weber, A.L., Chamberlin, R.H. & Kazemi, H. (1978). *Reversible Respiratory Disease in Beryllium Workers*, *American Review of Respiratory Disease* **117**, 1011–1017.

Further Reading

- Ayer, H. & Hunter, F. (1971). Mortality patterns in a group of former beryllium workers, *Proceedings of the American Conference of Governmental Industrial Hygienists 33rd Annual Meeting*, Toronto, 94–107.
- Schepers, G.W.H. (1964). Biological action of beryllium: reaction of the monkey to inhaled aerosols, *Industrial Medicine and Surgery* **33**, 1–16.
- Stokinger, H.E. (1981). In *Patty's Industrial Hygiene and Toxicology*, 3rd Edition, G. Clayton & D. Clayton, eds, John Wiley & Sons, New York, Vol. IIA, pp. 1537–1558.
- Stokinger, H.E., Sprague, G.F., Hall, R.H., Ashenburg, N.J., Scott, J.K. & Steadman, L.T. (1950). Acute inhalation toxicity of beryllium. I. Four definitive studies of beryllium sulfate at exposure concentrations of 100, 50, 10 and 1 mg per cubic meter, *Archives of Industrial Hygiene and Occupational Medicine* **1**, 379–397.

Related Articles

Air Pollution Risk

Detection Limits

Environmental Carcinogenesis Risk

FINIS CAVENDER

Asset–Liability Management for Life Insurers

Asset–liability management or asset–liability modeling (ALM) for a life insurer is a quantitative approach to assessing risks in both the liabilities and the assets and developing risk-management strategies incorporating asset allocations, capital provisions, derivatives, and reinsurance. An important objective is ensuring asset cash flows meet liability cash flows over long terms. ALM for a life insurer requires valuation and projection of complex cash flows contingent on a wide range of risks including human mortality, sickness, interest rates, equity returns, inflation, and exchange rates. With the increase in computing power, ALM has become a powerful risk assessment technique for life insurers.

Products and Risks of Life Insurers

Life insurance companies provide coverage for risks related to human mortality, survival, as well as health and disability. They pool large numbers of life insurance risks (*see Copulas and Other Measures of Dependency*) and as a result benefit from the law of large numbers in the variability and prediction of expected future claim payments. They also issue contracts that provide a long term savings benefit linked to mortality risks. Life insurance contracts are for long terms (*see Fair Value of Insurance Liabilities*) compared to nonlife insurance. They involve guarantees to make payments for many years in the future. Whole of life insurance and annuities involve payments depending on death or survival for as long as the oldest age of survival of the lives insured. This is the ultimate age in the life table for risk modeling purposes. For example, an annuity could involve a commitment to make payment to an age of 120 in the case of a life annuity. Obviously, such occurrences are relatively rare events, but improvements in mortality and also pandemics can impact mortality risk across a range of ages. Term insurance contracts provide cover for periods as long as twenty or thirty years. Long term savings products usually have terms of at least 10 years.

Life insurance contracts can be participating with-profits, with a bonus payable on a regular basis depending on the surplus arising on the insurance business (*see Bonus–Malus Systems*). Once the bonus is paid it becomes a guaranteed part of the life insurance contract, adding to the obligations of the company. Bonuses may also be paid on maturity or termination of the contract, in which case they are referred to as *terminal bonuses*. For nonparticipating contracts the payments are mostly fixed or linked to an index such as the consumer price index (CPI) in the case of indexed annuities. The amount of the payment can be estimated with a high degree of confidence. For a large number of lives insured, the expected claims payments on death or survival can be estimated. Mortality improvements can cause some uncertainty in the estimate of future payments, but an allowance is usually made for this.

Even though the policies are issued for long terms, in many cases, the policies terminate owing to lapse or surrender prior to the maturity date of the policy. A policy lapses when a policyholder does not pay the premium and no contractual payment is owed to the policyholder. A policyholder can also surrender their policy prior to the maturity date of the policy and receive a payment on the policy. This payment is called the *surrender value*. Often the surrender value is guaranteed. A minimum surrender value is usually required by life insurance law in many countries. Maturity benefits on savings products and also death benefits on life insurance contracts may also be guaranteed in the form of an investment guarantee. For example, a guarantee on a life insurance policy could be a minimum benefit of a return of premiums with a specified and fixed interest rate in the event of death or surrender after a minimum term. This is often a requirement of life insurance legislation in many jurisdictions. Guarantees in life insurance policies have the same features as financial option contracts such as puts and calls although they are far more complex since they depend on investment returns as well as mortality and withdrawals and extend over very long terms.

Lapses and surrenders are hard to forecast because they can depend on future economic conditions. For example, if there is a recession, then policyholders with savings policies may need to surrender the policies to access their savings to cover short-term needs. If earnings rates offered on competitor savings

products are higher, then this can also lead to surrenders of these policies.

Life insurance companies receive premiums from policies and invest these funds in assets in order to satisfy claim payments as and when they fall due. Assets include short-term cash investments, fixed interest securities, shares, and property investments, as well as various alternative investments such as hedge funds and insurance linked securities. They also insure their own risks by purchasing reinsurance (*see* **Reinsurance**) from specialist reinsurance companies.

ALM for Interest and Mortality Risk

Penman [1] noted the problems that arise in asset–liability management for life insurers resulting from too generous guaranteed surrender values. He also noted the increased importance of currencies, commodities, inflation, and income tax in asset–liability management in the early 1900s. He discussed the idea of matching the currency of the assets and the currency of the life insurance contracts and the role of property and ordinary shares in matching inflation. The importance of diversification by geographical area and type of security was also recognized. C. R. V. Coutts, in the discussion of the Penman paper, introduced an asset–liability management strategy known as *matching*, by which he meant holding investments that were repayable at a time to cover contract payments from the fund over the following forty or fifty years.

However, it was Redington [2] who developed the principle of immunization of a life insurance company against interest rate movements. Assets were to be selected so that the discounted mean term or duration of the asset and liability cash flows were equal, and that the spread of the assets cash flows around their discounted mean terms, referred to as convexity, should exceed the spread of the liabilities around their mean term (*see* **Statistical Arbitrage**). These ideas became the foundation of life insurer ALM for many years until the adoption of more advanced option pricing and modeling techniques following the Nobel prize winning work of Black and Scholes [3]. These risk modeling and assessment approaches have been extended in more recent years. The early ALM for life insurers was designed for interest rate risk management and fixed- interest or

interest sensitive assets and liabilities. As noted in Sherris [4], matching can be considered as a special case of asset–liability portfolio selection under risk and capital constraints for the life insurer. Panjer [5] covers ALM for interest rate risk and ALM portfolio selection of assets to allow for liabilities.

Modern ALM for life insurers extends beyond interest rate risk and measuring the duration of assets and liabilities to a broader range of risks that can adversely impact the financial performance of the company. Immunization techniques were developed for fixed cash flows whose values were dependent on interest rates. Modern option pricing models allowed these to be extended to interest sensitive cash flows.

Over the last decade or more, longevity has become an important risk to be assessed by life insurers, particularly for life annuity and pension policies. As individual life expectancy increases, the value of life annuities increases. Selecting assets to manage and match longevity risk is difficult since markets in long term securities are limited and there is only a limited market in mortality risk, unlike the markets for credit, interest rate and equity risk. Securitization is an increasingly important risk management technique in ALM for life insurers and mortality risk. The importance of mortality risk is highlighted by the rapidly growing collection of research related to this topic including ALM issues.

Quantifying Risks for Life Insurers

The financial risks of a life insurer are quantified using specialized ALM projection software to model and project both the assets and the liabilities. Since the liabilities consist of many individual policies, it is common to group those of similar type, such as term and whole of life insurance, as well as by age-group and other characteristics, to project the liability cash flows. Asset cash flows, including interest and dividends, as well as maturity payments are often projected for portfolios including shares, property, and fixed interest. The projections use an economic scenario generator that generates future distributions for interest rates, inflation, equity and property returns, and liability cash flows allowing for mortality, lapses, surrenders, and maturities, as well as the benefit guarantees.

Modern financial concepts and techniques from option and contingent claim valuation theory are

used to derive market consistent valuations of assets and liabilities by valuing projected cash flows on a risk adjusted basis. The models are also used to assess risk based capital and reinsurance requirements, both very important components of the risk management strategy of a life insurer by simulating future scenarios for the business and determining the scenarios where asset cash flows are insufficient to meet liability cash flows. Simulation (*see Value at Risk (VaR) and Risk Measures*) of future scenarios is commonly used in these projections, although the computational time taken to perform a complete ALM risk assessment requires the use of modern and fast methods of computation, including quasi-random numbers and other variance reduction methods that can produce accurate estimations with fewer scenarios.

Since most life insurance products contain guarantees to make future fixed or dependent payments, as well as other option features, standard cash flow projection techniques will not necessarily assess the risk in the cash flows, unless specific allowance is made for these options. Stochastic projections incorporating future random outcomes are essential in order to quantify adverse scenarios, which can often occur when policyholders exercise valuable options in their policies, and at the same time, options in the asset cash flows can have an adverse financial impact; for example, policyholders who exercise surrender options when interest rates are high and there are valuable alternative investments, and at the same time asset values drop. The actual valuation of life insurer obligations allowing for the guarantees and options will usually substantially exceed the value ignoring these options. ALM can be used to project the outcomes where the life insurer is most at risk and to also quantify the cost of managing or reducing the risk.

Dynamic modeling, including stochastic optimization, has the potential to improve the risk assessment and management of life insurers. These techniques

are used increasingly in ALM for nonlife insurers (*see Numerical Schemes for Stochastic Differential Equation Models; Continuous-Time Asset Allocation; Bayesian Analysis and Markov Chain Monte Carlo Simulation; Optimal Stopping and Dynamic Programming*). The long term nature of life insurance cash flows presents new challenges for the application of these techniques.

Conclusion

ALM is a critical component of a quantitative risk assessment for financial performance; it ensures that a life insurer can meet policy obligations. Risk assessment modeling has developed over a number of years to highly sophisticated cash flow projection and valuation models incorporating modern developments in financial risk modeling and valuation. A detailed coverage of many of these techniques is found in Ziemba and Mulvey [6].

References

- [1] Penman, W. (1933). A review of investment principles and practice, *Journal of the Institute of Actuaries* **LXIV**, 387–418.
- [2] Redington, F.M. (1952). Review of the principles of life office valuations, *Journal of the Institute of Actuaries* **78**, 286–315.
- [3] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**(3), 637–654.
- [4] Sherris, M. (1992). Portfolio selection and matching: a synthesis, *Journal of the Institute of Actuaries* **119**(Pt 1), 87–105.
- [5] Panjer, H.H. (ed) (1998). *Financial Economics: With Applications to Investments, Insurance, and Pensions*, The Actuarial Foundation, Schaumburg.
- [6] Ziemba, W.T. & Mulvey, J.M. (1998). *Worldwide Asset and Liability Modelling*, Cambridge University Press.

MICHAEL SHERRIS

Asset–Liability Management for Nonlife Insurers

In insurance, *asset–liability management* (ALM) is often treated as synonymous with *duration matching*. Matching refers to establishing and maintaining an investment portfolio – typically fixed income – that has the same average duration or even the same cash flow timing as the liability portfolio. The aim of a matching strategy is to protect the net value of the firm from interest rate changes, so that they affect the discounted values of assets and liabilities equally. Modern *enterprise risk analysis* (see **Enterprise Risk Management (ERM)**) and management provides the tools to design an investment strategy that reduces risks beyond interest rate hedging.

We define ALM to be a comprehensive analysis and management of the asset portfolio in light of current liabilities and future cash flows of a going-concern company, incorporating existing asset and liability portfolios as well as future premium flows. Besides duration matching, ALM considers additional risk factors beyond interest rate changes such as inflation risk, credit risk, and market risk. ALM also considers actions beyond the characteristics of a fixed income portfolio and seeks to identify and exploit hedges of any sort. For example, equities or real estate may provide a partial hedge against inflation-sensitive liabilities, at least over a long-enough time frame. *Reinsurance* (see **Reinsurance**), in this context, is a form of a hedge.

An example is foreign exchange risk. A typical hedging strategy is to keep assets in each currency equal to the liabilities, including unearned premiums. This is effective when the liabilities are known, but when they are random variables the mean of the liabilities might not be the proper target to hedge. Keeping some portion of capital in each currency might help reduce foreign exchange risk, and enterprise-wide, internal capital modeling can help evaluate the risk impact as part of an enterprise risk-management (ERM) program.

Insurance companies can benefit from the discipline of a more integrated analysis of the asset and liability portfolios in seeking better risk-return decisions. An enterprise-wide analysis of potential

risks and rewards affords an ideal opportunity to analyze the company's investment portfolio and underwriting portfolio in concert. Since insurance liabilities are far less liquid than assets, such analysis and management activity tend to focus on adjustments to the investment portfolio, given the constraints of the reserves and underwriting portfolio, to improve the risk-return characteristics of, say, annual earnings or a terminal (future) value of surplus. In this respect assets can be thought of as a way to hedge liability risk. However, management activity need not be confined to fine-tuning investment strategy. Future underwriting considerations, along with other hedges such as reinsurance, are risk-management variables at their disposal.

Venter *et al.* [1] presented a series of simple numerical examples illustrating that the optimal risk-return portfolio decisions are very different as the asset and liability considerations become more realistic and complex. The authors started with a stand-alone asset portfolio, then in a series of adjustments added a constant fixed duration liability, a liability that varied as to time and amount, and then added consideration of cash flows from current underwriting. As the various layers of complexity are added to the illustration, the nature of the inherent risks changes, as does the optimal investment portfolio:

- Looking at assets in isolation, short-term treasuries are considered risk-free, while higher yielding assets – stocks and bonds – are considered riskier. Traditional portfolio analysis would look at combinations of assets and measure their expected mean and variance, plotting return versus risk and searching out alternatives on an efficient frontier.
- Venter *et al.* point out that, when fixed liabilities are considered, holding shorter-term (that is, shorter than the liabilities) assets creates a new risk – a reinvestment risk. If interest rates drop, total investment income may prove insufficient to cover the liabilities. If interest rates increase, longer-term investments, too, present a new risk, if depressed assets need to be liquidated to fund liabilities. The risk to net assets, or surplus, can be reduced by cash flow matching. However there is a risk-return trade-off, as longer bonds usually have higher income. The accounting system in use may hide some of the risk as well.

- Adding in the complexity of liabilities that are variable as to amount and timing of cash flows makes precise cash flow matching impossible or transitory at best. Inflation-sensitive liabilities add additional complexity. A model incorporating both asset and liability fluctuations over time is required at this point to seek out optimal investment strategies.
- To make matters more difficult still, Venter *et al.*, introduce the notion of a company that is a going concern, with variable (positive or negative) cash flow from underwriting. Going concerns have greater flexibility. If, for example, conditions for liquidation are unfavorable, the company could pay claims from premium cash flows. At this level of complexity, an enterprise-wide model is truly needed, because, in addition to asset and liability models, a model of the current business operation is needed, including premium income, losses (including catastrophic losses), expenses, etc.

Venter *et al.*, did not address tax considerations, which can also have a profound impact on investment decisions. Recent studies have found that insurers consider cyclical changes in the investment portfolio between tax-exempt and taxable fixed income securities over the course of the underwriting cycle to be one of the principle drivers in investment strategy (as underwriting losses absorb taxable investment income). In addition to the integration of underwriting and investment results, such strategies rely on reallocation of assets to maximize income while avoiding alternative minimum taxes (AMT).

Little has been said up to this point about equity investments, but consideration of equities, too, adds some complexity and richness to the asset–liability analysis. Equities are considered risky in their own right and will imply a potentially worse downside risk to capital. Some believe that equities may provide a better inflation hedge for liabilities in an increasing loss cost environment. This proposition may be tested through the enterprise risk model, although the conclusion will be sensitive to input assumptions of the incorporated macroeconomic model.

In 2002, the Casualty Actuarial Society Valuation, Finance, and Investment Committee (VFIC) published a report [2] testing the optimality of duration matching (between assets and liabilities) investment strategies for insurance companies. To address this question, VFIC applied a simulation model to a

variety of scenarios – long-tailed *versus* short-tailed (with cat exposure), profitable *versus* unprofitable, and growing *versus* shrinking business. In doing so, VFIC attempted to tackle Venter’s most complex scenario discussed above.

Where Venter *et al.*, focused on changes in GAAP pretax surplus changes as the risk measure, VFIC looked at several different risk measures on both a statutory and a GAAP basis. Return, too, was considered on both accounting bases. In doing so, VFIC’s conclusion as to optimality was what one might expect in the real world: it depends. In fact, duration matching was among a family of optimal strategies, but the choice of specific investment strategies was dependent on the company’s choice of risk metrics, return metrics, and risk-return tolerances or preferences:

- Statutory accounting-based metrics implied little hedge from duration matching, as bonds were amortized and liabilities were not discounted.
- GAAP accounting-based metrics resulted in similar conclusions: though bonds were marked to market, there was no hedge to the liabilities.
- If metrics are calculated based on “true economics”, that is with bonds marked to market and liabilities discounted at current market rates, matching produces a low interest rate risk. In this case, a short investment strategy increases risk (creates mismatch in duration) and decreases return. Longer-duration investment strategies increase risk and return, making the trade-off more of a value judgment. In the economic case, the introduction of cash flows from operations greatly complicated the analysis and conclusions. However, this analysis did not consider the impact of inflation-sensitive liabilities.

An Asset–Liability Modeling Approach

It has been asserted earlier that an enterprise-wide model is the ideal, perhaps the only, way to model and ultimately manage an insurance company investment portfolio. ALM makes for an excellent application of such an integrated model.

The introductory comments highlight a series of modeling steps and management considerations that supply the necessary structure to the analysis of an insurer’s investment portfolio, given the liabilities it supports.

1. Start with models of asset classes (stocks, bonds, real estate, commodities), existing liabilities (loss reserves, receivables), and current business operations.
2. Define risk metric(s) for the analysis. Consideration should be given to accounting basis – statutory, GAAP, or economic. Risk can be defined on the basis of the volatility of periodic income or the volatility of ending surplus or equity. Decisions are therefore also necessary as to the time frames of the analysis (*see Axiomatic Measures of Risk and Risk-Value Models*).

Examples of risk metrics – either income based or balance sheet based – include the standard deviation and the probability of falling below a predetermined threshold. Balance sheet (surplus or equity) measures are more amenable to other metrics as well, such as value at risk, tail value at risk, *probability of ruin* (*see Ruin Probabilities: Computational Aspects*), or probability of impairment.

3. Similarly, management must define what constitutes “return”. Again, consideration must be given to accounting basis, and clearly risk and return must be defined in a compatible fashion. Return, too, can be income based or balance sheet based. For example, return can be periodic earnings or return on equity (ROE) (income measures), or it can be a terminal value of surplus (balance sheet).
4. Consideration must be given to the time horizon of the analysis and the relevant metrics. Single-period models are perhaps simpler, but may not adequately reflect the true nature of the business throughout a cycle. Multiperiod models can be more sophisticated – but also more difficult and complicated – especially if cycles and serial correlations are incorporated throughout.
5. The model will have to consider relevant constraints. For example, constraints might include limits on asset classes imposed by state regulators, or investments that drive rating agency or regulatory required capital scores too high, or restrictions based on the company’s own investment policy.
6. The model should be run for a variety of investment strategies, underwriting strategies, and reinsurance options under consideration. For each combination of scenarios, there will be thousands

of realizations. The selected risk and return metrics are calculated over these simulations.

7. An efficient frontier – a plot of the return metric versus the risk metric – can be constructed across the various portfolio scenarios (*see Figure 1*). The point in the risk-return space that defines the current portfolio should be represented. Portfolio moves should be explored where risk can be decreased without sacrificing return (A), where return can be increased without increasing risk (B), and points in between (C). Points on the frontier are considered optimal combinations of returns for given levels of risk. While a point on the curve should be an investment portfolio goal for the company, the selection of one point over another is dependant on the company’s preferences as defined by risk tolerances and return requirements.
8. It was noted at the outset of the process description that, since liabilities are more illiquid, the asset-liability analysis and management can be largely asset centric given the existing liabilities. The nature of the liabilities, however, can be adjusted – risks can be hedged – through reinsurance purchases. The effects of such purchases can have profound impacts on an ALM analysis, especially in a multiperiod model.

Various reinsurance structures should be modeled with the alternative asset portfolio options and the results compared. For example, Figure 2 compares the results of reinsurance and asset allocation decisions in the worst 1% of the simulations over time. The vertical axis represents company surplus, and the horizontal axis represents time. In this particular example, the analysis

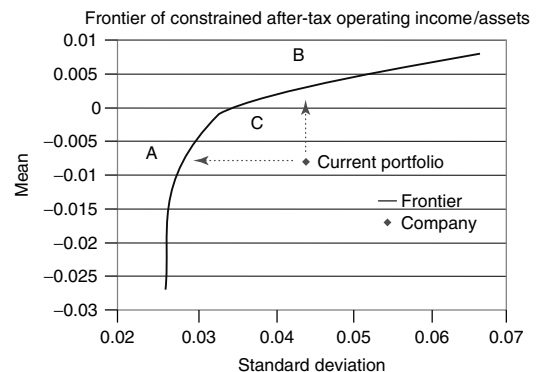


Figure 1 Efficient frontier

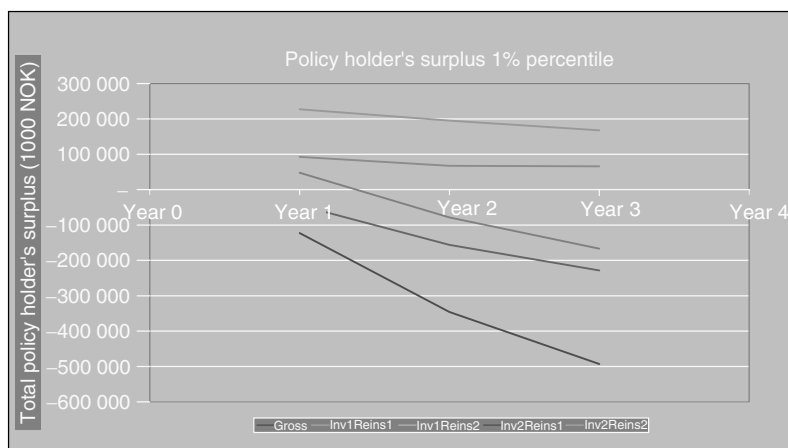


Figure 2 Reinsurance and asset allocation metrics

was conducted over a 3-year window. Beyond the gross (no reinsurance) option, the second and fourth options show better and more stable ending surplus values in the worst 1% of the simulations.

9. Having selected a targeted point on an efficient frontier and a companion reinsurance strategy, simulation output should be analyzed to identify those scenarios where even the preferred portfolio(s) performed poorly (since Figures 1 and 2 are drawn only at single points on the distribution of possible results). It is possible that an asset hedging strategy can be employed to reduce downside risks. It may also be the case that further investigation will highlight the type of prevailing conditions that can lead to substandard performance, for example, a large catastrophe that forces liquidation of assets, or a persistent soft market. Management can establish policy and monitoring mechanisms to identify the existence or the likelihood of such conditions and make appropriate adjustments.

Future Research

While enterprise-wide modeling is perhaps the only way to adequately address asset–liability management issues, there are a number of real-world issues that are subject of continuing research. For example, correlations among lines of insurance (either in current business or in past reserves), between assets and liabilities, and over time are poorly understood today.

And yet correlations can materially alter the risk of the optimal portfolio. Also, models of unpaid losses, while they can derive expected values and distributions, have not been developed as explanatory models. That is, unlike asset models, reserving models do not predict future loss payments with parameters linking losses to economic indices. Inflation sensitivity is often hypothesized on an accident year, a calendar year, or a payment year basis, but rarely explicitly developed from historic economic data and projected on the basis of, say, an economic scenario generator.

References

- [1] Ashab, M.Q., Bushel, A., Gradwell, J.W. & Venter, G.G. (1998). Implications of reinsurance and reserves on risk of investment asset allocation, *Casualty Actuarial Society Forum Summer*, 221–272.
- [2] CAS Valuation, Finance, and Investment Committee (2002). Interest rate risk: an evaluation of duration matching as a risk-minimizing strategy for property/casualty insurers, *Casualty Actuarial Society Forum Summer*, 135–168.

Related Articles

Correlated Risk

Dependent Insurance Risks

Large Insurance Losses Distributions

Ruin Probabilities: Computational Aspects

Statistical Arbitrage

PAUL BREHM AND GARY VENTER

Association Analysis

Association analysis aims to determine whether a *disorder* of interest correlates with *genetic variation* in the population. Such a correlation indicates that either the surveyed genetic variation predisposes to disease or acts as a proxy for the causal variant (see the section titled “Linkage Disequilibrium”). In this article, we provide a short introduction to association at a conceptual level, both analytically and genetically. Next we consider the two primary association approaches for mapping disease loci – case–control and family-based studies. The article concludes with recent developments in association, such as genome-wide association (GWA) that hold great promise for mapping of genetic variation related to common disease.

Conceptual Underpinnings

For association analysis to be conducted, a direct assay of genotypic variation (i.e., to know the genetic code at the locus of interest for all individuals in the analysis) is typically required. While many types of genetic variation exist in the human genome, the most commonly assayed for association are single nucleotide polymorphisms (SNPs). SNPs are composed of two *alleles* (forms) within the population. Given that humans are diploid (i.e., their chromosomes are organized in pairs, having inherited one of each pair from their mother and the other from their father), every individual has two copies of the SNP at a locus; this pair of SNPs constitutes the *genotype* at that locus. For didactic purposes, we restrict our considerations to SNPs, but the association techniques presented have been extended to genetic loci at which there are multiple alleles (e.g., microsatellites [1]).

Considering a SNP marker with two alleles (labeled A_1 and A_2), there are three possible genotypes (A_1A_1 , A_1A_2 , and A_2A_2). The genotypes with two copies of the same allele (i.e., A_1A_1 and A_2A_2) are termed *homozygotes* and the genotype with a copy of each allele is termed the *heterozygote* (A_1A_2). Given a sample of genotypes from a population of interest, it is possible to estimate the proportions of the alleles A_1 and A_2 (which are typically labeled p and q , respectively, with $q = 1 - p$), as well as the proportions of the genotypes A_1A_1 , A_1A_2 ,

and A_2A_2 . The terms *allele frequency* and *genotype frequency* are in common use for these quantities in genetics, and this convention will be used here. The expected frequencies for the genotype categories may be derived by assuming Hardy–Weinberg equilibrium (HWE), which occurs when mating in the population is random and the genotypes are equally viable [2, 3]. These expected frequencies are p^2 , $2pq$, and q^2 for A_1A_1 , A_1A_2 , and A_2A_2 , respectively. Testing for a deviation from HWE provides a good check for the quality of the genotype information and evidence for confounding or assortative (nonrandom) mating. However, HWE tests are usually restricted to controls, because deviation from HWE is conflated with true association signal in cases [4].

Case–Control Association Tests

With the genotype information available, it is possible to define a range of association tests for cases and control subjects. Much like any comparison between cases and controls, for a discrete predictor, a one degree of freedom Pearson’s χ^2 test on allele counts may be calculated from the data organized in Table 1.

Let the rows of Table 1 be indexed by i (where $i = 1$ represents cases and $i = 2$ represents controls), and the columns of by j (where $j = 1$ corresponds to A_1 and $j = 2$ corresponds to A_2). Then χ^2 is then calculated as:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(N_{ij} - M_{ij})^2}{M_{ij}} \quad (1)$$

where: N_{ij} is the allele count in row i , column j of Table 1; and M_{ij} is computed as the product of the sum of the elements in row i and the sum of the elements in column j , divided by N_{total} ($M_{ij} = N_{i*} * N_{*j} / N_{\text{total}}$, where $N_{*j} = N_{1j} + N_{2j}$; $N_{i*} = N_{i1} + N_{i2}$). It is also possible to calculate a two degree of freedom Pearson’s χ^2 from a table of genotype counts instead of allele counts.

The allelic test effectively employs an additive model, whereas the genotype test is saturated, because the three genotype categories are specified separately. It is also possible to use the information in Table 2 to estimate the odds ratio (see **Odds and Odds Ratio**) or increase in disease risk conferred by the risk allele in cases as compared to

Table 1 Notation for counts of alleles A_1 and A_2 in cases (d_1 and d_2 , respectively) and in controls (u_1 and u_2 , respectively)

	Allele counts		Total
	A_1	A_2	
Case	d_1	d_2	N_{case}
Control	u_1	u_2	N_{control}
Total	N_{A_1}	N_{A_2}	N_{total}

Table 2 Notation for counts of genotypes A_1A_1 , A_1A_2 , and A_2A_2 : $d_{11}d_{12}d_{22}$ are the respective genotype counts for cases, and u_{11} , u_{12} , and u_{22} are the corresponding genotype counts for controls

	Genotype counts			Total
	A_1A_1	A_1A_2	A_2A_2	
Case	d_{11}	d_{12}	d_{22}	N_{case}
Control	u_{11}	u_{12}	u_{22}	N_{control}
Total	N_{11}	N_{12}	N_{22}	N_{total}

controls:

$$OR = \frac{[(d_{12} + d_{22}) \times u_{11}]}{[d_{11} \times (u_{12} + u_{22})]} \quad (2)$$

The prevalence or the proportion of the population that is affected can be defined in terms of the *penetrance* of each genotype class:

$$K = p^2f_2 + 2pqf_1 + q^2f_0 \quad (3)$$

where K is the prevalence, p and q are the allele frequencies for A_1 and A_2 , respectively, and f_2 , f_1 , and f_0 represent the probability of expressing the trait for genotype categories A_1A_1 , A_1A_2 , and A_2A_2 . In the event of no association, $f_2 = f_1 = f_0 = K$ [5].

An alternative to these Pearson's χ^2 tests is to specify a logistic regression (*see Logistic Regression*) model for association analysis. For this regression model, the alleles for each individual are coded as an indicator variable, so the model is

$$\text{logit}(y_i) = \alpha + \beta x_i \quad (4)$$

where y_i is the case or control status, α is the intercept, β is the regression coefficient for the locus, and x_i is an indicator variable for the number of copies of A_2 . Note that only one of the two alleles is coded, because the second is linearly dependent upon

it (i.e., it is collinear). Essentially, this formulation specifies a multiplicative risk model for the effect of the allele in the population (i.e., the relative risk for genotype $A_1A_1 = R \times A_1A_2 = R^2 \times A_2A_2$).

An alternative approach is to parameterize the model so that the risk for each genotype is estimated directly:

$$\text{logit}(y_i) = \alpha + \beta_1 x_i + \beta_2 w_i \quad (5)$$

where y_i is the case or control status, α is the intercept, β_1 is the regression coefficient for A_1A_1 homozygote, x_i is an indicator variable coded 0/1 for the presence or absence of genotype A_1A_1 , β_2 is the regression coefficient for A_2A_2 homozygote, and w_i is an indicator variable coded 0/1 for the presence or absence of genotype A_2A_2 .

As noted, the first regression model follows a multiplicative genetic model, in that it is based on the number of alleles and does not consider the interaction between the alleles at the locus in question. However, it is possible to specify a model where the allelic effects operate according to a simple dominant or recessive mode of gene action. An allele is said to be dominant when only one copy of it is required for it to affect the phenotype; it is said to be recessive when two copies are required. Table 3 shows how models for dominant and recessive gene action may be parameterized.

In general, the regression models are preferable for association testing. Their main benefits are that (a) they are computationally efficient, (b) it is straightforward to incorporate other loci or covariates, and (c) it possible to test for interactions between the alleles at a locus, between different loci, or between covariates (such as environmental exposure) and alleles.

Table 3 Genotype frequency and penetrance patterns under different modes of gene action

	A_1A_1	A_1A_2	A_2A_2
Genotype frequency	p^2	$2pq$	q^2
Penetrance	f_2	f_1	f_0
Assuming recessive gene action	0	0	1
Assuming dominant gene action	1	1	0
Assuming multiplicative gene action	x^2	x	0

Population Stratification

One potential pitfall of the case–control study is population stratification or substructure. Between populations, allele frequencies tend to differ at a range of loci because of a phenomenon known as *genetic drift*. When such variation between populations is coupled with differences in the prevalence of a disease, spurious association may result. For example, consider the binary trait of “ever used chopsticks”. If cases and controls were sampled from a population consisting of American and Chinese individuals, then significant association would be likely to be found at any locus where the allele frequencies differ between these two groups [6]. This situation is illustrated in Table 4.

Locus A shows no association in either sample, yet there is highly significant association when the two datasets are combined and analyzed jointly. Also note that stratification might eliminate a genuine allelic effect, and cause a false negative finding. To control this particular problem, family-based association designs were proposed.

Family-Based Association

Unlike case–control designs for the identification of risk variants, family-based designs employ auto-control mechanisms. The first family-based design proposed was the transmission disequilibrium test (TDT), which utilizes Mendel’s law of segregation. The key principle is that the transmission of either allele from a heterozygous parent (i.e., genotype A_1A_2) is equally likely. However, affected offspring should be more likely to inherit risk alleles than non-risk. Thus, it is possible to test for overtransmission of risk variants using the McNemar χ^2 , as the transmission or nontransmission from heterozygous parents is a paired observation (see Table 5 and Figure 1).

Table 5 Transmission and nontransmission of alleles from heterozygous parents to affected offspring

	A_1	A_2
Transmission	t_1	t_2
Nontransmission	nt_1	nt_2

Note that $t_1 = nt_2$ and $t_2 = nt_1$

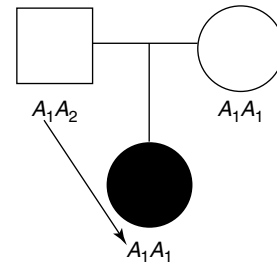


Figure 1 Example parent-offspring trio showing transmission of the risk allele

The test is computed using the formula

$$\chi^2 = \frac{(nt_1 - t_1)^2}{nt_1 + t_1} = \frac{(nt_2 - t_2)^2}{nt_2 + t_2} \quad (6)$$

Allele A_1 is transmitted from the heterozygous father to the affected daughter. If no association is present, A_1 and A_2 would have equal chance of transmission.

The TDT is robust to population stratification, because it only considers parents of a single genotype (i.e., heterozygotes).

Although the TDT may seem to be applicable only when parental genotypes are available, it is possible to use the same principle in the study of sibling pairs [7, 8]. Conceptually, if two siblings are discordant for the phenotype, then the affected sibling ought

Table 4 Demonstration of the effects of population stratification

Use of chopsticks									
	Sample 1 (Americans); $\chi^2 = 0, p = 1$			Sample 2 (Chinese); $\chi^2 = 0, p = 1$			Sample 3 (Americans + Chinese); $\chi^2 = 34.2, p = 4.9 \times 10^{-9}$		
A	Yes	No	Total	Yes	No	Total	Yes	No	Total
A_1	320	320	640	320	20	340	640	340	980
A_2	80	80	160	320	20	340	400	100	500
Total	400	400	800	640	40	680	1040	440	1480

to harbor more risk variants than the unaffected sibling. As the siblings are from the same parents, then they are *de facto* from the same stratum of the population. Thus, population stratification cannot be a problem. The basic test of significance for the sib-TDT employs a permutation test. The genotypes of a pair of sibships are randomly reassigned and the difference in marker allele frequencies between case and control siblings is tested. This procedure is repeated a large number of times to determine the empirical distribution of the test. Then, the true sample difference test statistic (i.e., when no randomization has been performed) is compared to the empirical distribution to determine its significance. The sib-TDT and the TDT information have been combined together into family-based association test [9–11] and the pedigree disequilibrium test [12].

Future Directions

Technologically, genetics is moving forward at an extremely rapid pace. Over the course of approximately the last decade, three main projects have pushed association analysis to the forefront of genetics: the Human Genome Project (HGP), The SNP consortium, and the international HapMap project (HapMap) [13, 14]. The HGP generated a consensus sequence, facilitating the SNP consortium the aim of which was SNP discovery. With a database of known SNPs, the HapMap project set out on identifying linkage disequilibrium (LD) information enabling further development of cost-effective genotyping platforms.

Linkage Disequilibrium

LD refers to the nonrandom distribution of alleles at two different loci. In effect, the genotype at one locus may be used to predict, to some extent, the genotype at the second locus [15]. This association between loci may be considered as a correlation between the alleles at two loci. A consequence of this correlation is that an association test conducted on a marker in LD with the causal variant will also show association. There are two main statistics for LD: D' [16]; and r^2 [17]. D' generally tracks with recombination rate, while r^2 is more appropriate for determining power for association. Both D' and r^2 aim to detect a deviation from the expected frequency of each haplotype (the allele at both loci

on the same strand of DNA) configuration from the product of the allele frequencies. Considered another way, these LD measures equal zero for a pair of loci (A_1/A_2 and B_1/B_2), when the probability of haplotypes $p_{A_1B_1} = p_{A_1} \times p_{B_1}$, $p_{A_1B_2} = p_{A_1} \times p_{B_2}$, $p_{A_2B_1} = p_{A_2} \times p_{B_1}$, and $p_{A_2B_2} = p_{A_2} \times p_{B_2}$.

A useful consequence of LD is that it is not necessary to type all variations across the genome to capture variation. In 2002, the HapMap project was launched as an international collaboration between the National Institutes of Health in the United States, Canada, Japan, China, the SNP consortium and the Wellcome trust [18]. Originally, the HapMap project aimed to identify and tag “haplotype blocks” that were observed across the genome [19]. Three populations were genotyped: an East Asian, comprised 45 unrelated Han Chinese individuals from Beijing and 45 unrelated Japanese individuals from Tokyo; a European, comprised 30 families of two parents and a single offspring from the *Center d'Etude du Polymorphisme Humain* (CEPH) collected in Utah; and a West African, comprised 30 families of two parents and a single offspring drawn from the Yoruba people of Ibadan, Nigeria.

Initially, HapMap had two main genotyping strategies for the three populations:

1. Densely type 10 “ENCODE” regions of 5 megabases (5 mb, 5000 base pairs) to gain insight into the fine-scale nature of the genome.
2. Genotype a SNP every 5 kb across the whole genome.

This genome-wide data provides a public resource for researchers interested in disease mapping. The HapMap project produced important new data in three principal areas: design and analysis of association studies, understanding human population history as shaped by mutation and recombination, and the illustration of global similarities and differences (www.hapmap.org). With the HapMap resource, it is now possible to tailor genotyping assays so that they capture all the common genomic polymorphisms at a fraction of the cost that would be required to type all the polymorphisms along the genome.

Differences between populations are potentially problematic for utilizing LD information [20, 21]. However, recent work has shown that population differences for LD are consistent with other variation

between populations. Broadly, continental groups can be considered jointly for the purposes of LD information with differences increasing with geographical distance [22–25].

The proportion of human genetic variation that needs to be captured for a study to be classified as a genome-wide association (GWA) scan is still a matter of debate [26]. Current SNP chip technology permits the assaying of approximately 1 million SNPs across the genome. Considering that in 2005, the first major success story of GWA was conducted using 100 K SNP chips, this rate of increase in density (an order of magnitude in 2 years) is dramatic. The first successful GWA identified a risk variant for age-related macular degeneration (AMD; certain variants in the complement factor *H* gene were demonstrated to predispose to the disorder) [27]. Since then, studies employing the 100 K SNP chips have identified variation in *NOS1AP* (also known as *CAPON*) that influences the QT interval on an electrocardiogram [28]. Importantly, both of these findings replicated significantly across a number of other studies and, as a result, are almost certainly true associations [28–33]. It seems likely that variants influencing many other conditions will be identified in future studies. Hopefully, these results in turn will lead to improved tests, treatments, and preventive measures for the major physical and psychological disorders with which so many people are afflicted.

Acknowledgments

MCN is supported by grants MH-65322 and DA-18673. SEM is supported by NHMRC (Australia) Sidney Sax Fellowship 443036.

References

- [1] Posthuma, D., de Geus, E.J., Boomsma, D.I. & Neale, M.C. (2004). Combined linkage and association tests in Mx, *Behavior Genetics* **34**, 179–196.
- [2] Hardy, G. (1908). Mendelian proportions in mixed populations, *Science* **28**, 49–50.
- [3] Weinberg, W. (1908). Über den nachweis der vererbung beim menschen, *Jahresh Ver Vaterl Naturk Württemb* **64**, 369–382.
- [4] Cardon, L.R. & Palmer, L.J. (2003). Population stratification and spurious allelic association, *Lancet* **361**, 598–604.
- [5] Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases, *Science* **273**, 1516–1517.
- [6] Hamer, D. & Sirota, L. (2000). Beware the chopsticks gene, *Molecular Psychiatry* **5**, 11–13.
- [7] Spielman, R.S. & Ewens, W.J. (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test, *American Journal of Human Genetics* **62**, 450–458.
- [8] Horvath, S. & Laird, N.M. (1998). A discordant-sibship test for disequilibrium and linkage: no need for parental data, *American Journal of Human Genetics* **63**, 1886–1897.
- [9] Laird, N. & Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies, *Nature Reviews Genetics* **7**, 385–394.
- [10] Lange, C., DeMeo, D., Silverman, E.K., Weiss, S.T. & Laird, N.M. (2004). PBAT: tools for family-based association studies, *American Journal of Human Genetics* **74**, 367–369.
- [11] Lange, C., DeMeo, D.L. & Laird, N.M. (2002). Power and design considerations for a general class of family-based association tests: quantitative traits, *American Journal of Human Genetics* **71**, 1330–1341.
- [12] Martin, E., Monks, S., Warren, L. & Kaplan, N. (2000). A test for linkage and association in general pedigrees: the pedigree disequilibrium test, *American Journal of Human Genetics* **67**, 146–154.
- [13] International Hapmap Consortium (2005). A haplotype map of the human genome, *Nature* **437**, 1299–1320.
- [14] Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J. & Donnelly, P. (2005). A haplotype map of the human genome, *Nature* **437**, 1299–1320.
- [15] Lewontin, R. & Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms, *Evolution* **14**, 458–472.
- [16] Lewontin, R. (1964). The interaction of selection and linkage I. General considerations; heterotic models, *Genetics* **49**, 49–67.
- [17] Hill, W.G. (1977). Correlation of gene frequencies between neutral linked genes in finite populations, *Theoretical Population Biology* **11**, 239–248.
- [18] Cousin, J. (2002). Human genome. HapMap launched with pledges of \$100 million, *Science* **298**, 941–942.
- [19] Jeffreys, A.J., Kauppi, L. & Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex, *Nature Genetics* **29**, 217–222.
- [20] Van Den Oord, E.J. & Neale, B.M. (2003). Will haplotype maps be useful for finding genes? *Molecular Psychiatry* **9**(3), 227–236.
- [21] Neale, B.M. & Sham, P.C. (2004). The future of association studies: gene-based analysis and replication, *American Journal of Human Genetics* **75**, 353–362.
- [22] Crawford, D.C., Carlson, C.S., Rieder, M.J., Carrington, D.P., Yi, Q., Smith, J.D., Eberle, M.A., Kruglyak, L. & Nickerson, D.A. (2004). Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations, *American Journal of Human Genetics* **74**, 610–622.

- [23] Mueller, J.C., Lohmussaar, E., Magi, R., Remm, M., Bettecken, T., Lichtner, P., Biskup, S., Illig, T., Pfeufer, A., Luedemann, J., Schreiber, S., Pramstaller, P., Pichler, I., Romeo, G., Gaddi, A., Testa, A., Wichmann, H.E., Metspalu, A. & Meitinger, T. (2005). Linkage disequilibrium patterns and tagSNP transferability among European populations, *American Journal of Human Genetics* **76**, 387–398.
- [24] de Bakker, P.I., Burt, N.P., Graham, R.R., Guiducci, C., Yelensky, R., Drake, J.A., Bersaglieri, T., Penney, K.L., Butler, J., Young, S., Onofrio, R.C., Lyon, H.N., Stram, D.O., Haiman, C.A., Freedman, M.L., Zhu, X., Cooper, R., Groop, L., Kolonel, L.N., Henderson, B.E., Daly, M.J., Hirschhorn, J.N. & Altshuler, D. (2006). Transferability of tag SNPs in genetic association studies in multiple populations, *Nature Genetics* **38**, 1298–1303.
- [25] Nejentsev, S., Godfrey, L., Snook, H., Rance, H., Nutland, S., Walker, N.M., Lam, A.C., Guja, C., Ionescu-Tirgoviste, C., Undlien, D.E., Ronningen, K.S., Tuomilehto-Wolf, E., Tuomilehto, J., Newport, M.J., Clayton, D.G. & Todd, J.A. (2004). Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene, *Human Molecular Genetics* **13**, 1633–1639.
- [26] Barrett, J.C. & Cardon, L.R. (2006). Evaluating coverage of genome-wide association studies, *Nature Genetics* **38**, 659–662.
- [27] Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C. & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration, *Science* **308**, 385–389.
- [28] Arking, D.E., Pfeufer, A., Post, W., Kao, W.H., Newton-Cheh, C., Ikeda, M., West, K., Kashuk, C., Akyol, M., Perz, S., Jalilzadeh, S., Illig, T., Gieger, C., Guo, C.Y., Larson, M.G., Wichmann, H.E., Marban, E., O'Donnell, C.J., Hirschhorn, J.N., Kaab, S., Spooner, P.M., Meitinger, T. & Chakravarti, A. (2006). A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization, *Nature Genetics* **38**, 644–651.
- [29] Edwards, A.O., Ritter III, R., Abel, K.J., Manning, A., Panhuysen, C. & Farrer, L.A. (2005). Complement factor H polymorphism and age-related macular degeneration, *Science* **308**, 421–424.
- [30] Hageman, G.S., Anderson, D.H., Johnson, L.V., Hancox, L.S., Taiber, A.J., Hardisty, L.I., Hageman, J.L., Stockman, H.A., Borchardt, J.D., Gehrs, K.M., Smith, R.J., Silvestri, G., Russell, S.R., Klaver, C.C., Barbazetto, I., Chang, S., Yannuzzi, L.A., Barile, G.R., Merriam, J.C., Smith, R.T., Olsh, A.K., Bergeron, J., Zernant, J., Merriam, J.E., Gold, B., Dean, M. & Allikmets, R. (2005). A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration, *Proceedings of the National Academy of Science USA* **102**, 7227–7232.
- [31] Haines, J.L., Hauser, M.A., Schmidt, S., Scott, W.K., Olson, L.M., Gallins, P., Spencer, K.L., Kwan, S.Y., Noureddine, M., Gilbert, J.R., Schnetz-Boutaud, N., Agarwal, A., Postel, E.A. & Pericak-Vance, M.A. (2005). Complement factor H variant increases the risk of age-related macular degeneration, *Science* **308**, 419–421.
- [32] Maller, J., George, S., Purcell, S., Fagerness, J., Altshuler, D., Daly, M.J. & Seddon, J.M. (2006). Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration, *Nature Genetics* **38**, 1055–1059.
- [33] Post, W., Shen, H., Damcott, C., Arking, D.E., Kao, W.H., Sack, P.A., Ryan, K.A., Chakravarti, A., Mitchell, B.D. & Shuldiner, A.R. (2007). Associations between genetic variants in the NOS1AP (CAPON) gene and cardiac repolarization in the old order Amish, *Human Heredity* **64**, 214–219.

Related Articles

Gene–Environment Interaction

Linkage Analysis

BENJAMIN M. NEALE, SARAH E. MEDLAND
AND MICHAEL C. NEALE

Attributable Fraction and Probability of Causation

One often sees measures that attempt to assess the public health impact of an exposure by measuring its contribution to the total incidence under exposure. For convenience, we will refer to the entire family of such fractional measures as *attributable fractions*. The terms *attributable risk percent* or just *attributable risk* are often used as synonyms, although “attributable risk” is also used to denote the risk difference [1–3]. Such fractions may be divided into two broad classes, which have been called *excess fractions* and *etiologic fractions*. The latter class corresponds to the concept of probability of causation.

A fundamental difficulty is that the two classes are usually confused, yet excess fractions can be much smaller than etiologic fractions, even if the disease is rare or other reasonable conditions are met. Another difficulty is that etiologic fractions are not estimable from epidemiologic studies alone, even if those studies are perfectly valid. Assumptions about the underlying biologic mechanism must be introduced to estimate etiologic fractions, and the estimates will be very sensitive to those assumptions.

To describe the situation, imagine we have a cohort of initial size N , and we are concerned with the disease experience over a specific time period, which is at issue in relation to a particular exposure (e.g., to a chemical, drug, pollutant, or food). We wish to contrast the disease experience of the cohort under its actual exposure history with what the experience would have been under a different (counterfactual or reference) exposure history. Usually the actual history involves some degree of exposure for all cohort members, and the reference history would involve no exposure, although the latter may instead involve only a different degree of exposure from the actual level (as is typical in air-pollution research; see **Air Pollution Risk**). For simplicity we will call the actual exposure history *exposed* and the counterfactual history *unexposed*, although the discussion applies to more general cases.

The following notation will be used:

A_1 : number of disease cases occurring under the actual exposure history;

A_0 : number of cases occurring under the counterfactual (reference) exposure history;

$R_1 = A_1/N$: incidence proportion (average risk) under the actual exposure history;

$R_0 = A_0/N$: incidence proportion under the counterfactual exposure history;

T_1 : total time at risk experienced by the cohort under the actual exposure history;

T_0 : total time at risk that would be experienced by the cohort under the counterfactual exposure history;

$I_1 = A_1/T_1$: incidence rate under the actual exposure history;

$I_0 = A_0/T_0$: incidence rate under the counterfactual exposure history. Then $R_1/R_0 = RR$ is the risk ratio and $I_1/I_0 = IR$ is the rate ratio for the effect of the actual exposure history relative to the counterfactual exposure history [4].

Throughout, it is assumed that the counterfactual history is one that is precisely defined (a placebo treatment, or withholding all treatment), and could have occurred physically even though it may not have been ethical or affordable. Without some such constraint, the meaning of “cause”, “effect”, and hence causal attribution becomes unclear [5–7].

Excess Fractions

One family of attributable fractions is based on recalculating an incidence difference as a proportion or fraction of the total incidence under exposure. One such measure is $(A_1 - A_0)/A_1$, the excess caseload owing to exposure, which has been called the *excess fraction* [8]. In a cohort, the fraction $(R_1 - R_0)/R_1 = (RR - 1)/RR$ of the exposed incidence proportion R_1 that is attributable to exposure may be called the *risk fraction*, and is exactly equal to the excess caseload fraction:

$$\begin{aligned} \frac{R_1 - R_0}{R_1} &= \frac{RR - 1}{RR} = \frac{A_1/N - A_0/N}{A_1/N} \\ &= \frac{A_1 - A_0}{A_1} \end{aligned} \quad (1)$$

The analogous relative excess for the incidence rate is the *rate fraction* or *assigned share* $(I_1 - I_0)/I_1 = (IR - 1)/IR$ [8, 9]. This rate fraction is often mistakenly equated with the excess caseload fraction $(A_1 - A_0)/A_1$. To see that the two fractions are not

equal, note that the rate fraction equals

$$\frac{A_1/T_1 - A_0/T_0}{A_1/T_1} \quad (2)$$

If exposure has an effect and the disease removes people from further risk (as when the disease is irreversible), then $A_1 > A_0$, $T_1 < T_0$, and hence the rate fraction will be greater than the excess fraction $(A_1 - A_0)/A_1$. If, however, the exposure effect on total time at risk is small, T_1 will be close to T_0 and the rate fraction will approximate the latter excess fraction [8].

Etiologic Fraction

The *etiologic fraction* is the fraction of cases “caused” by exposure, in that exposure had played some role in the mechanism leading to a person’s disease. We can estimate the total number of cases, and so we could estimate the etiologic fraction if we could estimate the number of cases that were caused by exposure. Unfortunately, and contrary to intuitions and many textbooks, the latter number is not estimable from ordinary incidence data, because the observation of an exposed case does not reveal the mechanism that caused the case. In particular, people who have the exposure can develop the disease from a mechanism that does not include the exposure.

As an example, a smoker may develop lung cancer through some mechanism that does not involve smoking (e.g., one involving asbestos or radiation exposure, with no contribution from smoking). For such lung-cancer cases, their smoking was incidental; it did not contribute to the cancer causation. The exposed cases include some cases of disease caused by the exposure (if the exposure is indeed a cause of disease), and some cases of disease that occur through mechanisms that do not involve the exposure.

Unfortunately, there is usually no way to tell which factors are responsible for a given case. Thus, the incidence of exposed cases of disease caused by exposure usually cannot be estimated [8, 10]. In particular, if I_1 is the incidence rate of disease in a population when exposure is present and I_0 is the rate in that population when exposure is absent, the rate difference $I_1 - I_0$ does not necessarily equal the rate of disease arising from mechanisms that include exposure as a component, and need not even be close

to that rate. Likewise, the rate fraction $(I_1 - I_0)/I_1$ need not be close to the etiologic fraction.

As a simple example of the potential discrepancy, suppose our cohort has $N = 3$ persons given a surgical procedure (the exposure) at age 50, and the time period of interest is from that point for 40 years following exposure (through age 90). Suppose persons 1, 2, and 3 died at ages 60, 70, and 80 as exposed (thus surviving 10, 20, and 30 years past age 50), but would have died at ages 75, 85, and 90 had the exposure (surgery) not been given (instead surviving 25, 35, and 40 years past 50). As all the three cohort members had their lives shortened by the exposure, the etiologic fraction is 1, the largest it can be. However, we have $A_1 = A_0 = 3$, for a risk fraction of $(1 - 1)/1 = 0$. Also, $T_1 = 10 + 20 + 30 = 60$, and $T_0 = 25 + 35 + 40 = 100$, for $I_1 = 3/60 = 0.05$, and $I_0 = 3/100 = 0.03$, and a rate fraction of $(0.05 - 0.03)/(0.05) = 0.20$, far less than 1.

Despite the potentially large discrepancy, excess fractions and rate fractions are often incorrectly interpreted as etiologic fractions. The preceding example shows that these fractions can be far less than the etiologic fraction. Under mechanisms in which exposure accelerates the occurrence of disease (e.g., for tumor promoters), the rate fraction will be close to zero if the rate difference is small relative to I_1 , but the etiologic fraction will remain close to 1, regardless of A_0 or I_0 . The rate fraction and etiologic fraction are equal under certain conditions, but these conditions are not testable with epidemiologic data and rarely have any supporting evidence or genuine biologic plausibility [11–14]. One condition sometimes cited is that exposure acts independently of background causes, which will be examined further below. Without such assumptions, however, the most we can say is that the excess fraction provides a lower bound on the etiologic fraction.

One condition that is irrelevant yet sometimes given is that the disease is rare. To see that this condition is irrelevant, note that the above example made no use of the absolute frequency of the disease; the excess and rate fractions could still be near 0 even if the etiologic fraction was near 1. Disease rarity brings the risk and rate fractions closer to one another, in the same manner it brings the risk and rate ratios close together (assuming exposure does not have a large effect on the person time [4]). However, disease rarity does not bring the rate fraction close to the etiologic fraction, as can be seen by modifying

the above simple example to include an additional 1000 cohort members who survive past age 90 [10].

Probability of Causation

To further illustrate the difference between excess and etiologic fractions, suppose exposure is sometimes causal and never preventive, so that $A_1 > A_0$, and a fraction F of the A_0 cases that would have occurred with or without exposure are caused by exposure when exposure does occur. In the above simple example, $F = 1$, but is usually much smaller. A fraction $1 - F$ of the A_0 cases would be completely unaffected by exposure, and the product $A_0(1 - F)$ is the number of cases unaffected by exposure when exposure occurs. Subtracting this product from A_1 (the number of cases when exposure occurs) gives $A_1 - A_0(1 - F)$ for the number of cases in which exposure plays an etiologic role when it occurs. The fraction of the A_1 cases caused by exposure is thus

$$\frac{A_1 - A_0(1 - F)}{A_1} = \frac{1 - (1 - F)}{RR} \quad (3)$$

If we randomly sample one case, this etiologic fraction formula equals the probability that exposure caused that case, or the *probability of causation* for the case. Although of great biologic and legal interest, this probability cannot be epidemiologically estimated if nothing is known about the fraction F [8, 10–14].

Now suppose exposure is sometimes preventive and never causal, so that $A_1 < A_0$, and a fraction F of the A_1 cases that would have occurred with or without exposure are caused by nonexposure when exposure does not occur. Then the product $A_1(1 - F)$ is the number of cases unaffected by exposure; subtracting this product from A_0 gives $A_0 - A_1(1 - F)$ for the number of cases in which exposure would play a preventive role. The fraction of the A_0 unexposed cases that were caused by nonexposure is thus

$$\frac{A_0 - A_1(1 - F)}{A_0} = \frac{1 - (1 - F)}{RR} \quad (4)$$

As with the etiologic fraction, this fraction cannot be estimated if nothing is known about F [4].

Biologic and Societal Considerations

Excess fractions require no biologic model for their estimation. Thus, they can be estimated from epidemiologic data using only the usual assumptions about study validity and that the exposure does not change the population at risk. In contrast, estimation of the etiologic fraction requires assumptions about the mechanism of exposure action, especially in relation to sufficient causes that act in the absence of exposure. At one extreme, mechanisms involving exposure would occur and act independently of other “background” mechanisms, in which case excess and etiologic fractions will be equal. At the other extreme, in which exposure advances the incidence time whenever a background mechanism is present, the excess fraction can be tiny but the etiologic fraction will be 100% [10–14]. Both extremes are rather implausible in typical settings, and there is rarely enough information to pin down the etiologic fraction, even if the excess fractions are known accurately.

The distinction is of great social importance because of the equality between the etiologic fraction and the probability of causation. The confusion between excess fractions and the probability of causation has led to serious distortions in regulatory and legal decision criteria [10, 14]. The distortions arise when criteria based solely on epidemiologic evidence (such as estimated relative risks) are used to determine whether the probability of causation meets some threshold. The most common mistake is to infer that the probability of causation is below 50% when the relative risk is inferred to be below 2. The reasoning is that $(RR - 1)/RR$ represents the probability of causation, and that this quantity is below 50% unless RR is at least 2. This reasoning is fallacious, however, because $(RR - 1)/RR$ is the excess caseload fraction among the exposed. Thus it may understate the probability of causation to an arbitrarily large degree, in the same manner as it understates the etiologic fraction, even if the RR estimate is highly valid and precise [10–12].

Terminology

More than with other concepts, there is profoundly inconsistent and confusing terminology across the literature on attributable fractions. Levin [15] used the term *attributable proportion* for his original measure

of population disease impact, which in our terms is an excess fraction or risk fraction. Many epidemiologic texts thereafter used the term *attributable risk* to refer to the risk difference $R_1 - R_0$ and called Levin's measure an *attributable risk percent* [1, 3]. By the 1970s, however, portions of the biostatistics literature began calling Levin's measure an "attributable risk" [16, 17], and unfortunately part of the epidemiologic literature followed suit.

Some epidemiologists struggled to keep the distinction by introducing the term *attributable fraction* for Levin's concept [18, 19]; others adopted the term *etiologic fraction* for the same concept and thus confused it with the fraction of cases caused by exposure [20]. The term *attributable risk* continues to be used for completely different concepts, such as the risk difference, the risk fraction, the rate fraction, and the etiologic fraction. On account of this confusion, it has been recommended that the term *attributable risk* be avoided entirely, and that the term *etiologic fraction* not be used for excess fractions [8].

References

- [1] MacMahon, B. & Pugh, T.F. (1970). *Epidemiology: Principles and Methods*, Little, Brown, Boston, 137–198, 175–184.
- [2] Szklo, M. & Nieto, F.J. (2006). *Epidemiology: Beyond the Basics*, 2nd Edition, Jones and Bartlett.
- [3] Koepsell, T.D. & Weiss, N.S. (2003). *Epidemiologic Methods*, Oxford, New York.
- [4] Greenland, S., Rothman, K.J. & Lash, T.L. (2008). Measures of effect and association, in *Modern Epidemiology*, 3rd Edition, K.J. Rothman, S. Greenland, & T.L. Lash, eds, Lippincott, Philadelphia, Chapter 4.
- [5] Greenland, S. (2002). Causality theory for policy uses of epidemiologic measures, in *Summary Measures of Population Health*, C.J.L. Murray, J.A. Salomon, C.D. Mathers, & A.D. Lopez, eds, Harvard University Press/WHO, Cambridge, Chapter 6.2, pp. 291–302.
- [6] Greenland, S. (2005). Epidemiologic measures and policy formulation: lessons from potential outcomes (with discussion), *Emerging Themes in Epidemiology* **2**, 1–4.
- [7] Hernán, M.A. (2005). Hypothetical interventions to define causal effects – afterthought or prerequisite? *American Journal of Epidemiology* **162**, 618–620.
- [8] Greenland, S. & Robins, J.M. (1988). Conceptual problems in the definition and interpretation of attributable fractions, *American Journal of Epidemiology* **128**, 1185–1197.
- [9] Cox, L.A. (1987). Statistical issues in the estimation of assigned shares for carcinogenesis liability, *Risk Analysis* **7**, 71–80.
- [10] Greenland, S. (1999). The relation of the probability of causation to the relative risk and the doubling dose: a methodologic error that has become a social problem, *American Journal of Public Health* **89**, 1166–1169.
- [11] Robins, J.M. & Greenland, S. (1989). Estimability and estimation of excess and etiologic fractions, *Statistics in Medicine* **8**, 845–859.
- [12] Robins, J.M. & Greenland, S. (1989). The probability of causation under a stochastic model for individual risks, *Biometrics* **46**, 1125–1138.
- [13] Beyea, J. & Greenland, S. (1999). The importance of specifying the underlying biologic model in estimating the probability of causation, *Health Physics* **76**, 269–274.
- [14] Greenland, S. & Robins, J.M. (2000). Epidemiology, justice, and the probability of causation, *Jurimetrics* **40**, 321–340.
- [15] Levin, M.L. (1953). The occurrence of lung cancer in man, *Acta Unio Internationalis Contra Cancrum* **9**, 531–541.
- [16] Walter, S.D. (1976). The estimation and interpretation of attributable risk in health research, *Biometrics* **32**, 829–849.
- [17] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research, Vol. I: The Analysis of Case-Control Data*, IARC, Lyon.
- [18] Ouellet, B.L., Roemeder, J.-M. & Lance, J.-M. (1979). Premature mortality attributable to smoking and hazardous drinking in Canada, *American Journal of Epidemiology* **109**, 451–463.
- [19] Deubner, D.C., Wilkinson, W.E., Helms, M.J., Tyroler, H.A. & Hames, C.G. (1980). Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia, *American Journal of Epidemiology* **112**, 135–143.
- [20] Miettinen, O.S. (1974). Proportion of disease caused or prevented by a given exposure, trait, or intervention, *American Journal of Epidemiology* **99**, 325–332.

Related Articles

Causality/Causation

Relative Risk

SANDER GREENLAND

Availability and Maintainability

Modern-day industry has become more mechanized and automated. This trend has also enhanced the business risks associated with stoppages or nonfunctioning of equipments and other engineering systems deployed in such industries. In general, mechanized and automatic systems are expected to perform safely with designed performance most of the time or even round the clock. However, owing to design deficiencies or the influence of operational and environmental stresses, these systems are not completely failure free and are unable to meet customer requirements in terms of system performance. This is often attributed to poorly designed reliability and maintainability characteristics, combined with poor maintenance and product support strategies. If the designed reliability and maintainability characteristics are poor and the maintenance preparedness is unable to meet the users' requirements, this may result in a poor availability performance.

Reliability and maintainability are commonly regarded as engineering characteristics of the technical system, dimensioned during the design and development phase of the system's life cycle, whereas maintenance support is dependent on the users' organizational and operational effectiveness, including their operating environments. All these factors have an impact on the risks (or uncertainties) related to the availability performance of the system. A high availability performance level is an important and critical requirement for most of the technical systems. The factors influencing availability performance of engineering systems include reliability, maintainability, and maintenance support (see Figure 1). Even though high reliability is desirable to reduce the costs of maintenance and maintenance preparedness, the benefits of improved maintainability should not be overlooked in achieving the required availability performance. High maintainability performance results in high availability performance as the system is easy to maintain and repair.

Availability Concepts

Availability is generally the most appropriate measure for the performance of repairable items [1].

Availability performance is described by the collective term *dependability* and is a function of reliability performance, maintainability performance, and maintenance support performance as illustrated in Figure 1.

Availability Performance

The formal definition provided by the International Electrotechnical Vocabulary (IEV) (IEV 191-02-05) is as follows [2]: "The ability of an item to be in a state to perform a required function under given conditions at a given instant of time or over a given time interval, assuming that the required external resources are provided".

This means that availability is related to the state of an item (system, equipment, component, etc.), whether or not it is operating. The availability of safety or backup systems is quite an important matter of interest, highlighting the need for system and component state or health information.

Related to availability performance, there are various measures of availability, e.g., instantaneous and mean availability, respectively, and asymptotic availability measures. The most frequently used measure of availability is defined as the ratio between the time during which the equipment is available for use and the total planned time for operation: $availability = \frac{up-time}{total\ planned\ time}$.

A more abstract availability measure can be defined as follows: "the probability that a system will be in a functioning state on demand".

This definition is very important from a safety point of view. It is often used in defining the state and condition of the safety equipment, and is used frequently in the oil and gas industry and nuclear power plants. The safety equipment monitors the plant and processes for any unsafe state. Upon the detection of an unsafe state, the safety equipment acts or initiates actions to prevent a hazardous situation from developing. When the safety equipment is in a nonoperational state when required to function, then it is termed *unavailable*. As the arrival of an unsafe state occurs at random, it is desirable that the safety equipment should be available all the time, i.e., it is desirable to have a high level of availability. For highly dependable systems it is often more appropriate to focus on unavailability rather than on availability [3]. The measures are closely related: $unavailability = 1 - availability$ e.g., the

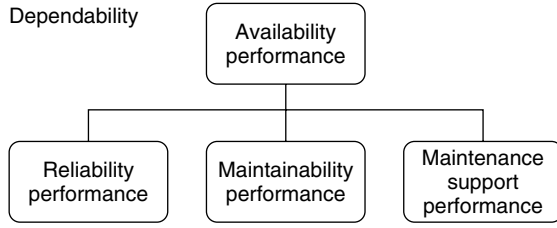


Figure 1 Dependability and availability concepts

effect of reducing the unavailability by 50%, from 0.0024 to 0.0012, is easier to comprehend than the effect of increasing the availability by 0.12%, from 0.9976 to 0.9988.

Similarly, this definition is useful to define the performance of the standby systems. In a production system with a standby production line, it is desirable that the standby system should start on demand. Such a demand is generated by the failure of the primary system or component.

For a continuously operating system, it is natural to define availability as “the probability that the system or a component is working at a specified time t ”. Upon the occurrence of failure, a repair process is initiated to bring the system back into the functioning state.

Many times, availability is also defined depending on whether the waiting time or the preventive maintenance times are included in or excluded from the calculation of the availability of the system. Inherent availability, achieved availability, and operational availability are some of the important measures used to quantify the extent to which an item is in an operational state, depending on whether the preventive maintenance time and the waiting time, apart from the active repair time, are taken into consideration. These categories of availability are defined as follows [4]:

1. Inherent availability (A_i)

The inherent availability is “the probability that a system or equipment, when used under stated conditions in an ideal support environment (e.g., readily available tools, spares, maintenance personnel, etc.), will operate satisfactorily at any point in time as required”. It excludes preventive or scheduled maintenance actions and logistics delay time, and is expressed as

$$A_i = \frac{MTBF}{MTBF + MTTR} \quad (1)$$

where $MTBF$ is the mean time between failure (*see Reliability Growth Testing; Reliability Data*) and $MTTR$ is the mean time to repair (*see Systems Reliability; Markov Modeling in Reliability*).

A_i is used for evaluation of different design alternatives to meet the design specification from system users.

2. Achieved availability (A_a)

Achieved availability is defined in a similar way as A_i – except that preventive (e.g., scheduled) maintenance is included. It excludes logistics delay time and administrative delay time and is expressed as

$$A_a = \frac{MTBM}{MTBM + \bar{M}} \quad (2)$$

where $MTBM$ is the mean time between maintenance and $\bar{M}$ the mean active maintenance time. $MTBM$ is a function of maintenance actions while $\bar{M}$ is a function of repair and service times.

A_a is used to assess the suitability of various preventive maintenance programs so as to meet the design specification from customers.

3. Operational availability (A_o)

Operational availability is “the probability that a system or an item of equipment, when used under stated conditions in an actual operational environment, will operate satisfactorily when called upon”. It is expressed as

$$A_o = \frac{MTBM}{MTBM + MDT} \quad (3)$$

where MDT is the mean maintenance downtime. The reciprocal of $MTBM$ is the frequency of maintenance, which in turn is a significant factor in determining logistic support requirements. MDT includes active maintenance time, logistics delay time, and administrative delay time.

A_o is used during the system operation phase to assess its performance in real situation.

Table 1 summarizes the various availability measures discussed earlier.

The term *availability* is used differently in different situations. If one is to improve an availability figure of merit as a design requirement for a given equipment design, and the designer has no control over the operational environment in which that equipment is to function, then A_a or A_i might be an

Table 1 Various availability measures

Concept	Calculation formula	Times included in calculations			Application area
		Reactive maintenance	Preventive and scheduled maintenance	Logistic and administrative delay	
General	$A = \frac{Up = time}{Total\ planned\ time}$	✓	✓	✓	Production
Inherent availability	$A_i = \frac{MTBF}{MTBF + MTTR}$	✓	–	–	System design
Achieved availability	$A_a = \frac{MTBM}{MTBM + \bar{M}}$	✓	✓	–	System design and maintenance planning
Operational availability	$A_o = \frac{MTBM}{MTBM + MDT}$	✓	✓	✓	Operation

appropriate figure of merit according to which equipment can be properly assessed. Conversely, if one is to assess a system in a realistic operational environment, then A_o is a preferred figure of merit to use for assessment purposes.

Furthermore, availability is defined differently by different groups depending on their background and interest. Many times there is no agreement on the meaning of downtime in the same organization, leading to different values of availability calculated for the same plant in the same period. Often, it is difficult to agree on what downtime is, as downtime to one person may not be downtime to another. To some people, downtime is related exclusively to machine health and does not include factors such as the waiting time for repair when a repairman is not available or the waiting time due to nonavailability of spare parts. There is no universally accepted definition of availability.

Example 1 Load haul dump (LHD) machines are used in many underground mines for loading and transportation of ore and other minerals.

The following data are extracted from the operation and maintenance card of an LHD machine deployed in a Swedish mine for the month of March, 2007. The machine was scheduled for 16h of operation per day for 23 days i.e., machine was scheduled for 368h of operation during the month; and no preventive maintenance actions were planned or performed.

All times are in hours.

Times between failure (TBF): 110.0, 13.0, 72.0, 4.0, 45.0, 56.0, 19.0.

Times to repair (TTR) including logistics and administrative delays: 7.5, 9.0, 6.3, 1.2, 2.6, 21.1, 1.3.

Active repair times (without logistics and administrative delays): 4.0, 6.5, 5.6, 1.0, 2.4, 11.0, 1.0.

In this case, A_i and A_a will be the same since no preventive maintenance action is performed.

To calculate A_a , we need to consider only the TBF (mean active operation time) and mean active repair time.

MTBM

= Mean times between failures(or maintenance)

$$= \frac{\text{Arithmetic sum of the TBF}}{\text{number of failures}}$$

$$= 319/7 = 45.6 \quad (4)$$

MDT

$$= \frac{\text{Arithmetic sum of the downtimes}}{\text{number of repairs}}$$

$$= 49/7 = 7 \quad (5)$$

Mean active repair time :

$$= \frac{\text{arithmetic sum of the active repair time}}{\text{number of repairs performed}}$$

$$= 31.5/7 = 4.5 \quad (6)$$

Achieved availability = A_a

$$= \frac{MTBM}{MTBM + \overline{M}} = 45.6/(45.6 + 4.5) = 91.1 \quad (7)$$

Operational availability = A_o

$$= \frac{MTBM}{MTBM + MDT} = 45.6/(45.6 + 7) = 86.7 \quad (8)$$

In an organization with world class logistics, i.e., without any logistics or administrative delays, the value of A_o should be approaching A_a .

Example 2 In the same mine, a detailed examination of the drill performance report showed two different figures for the availability of the heavy duty drill machines. One figure ($A_o = 81.5\%$) was reported by the production department and other figure ($A_o = 85\%$) by the contractor responsible for repair and maintenance of drill machines. The contract also has a provision for rewarding the contractor if he delivers better than minimum agreed availability target of 85% and penalizing the contractor if he fails to meet the minimum availability target of 85%.

A close study of all the facts and available records showed that the discrepancies in reporting availability figures were mainly owing to the fact that the maintenance contractor had not treated the high noise level in drill machine as a failure mode because the machine was functionally *ready* for operation. However, the operator refused to operate the machine saying that the high level of noise makes the machine unsuitable for operation, and declaring it *not ready* for operation.

Therefore, many times it is difficult to agree on the definition of machine breakdown, leading to different availability values of the same machine during the same period being provided by different parties.

From the above examples, we see that the availability performance of an engineering system is dependent on reliability performance, maintainability performance, and maintenance support performance.

Reliability Performance

Decisions regarding the reliability characteristics of an item of equipment are taken during the drawing board stage of equipment development. The aim is to prevent the occurrence of failures and also to eliminate their effects on the operational capability of the systems. The factors that have an important influence on the equipment reliability are

- period of use
- environment of use.

The formal definition of reliability according to IEV (191-02-06) is as follows [2]: “The ability of an item to perform required function under given conditions for a given time interval”.

A keyword is “a required function”. This means that it is possible to consider a number of “reliabilities” for a certain item, taking into account one requirement at a time. Of special interest are the safety functions or barriers related to the item under study. Another important word combination is “under given conditions”. It is essential that all the different foreseeable uses (and misuses) of the system under study should be considered when the requirements are identified. During the design of a system or item of equipment, the aim is to prevent the occurrence of failures as far as possible and also to reduce the effects of the failures that cannot be eliminated. The reliability of a system in operation is not determined solely by inherent design properties. A number of other factors such as operating environment, quality of repair and maintenance works performed, and so on, are taken into consideration.

Even though the system may have good reliability characteristics, the performance of the system may be poor because of scant attention paid to maintainability characteristics during the design phase.

Maintainability Performance

Maintainability is a design characteristic governed by design specification and is decided during the design and development phase of systems. It has a direct impact on availability in the sense that the time taken to locate faults and carry out routine preventive maintenance tasks contributes to the system downtime [4, 5]. Maintainability (*see Reliability Integrated Engineering Using Physics of Failure; No Fault Found*) and maintenance (*see Evaluation of Risk Communication Efforts; Repair, Inspection, and Replacement Models*) have always been important to the industry as they affect the system performance and thereby have a direct impact on the financial result of a company.

The objective of the maintainability input is to minimize the maintenance times and labor hours while maximizing the supportability characteristics in design (e.g., accessibility, diagnostic provisions, standardization, interchangeability), minimize the logistic

support resources required in the performance of maintenance (e.g., spare parts and supporting inventories, test equipment, maintenance personnel), and also minimize the maintenance cost.

The formal definition of maintainability is as follows (IEV 191-02-07) [2]: “The ability of an item under given conditions of use, to be retained in, or restored to, a state in which it can perform a required function, when maintenance is performed under given conditions and using stated procedures and resources”.

It is important to note that the ability to be retained in a satisfactory state is included, and not just the ability to be restored to a satisfactory state. Also note the important words *given conditions*. Often systems are designed, built, and tested in an environment with a comfortable temperature and good lighting. However, in real life, maintenance is often performed in a hostile environment, e.g., subarctic conditions or in tropical environments. The operating situation needs to be taken into account when designing for good maintainability. Robinson *et al.* [6] present an interesting discussion on the field maintenance interface between human engineering and maintainability engineering.

High maintainability performance and, in turn, high availability performance are obtained when the system is easy to maintain and repair. In general, the maintainability is measured by the mean repair time, often called *mean time to repair*, which includes the total time for fault finding, and the actual time spent in carrying out the repair. Both the way in which a system is designed and its installation have a direct bearing on the maintainability of the system. For capital-intensive equipment, the maintainability parameter influences the cost of life cycle more intensively than in the case of smaller equipment or throwaway consumer goods. Maintainability considerations are also very important for units operating under tough conditions, for instance, mining equipment. There exist numerous guidelines for taking maintainability into account [4, 7–10]. These guidelines are also very useful in the analysis of maintainability-related risks. The various aspects of maintainability may be grouped as shown in Figure 2.

Maintainability, as mentioned earlier, measures the ease and speed with which a system can be restored to an operational status after a failure occurs. Measures of maintainability are normally related to

the distribution of time needed for the performance of specified maintenance actions. For example, if it is said that a particular component has 90% maintainability in 1 h, this means that there is a 90% probability that the component will be repaired within an hour. Furthermore, there always exist risks associated with the performance of the organization that can be controlled and minimized by improving the maintainability of the system.

Sometimes the term *maintainability* is confused with maintenance. Maintainability is the ability of a product to be maintained, whereas maintenance constitutes a series of actions to be taken to restore a product to or retain it in an effective operational state. Maintainability is a design-dependent parameter. Maintenance is a result of design and operating environment. Ergonomic considerations in design for maintainability and also in the design of maintenance facilities have an important influence on the repair time and thereby on the availability of the system.

Various attempts have been made by researchers to develop procedures for evaluation of the maintainability of the various systems [11, 12].

Maintenance Support Performance

This final subconcept of availability is defined as follows (IEV 191-02-08) [2]: “The ability of a maintenance organization, under given conditions to provide upon demand the resources required to maintain an item, under a given maintenance policy”.

Thus, we can see that maintenance support performance is part of the wider concept of “product support”, including support to the product as well as support to the client.

Maintenance support is developed and maintained within the framework of corporate objectives. Defining and developing maintenance procedures, procurement of maintenance tools and facilities, logistic administration, documentation, and development and training programs for maintenance personnel are some of the essential features of a maintenance support system (Figure 3).

Availability, Maintainability, and Risk

The main factors to be considered when dealing with risks are the risk sources, or hazards, and the objects of harm. A risk source, or hazard, is defined as an

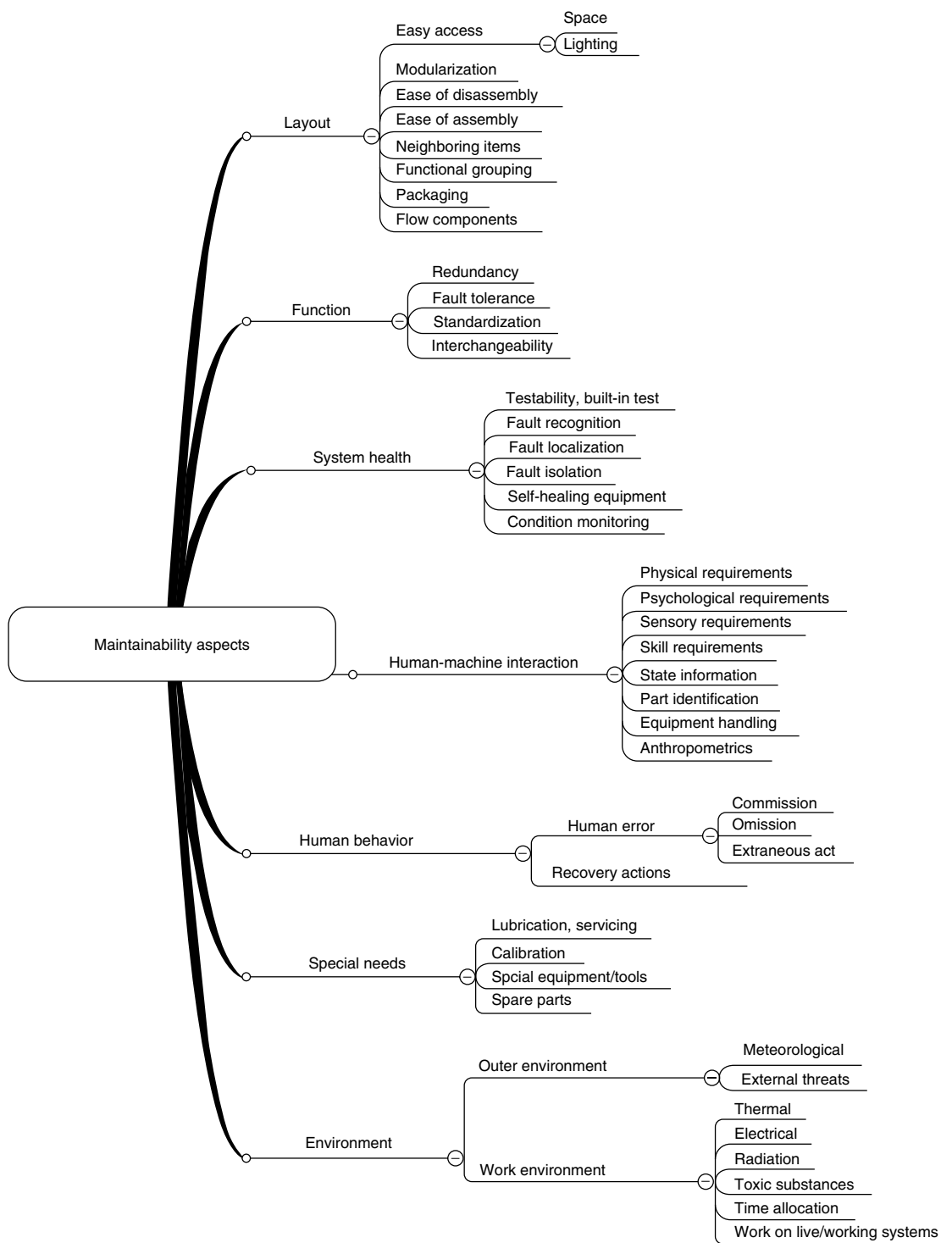


Figure 2 A number of factors influencing maintainability

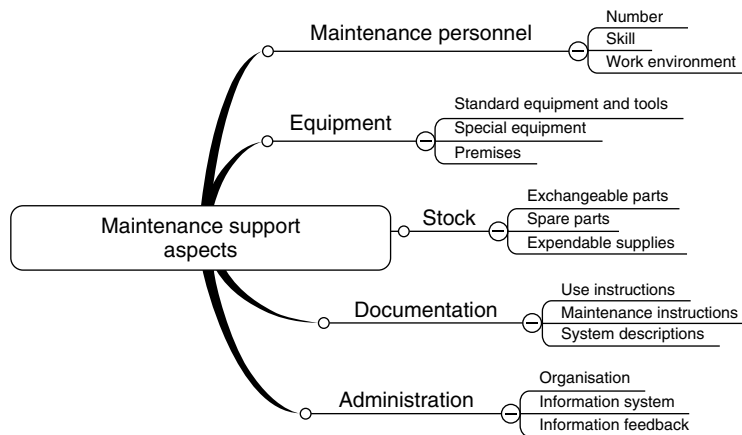


Figure 3 Some aspects of maintenance support

item or activity having a potential for an undesired consequence [13], and an object of harm is defined as any object, within the delimitations of the analysis, that can suffer a loss of some kind owing to the impact of a risk source. The risk sources related to availability and maintainability are of several types, mostly connected to the planning and performance of maintenance. The list of maintenance aspects is very helpful in this respect. The objects of harm range from the system under study, in its business and physical environment, down to components subject to repair or replacement.

The description of the risks identified can be made using scenarios starting with an initiating event from a risk source and ending with possible consequences for different objects of harm. The various mind maps, describing the factors influencing the reliability performance, maintainability performance, and maintenance support performance, respectively, may be used as checklists in the identification of risk sources. Experience from maintenance work within different fields may be used for the more concrete description of the risk sources. Some examples are as follows:

- Poor accessibility may lead to the use of inappropriate access paths, jeopardizing adjacent equipment, and components.
- Poor exchangeability, in the sense that physically exchangeable parts may have different performance characteristics, may lead to serious malfunctions.

- Disassembly difficulties may lead to the use of extra high forces, causing damage, e.g., to structure components.

Conclusions

Availability is the most commonly used measure of the state of plant and systems all over the world, and maintainability performance is the universal measure of the maintenance friendliness of systems. Not only are the availability and maintainability measures of system performance, but they also have considerable influence on the financial performance and survival of businesses. Designing and using systems that are maintenance friendly significantly improves the safety and also system availability performance. It is true that improved reliability is the best route to achieve higher availability performance; the benefits of improved maintainability should not be ignored. Many times, it is easier and cost effective to meet the availability performance requirement through improvement in maintainability characteristics than through improvement in reliability characteristics of the system.

References

- [1] Modarres, M. (2006). *Risk Analysis in Engineering: Techniques, Tools, and Trends*, CRC Press.
- [2] IEV 191 (2007). Dependability and quality of service, in *International Electrotechnical Vocabulary (IEV) Online*,

- Chapter 191, at <http://std.iec.ch/iec60050> (accessed Jan 2007).
- [3] Andrews, J.D. & Moss, T.R. (2006). *Reliability and Risk Assessment*, ISBN 1860582907, 2nd Edition, Professional Engineering Publisher, London.
 - [4] Blanchard, B.S., Verma, D. & Peterson, E.L. (1995). *Maintainability: A Key to Effective Serviceability and Maintenance Management*, John Wiley & Sons, New York.
 - [5] Dhillon, B.S. (2002). *Engineering Maintenance: A Modern Approach*, ISBN 1-58716-142-7, CRC Press.
 - [6] Robinson Jr, J.E., Deutsch, W.E. & Roger, J.G. (1970). The field maintenance interface between human engineering and maintainability engineering, *Human Factors* **12**(3), 253–259 (Special issue on maintenance and maintainability).
 - [7] IEC 60300-3-10 (2001). *International Standard IEC 60300-3-10: Dependability Management – Part 3 – 10: Application Guide – Maintainability*, Geneva.
 - [8] IEC 60706-2 (2006). *International Standard IEC 60706-2: Maintainability of Equipment – Part 2: Maintainability Requirements and Studies During the Design and Development Phase*, 2nd Edition, Geneva.
 - [9] IEC 60706-3 (2006). *International Standard IEC 60706-3: Maintainability of Equipment – Part 3: Verification and Collection, Analysis and Presentation of Data*, 2nd Edition, Geneva.
 - [10] Tjiparuro, Z. & Thompson, G. (2004). Review of maintainability design principles and their application to conceptual design, *Part E: Journal of Process Mechanical Engineering* **218**(2), 103–113.
 - [11] Smith, R.L., Westland, R.A. & Crawford, B.M. (1970). The status of maintainability models: a critical review, *Human Factors, Special Issue on Maintenance and Maintainability* **12**(3), 271–283.
 - [12] Wani, M.F. & Gandhi, O.P. (1999). Development of maintainability index for mechanical systems, *Reliability Engineering and System Safety* **65**, 259–277.
 - [13] IEC 60300-3-9 (1995). *International Standard IEC 60300-3-9: Dependability Management – Part 3: Application Guide – Section 9: Risk Analysis of Technological Systems*, Geneva.

UDAY KUMAR AND PER ANDERS AKERSTEN

Axiomatic Measures of Risk and Risk-Value Models

This article provides a review of measures of risk and risk-value models that have been developed over the past 10 years to provide a new class of decision making models based on the idea of risk-value trade-offs. The measurement of risk has been a critical issue in decision sciences, finance, and other fields for many years (*see* **Decision Modeling; Model Risk; Simulation in Risk Management; Environmental Health Risk**). We focus on a preference-dependent measure of risk that can be used to derive risk-value models within both an expected utility framework and a nonexpected utility framework. Although this measure of risk has some descriptive power for risk judgments, it is more normative in nature. We treat the issue of measures of perceived risk in a separate article (*see* **Axiomatic Models of Perceived Risk**).

Intuitively, individuals may consider their choices over risky alternatives by trading off between risk and return, where return is typically measured as the mean (or expected return) and risk is measured by some indicator of dispersion or possible losses. Markowitz [1–3] proposed variance as a measure of risk, and a mean–variance model for portfolio selection based on minimizing variance subject to a given level of mean return. Arguments have been put forth that mean–variance models are appropriate only if the investor’s utility function is quadratic or the joint distribution of returns is normal. However, these conditions are rarely satisfied in practice. Markowitz also suggested semivariance as an alternative measure of risk. Some other measures of risk, such as lower partial moment risk measures and absolute standard deviation, have also been proposed in the financial literature (e.g. [4]). However, without a common framework for risk models, it is difficult to justify and evaluate these different measures of risk as components of a decision making process.

Expected utility theory is generally regarded as the foundation of mean-risk models and risk-return models (e.g. [5–9]). However, expected utility theory has been called into question by empirical studies of risky choice (e.g. [10–14]). This suggests that an

alternative approach regarding the paradigm of risk-return trade-offs would be useful for predicting and describing observed preferences.

In the mainstream of decision research (*see* **Decision Modeling**), the role of risk in determining preference is usually considered implicitly. For instance, in the expected utility model [15], an individual’s attitude toward the risk involved in making choices among risky alternatives is defined by the shape of his or her utility function [16, 17]; and in some non-expected utility models, risk (or “additional” risk) is also captured by some nonlinear function over probabilities (e.g. *see* [12, 18–20]). Thus, these decision theories are not, at least explicitly, compatible with the choice behavior based on the intuitive idea of risk-return trade-offs as often observed in practice. Therefore, they offer little guidance for this type of decision making.

In this article, we review our risk-value studies and provide an axiomatic measure of risk that is compatible with choice behavior based on risk-value trade-offs. In particular, this framework unifies two streams of research: one in modeling risk judgments and the other in developing preference models. This synthesis provides risk-value models that are more descriptively powerful than other preference models and risk models that have been developed separately.

The remainder of this article is organized as follows. The next section describes a preference-dependent measure of risk with several useful examples. The section titled “Frameworks for Risk-Value Trade-off” reviews the basic framework of our risk-value studies and related preference conditions. The section titled “Generalized Risk-Value Models” presents three specific forms of risk-value models. The last section summarizes the applicability risk-value studies and discusses topics for future research.

The Standard Measure of Risk

As a first step in developing risk-value models, Jia and Dyer [21] propose a preference-dependent measure of risk, called a *standard measure of risk*. This general measure of risk is based on the converse expected utility of normalized lotteries with zero-expected values, so it is compatible with the measure of expected utility and provides the basis for linking risk with preference.

For lotteries with zero-expected values, we assume that the only choice attribute of relevance is risk. A riskier lottery would be less preferable and vice versa, for any risk averse decision maker. Therefore, the riskiness ordering of these lotteries should be simply the reverse of the preference ordering. However, if a lottery has a nonzero mean, then we assume that the risk of that lottery should be evaluated relative to a “target” or reference level. The expected value of the lottery is a natural reference point for measuring the risk of a lottery.

Therefore, we consider decomposing a X (i.e., a random variable) into its mean $\bar{X}$ and its standard risk, $X' = X - \bar{X}$, and the standard measure of risk is defined as follows:

$$R(X') = -E[u(X')] = -E[u(X - \bar{X})] \quad (1)$$

where $u(\cdot)$ is a utility function [15] and the symbol E represents expectation over the probability distribution of a lottery.

One of the characteristics of this standard measure of risk is that it depends on an individual's utility function. Once the form of the utility function is determined, we can derive the associated standard measure of risk over lotteries with zero means. More important, this standard measure of risk can offer a preference justification for some commonly used measures of risk, so that the suitability of those risk measures can be evaluated.

If a utility function is quadratic, $u(x) = ax - bx^2$, where $a, b > 0$, then the standard measure of risk is characterized by variance, $R(X') = bE[(X - \bar{X})^2]$. However, the quadratic utility function has a disturbing property; that is, it is decreasing in x after a certain point, and it exhibits increasing risk aversion. Since the quadratic utility function may not be an appropriate description of preference, it follows that variance may not be a good measure for risk (unless the distribution of a lottery is normal).

To obtain a related, but increasing, utility function, consider a third-order polynomial (or cubic) utility model, $u(x) = ax - bx^2 + c'x^3$, where $a, b, c' > 0$. When $b^2 < 3ac'$, the cubic utility model is increasing. This utility function is concave, and hence risk averse for low outcome levels (i.e., for $x < b/(3c')$), and convex, and thus risk seeking for high outcome values (i.e., for $x > b/(3c')$). Such a utility function may be used to model a preference structure consistent with the observation that a large number of individuals purchase both insurance (a moderate outcome-small

probability event) and lottery tickets (a small chance of a large outcome) in the traditional expected utility framework [22]. The associated standard measure of risk for this utility function can be obtained as follows:

$$R(X') = E[(X - \bar{X})^2] - cE[(X - \bar{X})^3] \quad (2)$$

where $c = c'/b > 0$. Model (2) provides a simple way to combine skewness with variance into a measure of risk. This measure of risk should be superior to the one based on variance alone, since the utility function implied by equation (2) has a more intuitive appeal than the quadratic one implied by variance. Further, since equation (2) is not consistent with increasing risk aversion, for prescriptive purposes, it is more appropriate than variance.

Markowitz [23] noted that an individual with the utility function that is concave for low outcome levels and convex for high outcome values will tend to prefer positively skewed distributions (with large right tails) over negatively skewed ones (with large left tails). The standard measure of risk (equation (2)) clearly reflects this observation; i.e., a positive skewness will reduce risk and a negative skewness will increase risk.

If an individual's preference can be modeled by an exponential or a linear plus exponential utility function, $u(x) = ax - be^{-cx}$, where $a \geq 0$, and $b, c > 0$, then its corresponding standard measure of risk (with the normalization condition $R(0) = 1$) is given by

$$R(X') = E[e^{-c(X - \bar{X})} - 1] \quad (3)$$

Bell [7] identified $E[e^{-c(X - \bar{X})}]$ as a measure of risk from the linear plus exponential utility model by arguing that the riskiness of a lottery should be independent of its expected value. Weber [24] also modified Sarin's [25] expected exponential risk model by requiring that the risk measure be location free.

If an individual is risk averse for gains but risk seeking for losses as suggested by Prospect theory [12, 26], then we can consider a piecewise power utility model as follows:

$$u(x) = \begin{cases} ex^{\theta_1}, & \text{when } x \geq 0 \\ -d|x|^{\theta_2}, & \text{when } x < 0 \end{cases} \quad (4)$$

where e, d, θ_1 , and θ_2 are nonnegative constants. Applying equation (1), the corresponding standard

measure of risk is given by

$$R(X') = dE^-[|X - \bar{X}|^{\theta_2}] - eE^+[|X - \bar{X}|^{\theta_1}] \quad (5)$$

where $E^-[|X - \bar{X}|^{\theta_2}] = \int_{-\infty}^{\bar{X}} |x - \bar{X}|^{\theta_2} f(x) dx$, $E^+[|X - \bar{X}|^{\theta_1}] = \int_{\bar{X}}^{\infty} (x - \bar{X})^{\theta_1} f(x) dx$, and $f(x)$ is the probability density of a lottery.

The standard measure of risk (equation (5)) includes several commonly used measures of risk as special cases in the financial literature. When $d > e > 0$, $\theta_1 = \theta_2 = \theta > 0$, and the distribution of a lottery is symmetric, we have $R(X') = (d - e)E[|X - \bar{X}|^\theta]$, which is associated with variance and absolute standard deviation if $\theta = 2$ and $\theta = 1$, respectively. This standard measure of risk is also related to the difference between the parameters d and e that reflect the relative effect of loss and gain on risk. In general, if the distribution of a lottery is not symmetric, this standard measure of risk will not be consistent with the variance of the lottery even if $\theta_1 = \theta_2 = 2$, but it is still related to the absolute standard deviation if $\theta_1 = \theta_2 = 1$ [27].

Konno and Yamazaki [28] have argued that the absolute standard deviation is more appropriate for use in portfolio decision making than the variance, primarily due to its computational advantages. Dennenberg [29] argues that the average absolute deviation is a better statistic for determining the safety loading (premium minus the expected value) for insurance premiums than the standard deviation. These arguments suggest that in some applied contexts, the absolute standard deviation may be a more suitable measure of risk than variance.

Another extreme case of equation (5) arises when $e = 0$ (i.e., the utility function is nonincreasing for gains); then the standard measure of risk is a lower partial moment risk model $R(X') = dE^-[|X - \bar{X}|^{\theta_2}]$. When $\theta_2 = 2$, it becomes a semivariance measure of risk [1]; and when $\theta_2 = 0$, it reduces to the probability of loss.

To summarize, there are other proposed measures of risk that are special cases of this standard measure of risk. The standard measure of risk is more normative in nature, as it is independent of the expected value of a lottery. To obtain more descriptive power and to capture perceptions of risk, we have also established measures of perceived risk that are based on a two-attribute structure: the mean of a lottery and its standard risk [30], as described in a separate article in this encyclopedia (see **Axiomatic Models of Perceived Risk**).

Frameworks for Risk-Value Trade-off

When we decompose a lottery into its mean and standard risk, then the evaluation of the lottery can be based on the trade-off between mean and risk. We assume a risk-value preference function $f(\bar{X}, R(X'))$, where f is increasing in $\bar{X}$ and decreasing in $R(X')$ if one is risk averse.

Consider an investor who wants to maximize his or her preference function f for an investment and also requires a certain level μ of expected return. Since f is decreasing in $R(X')$ and $\bar{X} = \mu$ is a constant, maximizing $f(\bar{X}, R(X'))$ is equivalent to minimizing $R(X')$; i.e., $\max \{f(\bar{X}, R(X')) | \bar{X} = \mu\} \Rightarrow \min \{R(X') | \bar{X} = \mu\}$. This conditional optimization model includes many financial optimization models as special cases that depend on the choice of different standard measures of risk; e.g., Markowitz's mean-variance model, the mean-absolute standard deviation model, and the mean-semivariance model. Some new optimization models can also be formulated based on our standard measures of risk (equations (2) and (5)).

In the conditional optimization problem, we do not need to assume an explicit form for the preference function f . The problem only depends on the standard measure of risk. However, we may argue that an investor should maximize his or her preference functions unconditionally in order to obtain the overall optimal portfolio. For an unconditional optimization decision, the investor's preference function must be specified. Here, we consider two cases for the preference function f : (a) when it is consistent with the expected utility theory; and (b) when it is based on a two-attribute expected utility foundation.

Let $\mathbf{P}$ be a convex set of all simple probability or lotteries $\{X, Y, \dots\}$ on a nonempty set $\mathbf{X}$ of outcomes, and Re be the set of real numbers (assuming $\mathbf{X} \in \text{Re}$ is finite). We define $\succ$ as a binary preference relation on $\mathbf{P}$.

Definition 1 For two lotteries $X, Y \in \mathbf{P}$ with $E(X) = E(Y)$, if $w_0 + X \succ w_0 + Y$ for some $w_0 \in \text{Re}$, then $w + X \succ w + Y$, for all $w \in \text{Re}$.

This is called the *risk independence condition*. It means that for a pair of lotteries with a common mean, the preference order between the two lotteries will not change when the common mean changes; i.e., preference between the pair of lotteries can be

determined solely by the ranking of their standard risks.

Result 1 *Assume that the risk-value preference function f is consistent with expected utility theory. Then f can be represented as the following standard risk-value form,*

$$f(\bar{X}, R(X')) = u(\bar{X}) - \varphi(\bar{X})[R(X') - R(0)] \quad (6)$$

if and only if the risk independence condition holds, where $\varphi(\bar{X}) > 0$ and $u(\cdot)$ is a von Neumann and Morgenstern [15] utility function.

Model (6) shows that an expected utility model could have an alternative representation if this risk independence condition holds. If one is risk averse, then $u(\cdot)$ is a concave function, and $R(X') - R(0)$ is always positive. $u(\bar{X})$ provides a measure of value for the mean. If we did not consider the riskiness of a lottery X , it would have the value $u(\bar{X})$. Since it is risky, $u(\bar{X})$ is reduced by an amount proportional to the normalized risk measure $R(X') - R(0)$. $\varphi(\bar{X})$ is a trade-off factor that may depend on the mean. If we further require the utility model to be continuously differentiable, then it must be either a quadratic, exponential, or linear plus exponential model [21].

There are also some other alternative forms of risk-value models within the expected utility framework under different preference conditions [8, 9, 31]. In addition, for nonnegative lotteries such as those associated with the price of a stock, Dyer and Jia [32] propose a relative risk-value framework in the form $X = \bar{X} \times (X/\bar{X})$ that decomposes the return into an average return $\bar{X}$ and a percentage-based risk factor $X/\bar{X}$. We find that this form of a risk-value model is compatible with the logarithmic (or linear plus logarithmic) and the power (or linear plus power) utility functions [32]. Recent empirical studies by Weber *et al.* [33] indicate that this formulation may also be useful as the basis for a descriptive model of the sensitivity of humans and animals to risk.

However, the notion of risk-value trade-offs within the expected utility framework is very limited; for example, consistency with expected utility, based on model (6), imposes very restrictive conditions on the relationship between the risk measure $R(X') = -E[u(X - \bar{X})]$, the value measure $u(\bar{X})$, and the trade-off factor $\varphi(\bar{X}) = u''(\bar{X})/u''(0)$ (for continuously differentiable utility models). In particular, the risk measure and the value measure

must be based on the same utility function. However, a decision maker may deviate from this “consistency” and have different measures for risk and value if his choice is based on risk-value trade-offs.

To be more realistic and flexible in the framework of risk-value trade-offs, we consider a two-attribute structure $(\bar{X}, X')$ for the evaluation of a risky alternative X . In this way, we can explicitly base the evaluation of lotteries on two attributes, mean and risk, so that the mean-risk (or risk-value) trade-offs are not necessarily consistent with the traditional expected utility framework.

We assume the existence of the von Neumann and Morgenstern expected utility axioms over the two-attribute structure $(\bar{X}, X')$ and require the risk-value model to be consistent with the two-attribute expected utility model, i.e., $f(\bar{X}, R(X')) = E[U(\bar{X}, X')]$, where U is a two-attribute utility function. As a special case, when the relationship between $\bar{X}$ and X' is a simple addition, the risk-value model reduces to a traditional expected utility model, i.e., $f(\bar{X}, R(X')) = E[U(\bar{X}, X')] = E[U(\bar{X} + X')] = E[U(X)] = aE[u(X)] + b$, where $a > 0$ and b are constants.

To obtain some separable forms of the risk-value model, we need to have a risk independence condition for the two-attribute structure. Let P^0 be the set of normalized lotteries with zero-expected values, and $>$ a strict preference relation for the two-attribute structure.

Definition 2 For $X', Y' \in P^0$, if there exists a $w_0 \in \text{Re}$ for which $(w_0, X') > (w_0, Y')$, then $(w, X') > (w, Y')$, for all $w \in \text{Re}$.

This two-attribute risk independence condition requires that if two lotteries have the same mean and one is preferred to the other, then transforming the lotteries by adding the same constant to all outcomes will not reverse the preference ordering. This condition is generally supported by our recent experimental studies [34].

Result 2 *Assume that the risk-value preference function f is consistent with the two-attribute expected utility model. Then f can be represented as the following generalized risk-value form,*

$$f(\bar{X}, R(X')) = V(\bar{X}) - \phi(\bar{X})[R(X') - R(0)] \quad (7)$$

if and only if the two-attribute risk independence condition holds, where $\phi(\bar{X}) > 0$ and $R(X')$ is the standard measure of risk.

In contrast to the standard risk-value model (6), the generalized risk-value model (7) provides the flexibility of considering $V(\bar{X})$, $R(X')$ and $\phi(\bar{X})$ independently. Thus we can choose different functions for the value measure $V(\bar{X})$, independent of the utility function. The expected utility measure is used only for the standard measure of risk. Even though expected utility theory has been challenged by some empirical studies for general lotteries, we believe that it should be appropriate for describing risky choice behavior within a special set of normalized probability distributions with the same expected values. For general lotteries with different means, however, our two-attribute risk-value model can deviate from the traditional expected utility preference. In fact, the generalized risk-value model can capture a number of decision paradoxes that violate the traditional expected utility theory [35].

If the utility function u is strictly concave, then $R(X') - R(0) > 0$ and model (7) will reflect risk averse behavior. In addition, if $V(\bar{X})$ is increasing and twice continuously differentiable, $\phi(\bar{X})$ is once continuously differentiable, and $\phi'(\bar{X})/\phi(\bar{X})$ is nonincreasing, then the generalized risk-value model (7) exhibits decreasing risk aversion if and only if $-V''(\bar{X})/V'(\bar{X}) < -\phi'(\bar{X})/\phi(\bar{X})$; and the generalized risk-value model (7) exhibits constant risk aversion if and only if $-V''(\bar{X})/V'(\bar{X}) = -\phi'(\bar{X})/\phi(\bar{X})$ is a constant. Thus, if a decision maker is decreasingly risk averse and has a linear value function, then we must choose a decreasing function for the trade-off factor $\phi(\bar{X})$.

The basic form of the risk-value model may be further simplified if some stronger preference conditions are satisfied. When $\phi(\bar{X}) = k > 0$, model (7) becomes the following additive form:

$$f(\bar{X}, R(X')) = V(\bar{X}) - k[R(X') - R(0)] \quad (8)$$

When $\phi(\bar{X}) = -V(\bar{X}) > 0$, then model (7) reduces to the following multiplicative form:

$$f(\bar{X}, R(X')) = V(\bar{X})R(X') \quad (9)$$

where $R(0) = 1$ and $V(0) = 1$. In this multiplicative model, $R(X')$ serves as a value discount factor due to risk.

We describe measures of perceived risk based on the converse interpretation of the axioms of risk-value models in a companion article in this collection (see **Axiomatic Models of Perceived Risk**). These perceived risk models are simply a negative linear transformation of the risk-value model (7) [30]. Our risk-value framework offers a unified approach to both risk judgment and preference modeling.

Generalized Risk-Value Models

According to the generalized risk-value model (7), the standard measure of risk, the value function, and the trade-off factor can be considered independently. Some examples of the standard measure of risk $R(X')$ are provided in the section titled “The Standard Measure of Risk”. The value measure $V(\bar{X})$ should be chosen as an increasing function and may have the same functional form as a utility model. For appropriate risk averse behavior, the trade-off factor $\phi(\bar{X})$ should be either a decreasing function or a positive constant; e.g., $\phi(\bar{X}) = ke^{-b(\bar{X})}$, where $k > 0$ and $b \geq 0$. We consider three types of risk-value models, namely, moments risk-value models, exponential risk-value models, and generalized disappointment models as follows. For the corresponding perceived risk of each risk-value model (see **Axiomatic Models of Perceived Risk**).

Moments Risk-Value Models

People often use mean and variance to make trade-offs for financial decision making because they provide a reasonable approximation for modeling decision problems and are easy to implement (see [1–3, 36, 37]). In the past, expected utility theory has been used as a foundation for mean–variance models. Risk-value theory provides a better foundation for developing moments models that include the mean–variance model as a special case.

As an example, the mean–variance model, $\bar{X} - kE[(X - \bar{X})^2]$, where $k > 0$, is a simple risk-value model with variance as the standard measure of risk and a constant trade-off factor. Sharpe [36, 37] assumed this mean–variance model in his analysis for portfolio selection and the capital asset pricing model. However, under the expected utility framework, this mean–variance model is based on the assumptions that the investor has an exponential utility function and that returns are jointly normally distributed.

According to our risk-value theory, this mean–variance model is constantly risk averse. To obtain a decreasing risk averse mean–variance model, we can simply use a decreasing function for the trade-off factor:

$$f(\bar{X}, R(X')) = \bar{X} - ke^{-b\bar{X}}E[(X - \bar{X})^2] \quad (10)$$

where $b, k > 0$.

For many decision problems, mean–variance models are an over simplification. On the basis of this risk-value framework, we can specify some richer moment models for risky decision making. First, let us consider the moment standard measure of risk (equation (2)) for the additive risk-value model (8):

$$f(\bar{X}, R(X')) = \bar{X} - k \{E[(X - \bar{X})^2] - cE[(X - \bar{X})^3]\} \quad (11)$$

where $c, k > 0$. The three moments model (11) can be either risk averse or risk seeking, depending on the distribution of a lottery. For symmetric bets or lotteries not highly skewed (e.g., an insurance policy) such that $E[(X - \bar{X})^2] > cE[(X - \bar{X})^3]$, model (11) will be risk averse. But for highly positive skewed lotteries (e.g., lottery tickets) such that the skewness overwhelms the variance, i.e., $E[(X - \bar{X})^2] < cE[(X - \bar{X})^3]$, then model (11) will exhibit risk seeking behavior. Therefore, an individual with preferences described by the moments model (11) would purchase both insurance and lottery tickets simultaneously.

Markowitz [23] noticed that individuals have the same tendency to purchase insurance and lottery tickets whether they are poor or rich. This observed behavior contradicts a common assumption of expected utility theory that preference ranking is defined over ultimate levels of wealth. But whether our risk-value model is risk averse or risk seeking is determined only by the standard measure of risk, which is independent of an individual's wealth level (refer to the form of risk-value model (4)). In particular, for the three moments model (10), the change of wealth level only causes a parallel shift for $f(\bar{X}, R(X'))$ that will not affect the risk attitude and the choice behavior of this model. This is consistent with Markowitz's observation.

Exponential Risk-Value Models

If the standard measure of risk is based on exponential or linear plus exponential utility models, then the

standard measure of risk is given by equation (3). To be compatible with the form of the standard measure of risk, we can also use exponential functions, but with different parameters, for the value measure $V(\bar{X})$ and the trade-off factor $\phi(\bar{X})$, for the generalized risk-value model (7) which then leads to the following model:

$$f(\bar{X}, R(X')) = -he^{-a\bar{X}} - ke^{-b\bar{X}}E[e^{-c(X-\bar{X})} - 1] \quad (12)$$

where a, b, c, h , and k are positive constants. When $a = b = c$ and $h = k$, this model reduces to an exponential utility model; otherwise, these two models are different. When $b > a$, model (12) is decreasing risk averse even though the traditional exponential utility model exhibits constant risk aversion.

As a special case, when $a = b$ and $h = k$, model (12) reduces to the following simple multiplicative form:

$$f(\bar{X}, R(X')) = ke^{-a\bar{X}}E[e^{-c(X-\bar{X})}] \quad (13)$$

This model is constantly risk averse, and therefore has the same risk attitude as an exponential utility model. It has more flexibility since there are two different parameters. This simple risk-value model can be used to explain some well known decision paradoxes [35].

Choosing a linear function or a linear plus exponential function for $V(\bar{X})$ leads to the following models:

$$f(\bar{X}, R(X')) = \bar{X} - ke^{-b\bar{X}}E[e^{-c(X-\bar{X})} - 1] \quad (14)$$

$$f(\bar{X}, R(X')) = \bar{X} - he^{-a\bar{X}} - ke^{-b\bar{X}}E[e^{-c(X-\bar{X})} - 1] \quad (15)$$

Model (14) is decreasingly risk averse. Model (15) includes a linear plus exponential utility model as a special case when $a = b = c$ and $h = k$. It is decreasingly risk averse if $b \geq a$.

Generalized Disappointment Models

Bell [38] proposed a disappointment model for decision making under uncertainty. According to Bell, disappointment is a psychological reaction to an outcome that does not meet a decision maker's expectation. Bell used the mean of a lottery as a decision

maker's psychological expectation. If an outcome smaller than the expected value occurs, the decision maker would be disappointed. Otherwise, the decision maker would be elated. Although Bell's development of the disappointment model has an intuitive appeal, his model is only applicable to lotteries with two outcomes.

Jia *et al.* [27] use the risk-value framework to develop a generalized version of Bell's [38] disappointment model. Consider the following piecewise linear utility model:

$$u(x) = \begin{cases} ex & \text{when } x \geq 0 \\ dx & \text{when } x < 0 \end{cases} \quad (16)$$

where $d, e > 0$ are constant. Decision makers who are averse to downside risk or losses should have $d > e$, as illustrated in Figure 1. The standard measure of risk for this utility model can be obtained as follows:

$$\begin{aligned} R(X') &= dE^- [|X - \bar{X}|] - eE^+ [|X - \bar{X}|] \\ &= [(d - e)/2]E[|X - \bar{X}|] \end{aligned} \quad (17)$$

where $E^- [|X - \bar{X}|] = \sum_{x_i < \bar{X}} p_i |x_i - \bar{X}|$ and $E^+ [|X - \bar{X}|] = \sum_{x_i > \bar{X}} p_i (x_i - \bar{X})$, and $E[|X - \bar{X}|]$ is the absolute standard deviation. Following Bell's [38] basic idea, $dE^- [|X - \bar{X}|]$ represents a general measure of expected disappointment and $eE^+ [|X - \bar{X}|]$ represents a general measure of expected elation. The overall psychological satisfaction is measured by $-R(X')$, which is the converse of the standard measure of risk (equation (17)).

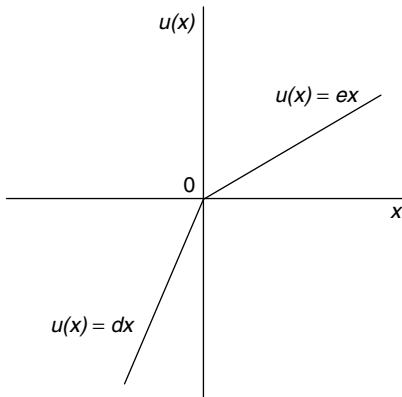


Figure 1 A piecewise linear utility function

If we assume a linear value measure and a constant trade-off factor, then we can have the following risk-value model based on the measure of disappointment risk (equation (17)):

$$\begin{aligned} f(\bar{X}, R(X')) &= \bar{X} - \{dE^- [|X - \bar{X}|] - eE^+ [|X - \bar{X}|]\} \\ &= \bar{X} - [(d - e)/2]E[|X - \bar{X}|] \end{aligned} \quad (18)$$

For a two-outcome lottery, model (18) reduces to Bell's disappointment model. Thus, we call the risk-value model (18) a "generalized disappointment model". Note that the risk-value model (18) will be consistent with the piecewise linear utility model (16) when the lotteries considered have the same means. It is a risk averse model when $d > e$.

Using his two-outcome disappointment model, Bell gave an explanation for the common ratio effect [38]. Our generalized disappointment model (18) can explain the Allais Paradox [10, 11], which involves an alternative with three outcomes [27]. Bell's model and (18) both imply constant risk aversion and are not appropriate for decreasingly risk averse behavior. To obtain a disappointment model with decreasing risk aversion, we can use a decreasing function for the trade-off factor:

$$f(\bar{X}, R(X')) = \bar{X} - ke^{-b\bar{X}}E[|X - \bar{X}|] \quad (19)$$

Bell's disappointment model (18) implies that disappointment and elation are proportional to the difference between the expected value and an outcome. We can use risk model (5) to incorporate nonlinear functions for disappointment and elation in a more general form of disappointment model:

$$\begin{aligned} f(\bar{X}, R(X')) &= \bar{X} - dE^- [|X - \bar{X}|^{\theta_2}] \\ &\quad - eE^+ [|X - \bar{X}|^{\theta_1}] \end{aligned} \quad (20)$$

When $\theta_1 = \theta_2 = 1$, this model reduces to model (18). When $e = 0$ and $\theta_2 = 2$, model (20) becomes a mean-semivariance model.

Finally, our generalized disappointment models are different from Loomes and Sugden's development [39]. In their basic model, disappointment (or elation) is measured by some function of the difference between the utility of outcomes and the expected utility of a lottery. They also assume a linear "utility" measure of wealth and the same sensation intensity for both disappointment and elation, so that

their model has the form, $\bar{X} + E[D(X - \bar{X})]$, where $D(x - \bar{X}) = -D(\bar{X} - x)$, and D is continuously differentiable and convex for $x > \bar{X}$ (thus concave for $x < \bar{X}$). Even though this model is different from our generalized disappointment models (20), it is a special case of our risk-value model with a linear measure of value, a constant trade-off factor, and a specific form of the standard measure of risk (i.e., $R(X') = -E[D(X - \bar{X})]$, where $D(x - \bar{X}) = -D(\bar{X} - x)$). Loomes and Sugden used this model to provide an explanation for the choice behavior that violates Savage's sure-thing principle [40].

Conclusion

We have summarized our efforts to incorporate the intuitively appealing idea of risk-value trade-offs into decision making under risk. The risk-value framework ties together two streams of research: one in modeling risk judgments and the other in developing preference models, and unifies a wide range of decision phenomena including both normative and descriptive aspects.

This development also refines and generalizes a substantial number of previously proposed decision theories and models, ranging from the mean-variance model in finance to the disappointment models in decision sciences. It is also possible to create many new risk-value models. Specifically, we have discussed three classes of decision models based on this risk-value theory: moments risk-value models, exponential risk-value models and generalized disappointment risk-value models. These models are very flexible in modeling preferences. They also provide new resolutions for observed risky choice behavior and the decision paradoxes that violate the independence axiom of the expected utility theory.

The most important assumption in this study is the two-attribute risk independence condition that leads to a separable form of risk-value models. Although some other weaker condition could be used to derive a risk-value model that has more descriptive power, this reduces the elegance of the basic risk-value form, and increases operational difficulty. Butler *et al.* [34] conducted an empirical study of this key assumption, and found some support for it. This study also highlighted some additional patterns of choices indicating that the translation of lottery pairs from the positive domain to the negative domain

often results in the reversal of preference and risk judgments. To capture this phenomenon, we have extended risk independence conditions to allow the trade-off factor in the risk-value models to change sign, and therefore to infer risk aversion in the positive domain and risk seeking in the negative domain. These generalized risk-value models provide additional insights into the reflection effects [12] and related empirical results [14, 26, 41].

Even though some other nonexpected utility theories that have been proposed (e.g., Prospect theory and other weighted utility theories) may produce the same predictions for the decision paradoxes as risk-value theory, it offers a new justification for them based on the appealing and realistic notion of risk-value trade-offs. In particular, since the role of risk is considered *implicitly* in these decision theories and models, they are not compatible with choice behavior that is based on the risk and mean return trade-offs often encountered in financial management, psychology, and other applied fields. Therefore, these theories and models offer little guidance in practice for this type of decision making. We believe that the potential for contributions of these risk-value models in finance is very exciting. Applications of our risk-value models in other fields such as economics, marketing, insurance and risk management are also promising.

Acknowledgment

This article summarizes a stream of research on risk and risk-value models. In particular, we have incorporated materials that appeared previously in the following papers: (a) Jia, J. & Dyer, JS (1995). Risk-value theory, Working Paper, Graduate School of Business, University of Texas at Austin, TX; (b) Jia, J. & Dyer, JS (1996). A standard measure of risk and risk-value models, *Management Science* 42:1961–1705; (c) Jia, J., Dyer, JS. & Butler, JC (1999). Measures of perceived risk, *Management Science* 45:519–532; (d) Jia, J. & Dyer, JS., Decision making based on risk-value theory, in “The Mathematics of Preference, Choice, and Order: Essays in Honor of Peter C. Fishburn”, Edited by SJ Brams, WV Gehrlein and FS Roberts, Springer, New York, 2008.

References

- [1] Markowitz, H.M. (1959). *Portfolio Selection*, John Wiley & Sons, New York.

- [2] Markowitz, H.M. (1987). *Mean-variance Analysis in Portfolio Choice and Capital Markets*, Basil Blackwell, New York.
- [3] Markowitz, H.M. (1991). Foundations of portfolio theory, *Journal of Finance* **XLVI**, 469–477.
- [4] Stone, B. (1973). A general class of 3-parameter risk measures, *Journal of Finance* **28**, 675–685.
- [5] Fishburn, P.C. (1977). Mean-risk analysis with risk associated with below-target returns, *American Economic Review* **67**, 116–126.
- [6] Meyer, J. (1987). Two-moment decision models and expected utility maximization, *American Economic Review* **77**, 421–430.
- [7] Bell, D.E. (1988). One-switch utility functions and a measure of risk, *Management Science* **34**, 1416–1424.
- [8] Bell, D.E. (1995). Risk, return, and utility, *Management Science* **41**, 23–30.
- [9] Sarin, R.K. & Weber, M. (1993). Risk-value models, *European Journal of Operational Research* **70**, 135–149.
- [10] Allais, M. (1953). Le comportement de l'homme rationnel devant le risque, critique des postulats et axiomes de l'école américaine, *Econometrica* **21**, 503–546.
- [11] Allais, M. (1979). The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American school, in *Expected Utility Hypotheses and the Allais Paradox*, M. Allais & O. Hagen, eds., D. Reidel, Dordrecht, Holland, pp. 27–145.
- [12] Kahneman, D.H. & Tversky, A. (1979). Prospect theory: an analysis of decision under risk, *Econometrica* **47**, 263–290.
- [13] Machina, M.J. (1987). Choice under uncertainty: problems solved and unsolved, *Economic Perspectives* **1**, 121–154.
- [14] Weber, E.U. (2001). Decision and choice: risk, empirical studies, in *International Encyclopedia of the Social Sciences*, N. Smelser & P. Baltes, eds, Elsevier Science, Oxford, pp. 13347–13351.
- [15] von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.
- [16] Pratt, J.W. (1964). Risk aversion in the small and in the large, *Econometrica* **32**, 122–136.
- [17] Dyer, J.S. & Sarin, R.K. (1982). Relative risk aversion, *Management Science* **28**, 875–886.
- [18] Quiggin, J. (1982). A theory of anticipated utility, *Journal of Economic Behavior and Organization* **3**, 323–343.
- [19] Tversky, A. & Kahneman, D.H. (1992). Advances in prospect theory: cumulative representation of uncertainty, *Journal of Risk and Uncertainty* **5**, 297–323.
- [20] Wu, G. & Gonzalez, R. (1996). Curvature of the probability weighting function, *Management Science* **42**, 1676–1690.
- [21] Jia, J. & Dyer, J.S. (1996). A standard measure of risk and risk-value models, *Management Science* **42**, 1691–1705.
- [22] Friedman, M. & Savage, L.P. (1948). The utility analysis of choices involving risk, *Journal of Political Economy* **56**, 279–304.
- [23] Markowitz, H.M. (1952). The utility of wealth, *Journal of Political Economy* **60**, 151–158.
- [24] Weber, M. (1990). *Risikoentscheidungskalkule in der Finanzierungstheorie*, Poeschel, Stuttgart.
- [25] Sarin, R.K. (1987). Some extensions of Luce's measures of risk, *Theory and Decision* **22**, 25–141.
- [26] Fishburn, P.C. & Kochenberger, G.A. (1979). Two-piece von Neumann-Morgenstern utility functions, *Decision Sciences* **10**, 503–518.
- [27] Jia, J., Dyer, J.S. & Butler, J.C. (2001). Generalized disappointment models, *Journal of Risk and Uncertainty* **22**, 59–78.
- [28] Konno, H. & Yamazaki, H. (1992). Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market, *Management Science* **37**, 519–531.
- [29] Dennenberg, D. (1990). Premium calculation: why standard deviation should be replaced by absolute deviation, *ASTIN Bulletin* **20**, 181–190.
- [30] Jia, J., Dyer, J.S. & Butler, J.C. (1999). Measures of perceived risk, *Management Science* **45**, 519–532.
- [31] Dyer, J.S. & Jia, J. (1998). Preference conditions for utility models: a risk-value perspective, *Annals of Operations Research* **80**, 167–182.
- [32] Dyer, J.S. & Jia, J. (1997). Relative risk-value models, *European Journal of Operational Research* **103**, 170–185.
- [33] Weber, E.U., Shafir, S. & Blais, A. (2004). Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation, *Psychological Review* **111**, 430–445.
- [34] Butler, J., Dyer, J. & Jia, J. (2005). An empirical investigation of the assumptions of risk-value models, *Journal of Risk and Uncertainty* **30**, 133–156.
- [35] Jia, J. (1995). Measures of risk and risk-value theory, unpublished Ph.D. Dissertation, University of Texas at Austin, Texas.
- [36] Sharpe, W.F. (1970). *Portfolio Theory and Capital Markets*, McGraw-Hill, New York.
- [37] Sharpe, W.F. (1991). Capital asset prices with and without negative holdings, *Journal of Finance* **XLVI**, 489–509.
- [38] Bell, D.E. (1985). Disappointment in decision making under uncertainty, *Operations Research* **33**, 1–27.
- [39] Loomes, G. & Sugden, R. (1986). Disappointment and dynamic consistency in choice under uncertainty, *Review of Economic Studies* **LIII**, 271–282.
- [40] Savage, L.J. (1954). *The Foundations of Statistics*, John Wiley & Sons, New York.
- [41] Payne, J.W., Laughhunn, D.J. & Crum, R. (1981). Further tests of aspiration level effects in risky choice behavior, *Management Science* **27**, 953–958.

Further Reading

Payne, J.W., Laughhunn, D.J. & Crum, R. (1980). Translation of gambles and aspiration level effects in risky choice behavior, *Management Science* **26**, 1039–1060.

Related Articles

Credit Risk Models

Individual Risk Models

Utility Function

Value at Risk (VaR) and Risk Measures

JIANMIN JIA, JAMES S. DYER AND
JOHN C. BUTLER

Axiomatic Models of Perceived Risk

In a separate article in this encyclopedia (*see* **Axiomatic Measures of Risk and Risk-Value Models**), we discussed a standard measure of risk based on the converse of the expected utility of lotteries with zero-expected values, and showed how it could be used to derive risk-value models for risky choices within both an expected utility framework and a non-expected utility framework (see also [1–3]). Though this standard measure of risk has some descriptive power for risk judgments, it is more normative in nature. In particular, since the standard measure of risk eliminates the effect of the mean of a lottery, it only measures the “pure” risk of the lottery, and may not be appropriate for modeling perceptions of risk. The purpose of this article is to describe a two-attribute structure of perceived risk that allows the mean to impact judgments of risk perception and that can also be related to risk-value models for risky choice.

Over the last 30 years, researchers have expended much effort toward developing and testing models of the perceived riskiness of lotteries. Pollatsek and Tversky [4] provide an early summary of risk research that is still meaningful today:

The various approaches to study of risk share three basic assumptions. 1. Risk is regarded as a property of options (e.g., gambles, courses of action) that affects choices among them. 2. Options can be meaningfully ordered with respect to their riskiness. 3. The risk of an option is related in some way to the dispersion, or the variance, of its outcomes.

As stated in assumption 1, risk is a characteristic of a lottery that affects decisions. This is the primary motivation for studying the nature of perceived risk. A measure of perceived risk may be used as a variable in preference models, such as Coombs’ portfolio theory [5–8], in which a choice among lotteries is a compromise between maximizing expected value and optimizing the level of perceived risk. However, the “risk” measure in Coombs’ Portfolio theory is left essentially undefined, and is considered to be an independent theory. This has stimulated a long stream of research on the measure of perceived risk.

Empirical studies have demonstrated that people are consistently able to order lotteries with respect to

their riskiness, and that risk judgments satisfy some basic axioms (e.g., see [9, 10]). Thus, as stated in assumption 2, the term *riskiness* should be both a meaningful and measurable characteristic of lotteries.

There have been many refinements to assumption 3 as experimental results have exposed how perceived-risk judgments change as a function of the characteristics of the lotteries considered. Some stylized facts regarding risk judgments include the following:

- Perceived risk increases when there is an increase in range, variance, or expected loss, e.g., see [11].
- Perceived risk decreases if a constant positive amount is added to all outcomes of a lottery [9, 12].
- Perceived risk increases if all outcomes of a lottery with zero mean are multiplied by a positive constant greater than one [6].
- Perceived risk increases if a lottery with zero-mean is repeated many times [6].

These empirically verified properties provide basic guidelines for developing and evaluating measures of perceived risk.

In the following section, we give a review of previously proposed models of perceived risk and discuss their performance in empirical studies. In the section titled “Two-Attribute Models for Perceived Risk”, we present our measures of perceived risk based on a two-dimensional structure of the standard risk of a lottery and its mean, and show that it includes many of these previously proposed models as special cases. Finally, in the concluding section, we provide a summary and discuss the implications of the two-attribute perceived-risk structure in decision making under risk.

Review of Perceived Risk Models

The literature contains various attempts to define risk based on different assumptions about how perceived risk is formulated and how it evolves as the risky prospects under consideration are manipulated. In this section, we review some previously proposed models for perceived risk and their key assumptions. We focus on those that are closely related to a preference-dependent measure of risk that is compatible with traditional expected utility theory.

A more detailed review of perceived-risk studies, including some measures of risk that are not closely related to preference, is provided by Brachinger and Weber [13]. We also exclude from our discussion the “coherent measures of risk” developed in the financial mathematics literature by Artzner *et al.* [14, 15] which produce results in monetary units (for example, in dollars) that estimate the “expected shortfall” associated with a portfolio. For a recent review of the latter, see Acerbi [16].

Studies by Coombs and His Associates

In early studies, risk judgments were represented by using the moments of a distribution and their transformations, i.e., distributive models. Expected value, variance, skewness, range, and the number of repeated plays have been investigated as possible determinants of risk judgments [6, 11, 17]. Coombs and Huang [7] considered several composition functions of three indices corresponding to transformations on two-outcome lotteries, and their paper supported a distributive model that is based on a particular structure of the joint effect of these transformations on perceived risk. However, evidence to the contrary of such a distributive model was also found by Barron [18].

Coombs and Lehner [12] used distribution parameters as variables in the distributive model to test if moments of distributions are useful in assessing risk. For a lottery $(b + 2(1 - p)a, p; b - 2pa)$ (this means that the lottery has an outcome $(b + 2(1 - p)a)$ with probability p and an outcome $(b - 2pa)$, otherwise), which has a mean equal to b and range $2a$, the distributive model is represented by

$$R(a, b, p) = [\phi_1(a) + \phi_2(b)]\phi_3(p) \quad (1)$$

where R is a riskiness function and ϕ_1, ϕ_2 , and ϕ_3 are real-valued monotonic functions defined on a, b , and p , respectively. Coombs and Lehner [12] showed that the distributive model (1) is not acceptable as a descriptive model of risk judgment. They concluded that complex interactions between the properties of risky propositions prevent a simple polynomial expression of the variables a, b , and p , from capturing perceived riskiness.

Coombs and Lehner [19] further considered perceived risk as a direct function of outcomes and probabilities, with no intervening distribution parameters. They assumed a bilinear model. In the case

of just three outcomes (positive, negative, and zero), perceived risk is represented by the following model:

$$R(X) = \phi_1(p)\phi_2(w) + \phi_3(q)\phi_4(l) \quad (2)$$

where w and l represent the positive and negative outcomes, with probabilities p and q ($p + q = 1$), respectively; R and ϕ_i ($i = 1, 2, 3$, and 4) are real-valued functions and X represents the lottery. The model assumes that a zero outcome and its associated probability have no direct effect on perceived risk. The form of model (2) is similar to Prospect theory [20]. Coombs and Lehner’s [19] experiment supported the notion that perceived risk can be decomposed into contributions from good and bad components, and the bad components play a larger role than the good ones.

Pollatsek and Tversky’s Risk Theory

An important milestone in the study of perceived risk is the axiomatization of risk theory developed by Pollatsek and Tversky [4]. They assumed four axioms for a risk system: (a) weak ordering; (b) cancellation (or additive independence); (c) solvability; and (d) an Archimedean property. Let P denote a set of simple probability distributions or lotteries $\{X, Y, Z, \dots\}$ and $\geq_R$ be a binary risk relation (meaning at least as risky as). For convenience, we use X, Y , and Z to refer to random variables, probability distributions, or lotteries interchangeably. Pollatsek and Tversky showed that the four axioms imply that there exists a real-valued function R on P such that for lotteries X and Y : (i) $X \geq_R Y$ if and only if $R(X) \geq R(Y)$; (ii) $R(X \circ Y) = R(X) + R(Y)$, where “ $\circ$ ” denotes the binary operation of adding independent random variables; that is, the convolution of their density functions.

Pollatsek and Tversky considered three additional axioms: (e) positivity; (f) scalar monotonicity; and (g) continuity. These three additional axioms imply that R is a linear combination of mean and variance:

$$R(X) = -\theta \bar{X} + (1 - \theta)E[(X - \bar{X})^2] \quad (3)$$

where $0 < \theta < 1$.

However, the empirical validity of equation (3) was criticized by Coombs and Bowen [21] who showed that factors other than mean and variance,

such as skewness, affect perceived risk. In Pollatsek and Tversky's system of axioms, the continuity condition based on the central limit theorem is directly responsible for the form of the mean–variance model (3). Coombs and Bowen [21] showed that skewness impacts perceived risk even under multiple plays of a lottery, when the effect of the central limit theorem modifies the effect of skewness.

Another empirically questionable axiom is the additive independence condition, which says that, for X, Y , and Z in P , $X \geq_R Y$ if and only if $X \circ Z \geq_R Y \circ Z$. Fishburn [22] provides the following example of a setting where additive independence is unlikely to hold. Many people feel that a lottery $X = (\$1000, 0.01; -\$10\,000)$ (i.e., X has probability 0.01 of a \$1000 gain, and a \$10 000 loss otherwise) is riskier than $Y = (\$2000, 0.5; -\$12\,000)$. Consider another degenerate lottery Z that has a sure \$11 000 gain. Since $X \circ Z$ yields at least a \$1000 gain while $Y \circ Z$ results in a loss of \$1000 with probability 0.5, it seems likely that most people would consider $Y \circ Z$ to be riskier than $X \circ Z$. This risk judgment pattern is inconsistent with additive independence. Empirical studies have also failed to support the independence condition [23, 24].

Nevertheless, some of Pollatsek and Tversky's axioms, such as positivity and scalar monotonicity, are very appealing. Because they are important to the present article, we briefly introduce them here. According to the positivity axiom, if K is a degenerate lottery with an outcome $k > 0$, then $X \geq_R X \circ K$ for all X in P . In other words, the addition of a positive sure-amount to a lottery would decrease its perceived risk. This quality is considered an essential property of perceived risk and has been confirmed by several empirical studies (e.g. [9, 12]).

Another appealing axiom in Pollatsek and Tversky's theory is scalar monotonicity, which says, for all X, Y in P with $E(X) = E(Y) = 0$, (a) $\beta X \geq_R X$ for $\beta > 1$; (b) $X \geq_R Y$ if and only if $\beta X \geq_R \beta Y$ for $\beta > 0$. This axiom asserts that, for lotteries with zero expectation, risk increases when the lottery is multiplied by a real number $\beta > 1$ (also see [6]), and that the risk ordering is preserved upon a scale change of the lotteries (e.g., dollars to pennies). Pollatsek and Tversky regarded the positivity axiom and part (a) of the monotonicity axiom as necessary assumptions for any theory of risk.

In a more recent study, Rotar and Sholomitsky [25] weakened part (b) of the scalar monotonicity axiom (coupled with some other additional conditions) to arrive at a more flexible risk model that is a finite linear combination of cumulants of higher orders. This generalized risk model can take into account additional characteristics of distributions such as skewness and other higher-order moments. However, because Rotar and Sholomitsky's risk model still retains the additive independence condition as a basic assumption, their model would be subject to the same criticisms regarding the additivity of risk.

Luce's Risk Models and Others

Subsequent to the criticisms of Pollatsek and Tversky's risk measure, Luce [26] approached the problem of risk measurement in a different way. He began with a multiplicative structure of risk. First, Luce considered the effect of a change of scale on risk, multiplying all outcomes of a lottery by a positive constant. He assumed two simple possibilities, an additive effect and a multiplicative effect, presented as follows:

$$R(\alpha * X) = R(X) + S(\alpha) \quad (4)$$

or

$$R(\alpha * X) = S(\alpha) R(X) \quad (5)$$

where α is a positive constant and S is a strictly increasing function with $S(1) = 0$ for equation (4) and $S(1) = 1$ for equation (5). Luce's assumptions (4) and (5) are related to Pollatsek and Tversky's [4] scalar monotonicity axiom. However, an important difference is that Pollatsek and Tversky only applied this assumption to the lotteries with zero-expected values. As we see later, this may explain why Luce's models have not been supported by experiments.

Next, Luce considered two ways in which the outcomes and probabilities of a lottery could be aggregated into a single number. The first aggregation rule is analogous to the expected utility form and leads to an expected risk function:

$$R(X) = \int_{-\infty}^{\infty} T(x) f(x) dx = E[T(X)] \quad (6)$$

where T is some transformation function of the random variable X and $f(x)$ is the density of

lottery X . In the second aggregation rule, the density goes through a transformation before it is integrated,

$$R(X) = \int_{-\infty}^{\infty} T^+[f(x)] dx \quad (7)$$

where T^+ is some nonnegative transformation function. The combinations of the two decomposition rules (4) and (5) and the two aggregation rules (6) and (7) yield four possible measures of risk as follows:

1. by rules (4) and (6),

$$R(X) = a \int_{x \neq 0} \log |x| dx + b_1 \int_{-\infty}^0 f(x) dx + b_2 \int_0^{\infty} f(x) dx, \quad a > 0 \quad (8)$$

2. by rules (5) and (6),

$$R(X) = a_1 \int_{-\infty}^0 |x|^\theta dx + a_2 \int_0^{\infty} x^\theta dx, \quad \theta > 0 \quad (9)$$

3. by rules (4) and (7),

$$R(X) = -a \int_{-\infty}^{\infty} f(x) \log f(x) dx + b, \quad a > 0, b = 0 \quad (10)$$

4. by rules (5) and (7),

$$R(X) = a \int_{-\infty}^{\infty} f(x)^{1-\theta} dx, \quad a, \theta > 0 \quad (11)$$

Luce's risk models did not receive positive support from an experimental test by Keller *et al.* [9]. As Luce himself noted, an obvious drawback of models (10) and (11) is that both measures require that risk should not change if we add or subtract a constant amount to all outcomes of a lottery. This is counter to intuition and to empirical evidence [9, 12]. Luce's absolute logarithmic measure (equation 8) is also neither empirically valid [9] nor prescriptively valid [27]. In the experiment by Keller *et al.* [9], only the absolute power model (9) seems to have some promise as a measure of perceived risk.

Following Luce's work, Sarin [27] considered the observation that when a constant is added to all outcomes of a lottery, perceived risk should decrease.

He assumed that there is a strictly monotonic function S such that for all lotteries and any real number β ,

$$R(\beta \circ X) = S(\beta)R(X) \quad (12)$$

Together with the expectation principle (rule 6), this yields an exponential form of risk model,

$$R(X) = kE(e^{cX}) \quad (13)$$

where $kc < 0$.

Luce and Sarin's models employ the expectation principle, which was first postulated for risk by Huang [28]. The expectation principle – an application of the independence axiom of expected utility theory – means that the risk measure R is linear under convex combinations:

$$R(\lambda X + (1 - \lambda)Y) = \lambda R(X) + (1 - \lambda)R(Y) \quad (14)$$

where $0 < \lambda < 1$. Empirical studies have suggested that this assumption may not be valid for risk judgments [9, 19]. Weber and Bottom [10] showed, however, that the independence axiom is not violated for risk judgments, but that the culprit is the so-called probability accounting principle (see [29]). These findings cast doubt on any perceived-risk models based on the expectation principle, including Luce's logarithmic model (8) and power model (9), and Sarin's exponential model (13).

Sarin also generalized the simple expectation principle using Machina's [30] nonexpected utility theory and extended Luce's models (8 and 9) into more complicated risk models. However, since risk judgment is not identical to choice preference under risk, Sarin's proposal needs to be tested empirically.

Luce and Weber [29] proposed a revision of Luce's original power model (8) based on empirical findings. This conjoint expected risk (CER) model has the following form:

$$\begin{aligned} R(X) = & A_0 \Pr(X = 0) \\ & + A_+ \Pr(X > 0) + A_- \Pr(X < 0) \\ & + B_+ E[X^{K_+} | X > 0] \Pr(X > 0) \\ & + B_- E[|X|^{K_-} | X < 0] \Pr(X < 0) \end{aligned} \quad (15)$$

where A_0 , A_+ , and A_- are probability weights, and B_+ and B_- are weights of the conditional expectations, raised to some positive powers, K_+ and K_- . The major advantage of the CER model is that it

allows for asymmetric effects of transformations on positive and negative outcomes. Weber [31] showed that the CER model describes risk judgments reasonably well. One possible drawback of the CER model is that the lack of parsimony provides the degrees of freedom to fit any set of responses.

Weber and Bottom [10] tested the adequacy of the axioms underlying the CER model and found that the conjoint structure assumptions about the effect of change of scale transformations on risk hold for negative-outcome lotteries, but not for positive-outcome lotteries. This suggests that the multiplicative independence assumption (i.e., for positive (or negative)-outcome-only lotteries X and Y , $X \geq_R Y$ if and only if $\beta X \geq_R \beta Y$ for $\beta > 0$) may not be valid. Note that Pollatsek and Tversky's [4] scalar monotonicity axiom is identical to multiplicative independence, but is only assumed to hold for lotteries with zero-expected values.

Fishburn's Risk Systems

Fishburn [31, 32] explored risk measurement from a rigorous axiomatic perspective. In his two-part study on axiomatizations of perceived risk, he considered lotteries separated into gains and losses relative to a target, say 0. Then the general set P of measures can be written as $P = \{(\alpha, p; \beta, q) : \alpha \geq 0, \beta \geq 0, \alpha + \beta \leq 1, p \text{ in } P^-, q \text{ in } P^+\}$, where α is the loss probability, p is the loss distribution given a loss, β is the gain probability, q is the gain distribution given a gain, $1 - p - q$ is the probability for the target outcome 0, and P^- and P^+ are the sets of probability measures defined on loss and gain, respectively. The risk measure in this general approach satisfies $(\alpha, p; \beta, q) \geq_R (\gamma, r; \delta, s)$ if and only if $R(\alpha, p; \beta, q) \geq R(\gamma, r; \delta, s)$.

Fishburn assumed that there is no risk if and only if there is no chance of getting a loss, which implies $R = 0$ if and only if $\alpha = 0$. This rules out additive forms of risk measures like $R(\alpha, p; \beta, q) = R_1(\alpha, p) + R_2(\beta, q)$, but allows forms that are multiplicative in losses and gains; e.g., $R(\alpha, p; \beta, q) = \rho(\alpha, p) \times \tau(\beta, q)$. If ρ and τ are further decomposable, then the risk measure can be

$$R(\alpha, p; \beta, q) = \left[\rho_1(\alpha) \int_{x < 0} \rho_2(x) dp(x) \right] \times \left[1 - \tau_1(\beta) \int_{y > 0} \tau_2(y) dq(y) \right] \quad (16)$$

According to Fishburn [33], the first part of this model measures the "pure" risk involving losses and the second measures the effect of gains on the risk. In the multiplicative model (16), gains proportionally reduce risk independent of the particular (α, p) involved (unless the probability of a loss, α , is zero, in which case there is no risk to be reduced). Fishburn did not suggest functional forms for the free functions in his models, so it is difficult to test them empirically.

In summary, since the pioneering work of Coombs and his associates' on perceived risk, several formal theories and models have been proposed. But none of these risk models is fully satisfactory. As Pollatsek and Tversky [4] wrote, "... our intuitions concerning risk are not very clear and a satisfactory operational definition of the risk ordering is not easily obtainable." Nevertheless, empirical studies have observed a remarkable consistency in risk judgments [9, 10] suggesting the existence of robust measures of perceived risk.

Two-Attribute Models for Perceived Risk

In this section, we propose two-attribute models for perceived risk based on the mean of a lottery and the standard measure of risk that is discussed in another article in this collection (see **Axiomatic Measures of Risk and Risk-Value Models**). In particular, our measures of perceived risk can be incorporated into preference models based on the notion of risk-value trade-offs. For the purpose of application, we suggest several explicit functional forms for the measures of perceived risk.

A Two-Attribute Structure for Perceived Risk

A common approach in previous studies of perceived risk is to look for different factors underlying a lottery that are responsible for risk perceptions, such as mean and variance or other risk dimensions, and then consider some separation or aggregation rules to obtain a risk measurement model (see [34] for a review). Jia and Dyer [2] decomposed a lottery X into its mean $\bar{X}$ and its standard risk, $X' = X - \bar{X}$, and proposed a standard measure of risk based on expected utility theory:

$$R(X') = -E[u(X')] = -E[u(X - \bar{X})] \quad (17)$$

where $u(\cdot)$ is a von Neumann and Morgenstern [35] utility function. The mean of a lottery serves as a status quo for measuring the standard risk (see **Axiomatic Measures of Risk and Risk-Value Models**) for a summary of this development.

The standard measure of risk has many desirable properties that characterize the “pure” risk of lotteries. It can provide a suitable measure of perceived risk for lotteries with zero-expected values as well. However, the standard measure of risk would not be appropriate for modeling people’s perceived risk for general lotteries since the standard measure of risk is independent of expected value or any certain pay-offs. That is, if $Y = X + k$, where k is a constant, then $Y' = Y - \bar{Y} = X - \bar{X} = X'$. As we discussed earlier, empirical studies have shown that people’s perceived risk decreases as a positive constant amount is added to all outcomes of a lottery.

To incorporate the effect of the mean of a lottery on perceived risk, we consider a two-attribute structure for evaluating perceived risk; that is, $(\bar{X}, X')$. In fact, a lottery X can be represented by its expected value $\bar{X}$ and the standard risk X' exclusively, e.g., $X = \bar{X} + X'$. Thus, $(\bar{X}, X')$ is a natural extension of the representation of the lottery X . This two-attribute structure has an intuitive interpretation in risk judgment. When people make a risk judgment, they may first consider the variation or uncertainty of the lottery, measured by X' , and then take into account the effect of expected value on the uncertainty perceived initially, or *vice versa*.

To develop our measure of perceived risk formally, let $\mathbf{P}$ be the set of all simple probability distributions, including degenerate distributions, on a nonempty product set, $\mathbf{X}_1 \times \mathbf{X}_2$, of outcomes, where $\mathbf{X}_i \subseteq \text{Re}$, $i = 1, 2$, and Re is the set of real numbers. For our special case, the outcome of a lottery X on $\mathbf{X}_1$ is fixed at its mean $\bar{X}$; thus, the marginal distribution on $\mathbf{X}_1$ is degenerate with a singleton outcome $\bar{X}$. Because $\bar{X}$ is a constant, the two “marginal distributions” $(\bar{X}, X')$ are sufficient to determine a unique distribution in $\mathbf{P}$.

Let $>_{\bar{R}}$ be a strict risk relation, $\sim_{\bar{R}}$ an indifference risk relation, and $\geq_{\bar{R}}$ a weak risk relation on $\mathbf{P}$. We assume a two-attribute case of the expectation principle and other necessary conditions analogous to those of multiattribute utility theory (e.g., see [36, 37]) for the risk ordering $>_{\bar{R}}$ on $\mathbf{P}$, such that for all $(\bar{X}, X'), (\bar{Y}, Y') \in \mathbf{P}$, $(\bar{X}, X') >_{\bar{R}} (\bar{Y}, Y')$ if and only if $R_p(\bar{X}, X') > R_p(\bar{Y}, Y')$, where R_p is defined as

follows:

$$R_p(\bar{X}, X') = E[U_R(\bar{X}, X')] \quad (18)$$

and U_R is a real-valued function unique up to a positive linear transformation. Note that because the marginal distribution for the first attribute is degenerate, the expectation, in fact, only needs to be taken over the marginal distribution for the second attribute, which, in turn, is the original distribution of a lottery X normalized to a mean of zero.

Basic Forms of Perceived-Risk Models

Model (18) provides a general measure of perceived risk based on two attributes, the mean and standard risk of a lottery. To obtain separable forms of perceived-risk models, we make the following assumptions about risk judgments.

Assumption 1. For $X', Y' \in \mathbf{P}^0$, if there exists a $w^o \in \text{Re}$ for which $(w^o, X') >_{\bar{R}} (w^o, Y')$, then $(w, X') >_{\bar{R}} (w, Y')$ for all $w \in \text{Re}$.

Assumption 2. For $X', Y' \in \mathbf{P}^0$, $(0, X') >_{\bar{R}} (0, Y')$ if and only if $X' >_R Y'$.

Assumption 3. For $(\bar{X}, X') \in \mathbf{P}$, then $(\bar{X}, X') >_{\bar{R}} (\bar{X} + \Delta, X')$ for any constant $\Delta > 0$.

Assumption 1 is an independence condition, which says that the risk ordering for two lotteries with the same mean will not switch when the common mean changes to any other value. Compared with Pollatsek and Tversky’s [4] additive independence condition, Assumption 1 is weaker since it features a pair of lotteries with the same mean and a common constant rather than a common lottery. Coombs [8] considered a similar assumption for a riskiness ordering; i.e., $X \geq_R Y$ if and only if $X + k \geq_R Y + k$, where $E(X) = E(Y)$ and k is a constant. However, our formulation is based on a two-attribute structure, which leads to a separable risk function for $\bar{X}$ and X' , as we shall discuss.

Assumption 2 postulates a relationship between the two risky binary relations, $>_{\bar{R}}$ and $>_R$ (where $>_R$ is a strict risk relation on $\mathbf{P}^0$), so that for any zero-expected lotteries, the risk judgments made by $R_p(0, X')$ and by the standard measure of risk $R(X')$ are consistent. The last assumption implies that if two lotteries have the same “pure” risk, X' , then the lottery with a larger mean will be perceived less

risky than the one with a lower mean as suggested by previous studies (e.g. [9, 12]).

Result 1 [38]. The two-attribute perceived-risk model (18) can be decomposed into the following form:

$$R_p(\bar{X}, X') = g(\bar{X}) + \psi(\bar{X})R(X') \quad (19)$$

if and only if assumptions 1–3 are satisfied, where $\psi(\bar{X}) > 0$, $g'(\bar{X}) < -\psi'(\bar{X})R(X')$, and $R(X')$ is the standard measure of risk.

According to this result, perceived risk can be constructed by a combination of the standard measure of risk and the effect of the mean. Result 1 postulates a constraint on the choice of functions $g(\bar{X})$ and $\psi(\bar{X})$ in model (19). If $\psi(\bar{X})$ is a constant, then the condition $g'(\bar{X}) < -\psi'(\bar{X})R(X')$ becomes $g'(\bar{X}) < 0$; i.e., $g(\bar{X})$ is a decreasing function of $\bar{X}$. Otherwise, a nonincreasing function $g(\bar{X})$ and a decreasing function $\psi(\bar{X})$ should be sufficient to satisfy the condition $g'(\bar{X}) < -\psi'(\bar{X})R(X')$ when $R(X') > 0$.

For risk judgments, we may require that any degenerate lottery should have no risk (e.g., [39–41]). The concept of risk would not be evoked under conditions of certainty; no matter how bad a certain loss may be, it is a sure thing and, therefore, riskless. This point of view can be represented by the following assumption.

Assumption 4. For any $w \in \text{Re}$, $(w, 0) \sim_{\bar{R}} (0, 0)$.

Result 2 [38]. The two-attribute perceived-risk model (19) can be represented as follows:

$$R_p(\bar{X}, X') = \psi(\bar{X})[R(X') - R(0)] \quad (20)$$

if and only if assumptions 1–4 are satisfied, where $\psi(\bar{X}) > 0$ is a decreasing function of the mean $\bar{X}$, $\psi(\bar{X})[R(X') - R(0)]$ is the standard measure of risk, and $R(0) = -u(0)$ is a constant.

When $g(\bar{X}) = -R(0)\psi(\bar{X})$ as required by Assumption 4, the general risk model (19) reduces to the multiplicative risk model (20). This multiplicative risk model captures the effect of the mean on perceived riskiness in an appealing way; increasing (decreasing) the mean reduces (increases) perceived riskiness in a proportional manner.

Finally, note that the two-attribute perceived-risk models (19) and (20) are not simple expected forms; we decompose a lottery into a two-attribute structure and only assume the expectation principle holds for normalized lotteries with zero-expected values. For general lotteries with non zero-expected values, the

underlying probabilities of lotteries can influence the standard measure of risk in a nonlinear fashion via $g(\bar{X}) = -R(0)\psi(\bar{X})$, $\psi(\bar{X})$. Thus, models (19) and (20) avoid the drawbacks of expected risk models because the two-attribute expected utility axioms will not generally result in linearity in probability in the perceived-risk models.

Relationship between Perceived Risk and Preference

An important feature of the two-attribute approach to modeling risk is that the derived measures of perceived risk can be treated as a stand alone primitive concept and can also be incorporated into preference models in a clear fashion. As summarized in the complementary article, we proposed a risk-value theory for preference modeling also by decomposing a lottery into a two-attribute structure, the mean of the lottery, and its standard risk (see **Axiomatic Measures of Risk and Risk-Value Models**).

A general form of the risk-value model can be represented as follows:

$$\bar{X} - \phi(\bar{X})[R(X') - R(0)] \quad (21)$$

where $f(\bar{X}, X')$ represents a preference function based on the mean of a lottery and its standard risk, $V(\bar{X})$ is a subjective value measure for the mean of a lottery, $\phi(\bar{X}) > 0$ is a trade-off factor that may depend on the mean, and the other notations are the same as in equation (20). In general, a decreasing trade-off factor $\phi(\bar{X})$ is required in risk-value theory, which implies that the intensity of the risk effect on preference decreases as a positive constant amount is added to all outcomes of a lottery. Since the risk-value model (21) is based on the two-attribute expected utility axioms, and the perceived-risk model (19) is derived by using the reverse interpretation of the same axioms, the two types of models must be a negative linear transformation of each other, i.e., $\bar{X} - \phi(\bar{X})[R(X') - R(0)]R_p(\bar{X}, X')$, where $a > 0$ and b are constants. Several previously proposed measures of perceived risk also have the implication that their converse forms may be used for preference modeling (e.g. [4, 27, 30]).

The relationships between the functions in models (19) and (21) can be clarified by transforming the perceived-risk model (19) into another representation similar to the risk-value model (21).

When $X = \bar{X}$, equation (19) becomes $R_p(\bar{X}, X') = g(\bar{X}) + \psi(\bar{X})R(X')$. Let $h(\bar{X}) = R_p(\bar{X}, X')$, then $g(\bar{X}) = g(\bar{X}) + \psi(\bar{X})R(X')$. Substituting this into equation (19), we obtain an alternative representation of the perceived-risk measure, $R_p(\bar{X}, X') = g(\bar{X}) + \psi(\bar{X})R(X')$. On the basis of our risk-value theory (result 1), we can have $h(\bar{X}) = -aV(\bar{X}) + b$, and $\psi(\bar{X}) = a\phi(\bar{X})$, where $a > 0$ and b are constants.

The measure of perceived risk (equation 20) has more intuitive appeal in constructing preference based on risk-value trade-offs. Substituting $\phi(\bar{X}) = (1/a)\psi(\bar{X})$ into equation (21), we have $\bar{X} - \phi(\bar{X})[R(X') - R(0)] = V(\bar{X}) - (1/a)\psi(\bar{X})[R(X') - R(0)] = R_p(\bar{X}, X')$. This representation is consistent with an explicit trade-off between perceived risk and value in risky decision making. This provides a clear link between a riskiness ordering and a preference ordering, and shows an explicit role of risk perceptions in decision making under risk.

Some Examples

When $g(\bar{X})$ is linear, $\psi(\bar{X})$ is constant, and $R(X')$ is the variance, the risk model (19) reduces to Pollatsek and Tversky's [4] mean–variance model (1), as a special case. But Pollatsek and Tversky's risk model may be considered oversimplified. To obtain Rotar and Sholomitsky's [25] generalized moments model, the standard measure of risk should be based on a polynomial utility model.

We can select some appropriate functional forms for $g(\bar{X})$, $\psi(\bar{X})$, and $R(X')$ to construct specific instances of equation (19). In **Axiomatic Measures of Risk and Risk-Value Models**, we have proposed some explicit models for the standard measure of risk $R(X')$. Those models can be used directly in constructing functional forms of perceived-risk models (19) and (20). An example for $\psi(\bar{X})$ is $\psi(\bar{X}) = ke^{-b\bar{X}}$, where $k > 0$ and $b \geq 0$ (when $b = 0$, $\psi(\bar{X})$ becomes a constant k), and a simple choice for $g(\bar{X})$ is $g(\bar{X}) = -a\bar{X}$, where $a > 0$ is a constant. Some functional forms of the perceived-risk model (19) based on these choices of $\psi(\bar{X})$ and $g(\bar{X})$ are the following:

$$R_p(\bar{X}, X') = -a\bar{X} + ke^{-b\bar{X}}E[e^{-c(X-\bar{X})}] \quad (22)$$

$$R_p(\bar{X}, X') = -a\bar{X} + ke^{-b\bar{X}}\{E[(X - \bar{X})^2] - cE[(X - \bar{X})^3]\} \quad (23)$$

$$R_p(\bar{X}, X') = -a\bar{X} + e^{-b\bar{X}}\{dE^- [|X - \bar{X}|^{\theta_2}] - eE^+ [|X - \bar{X}|^{\theta_1}]\} \quad (24)$$

where $a, b, c, d, e, k, \theta_1$, and θ_2 are constants, $E^- [|X - \bar{X}|^{\theta_2}] = \sum_{x_i < \bar{X}} p_i |x_i - \bar{X}|^{\theta_2}$, $E^+ [|X - \bar{X}|^{\theta_1}] = \sum_{x_i > \bar{X}} p_i (x_i - \bar{X})^{\theta_1} = \sum_{x_i > \bar{X}} p_i (x_i - \bar{X})^{\theta_1}$, and p_i is the probability associated with the outcome x_i . When $b = 0$, these perceived-risk models become additive forms. For consistency with their corresponding risk-value models, we refer to equation (22) as the exponential risk model, equation (23) as the moments risk model, and equation (24) as the disappointment risk model. The latter was introduced by Bell [42] and explored in more detail by Jia *et al.* [43].

Similarly, some examples of the multiplicative form of risk model (20) are given as follows:

$$R_p(\bar{X}, X') = ke^{-b\bar{X}}E[e^{-c(X-\bar{X})} - 1] \quad (25)$$

$$R_p(\bar{X}, X') = ke^{-b\bar{X}}\{E[(X - \bar{X})^2] - cE[(X - \bar{X})^3]\} \quad (26)$$

$$R_p(\bar{X}, X') = e^{-b\bar{X}}\{dE^- [|X - \bar{X}|^{\theta_2}] - eE^+ [|X - \bar{X}|^{\theta_1}]\} \quad (27)$$

Research on financial risk and psychological risk (i.e., perceived risk) has been conducted separately in the past. The risk-value framework is able to provide a unified approach for dealing with both types of risk. The standard measure of risk is more normative in nature and should be useful in financial modeling. For instance, the standard measure of risk in perceived-risk models (24) and (27) includes many financial risk measures as special cases [2]. Our perceived-risk models show how financial measures of risk and psychological measures of risk can be related. In particular, for a given level of the mean value, minimizing the perceived risk will be equivalent to minimizing the standard risk since the expressions for $g(\bar{X})$ and $\psi(\bar{X})$ in equation (19) become constants. Our measures of perceived risk provide a clear way to simplify the decision criterion of minimizing perceived risk as suggested, but never operationalized, in Coombs' portfolio theory.

Conclusions

In this article, we have reviewed previous studies about perceived risk and focused on a two-attribute structure for perceived risk based on the mean of a lottery and its standard risk. Some of these risk measures also take into account the asymmetric effects of losses and gains on perceived risk. These measures of perceived risk can unify a large body of empirical evidence about risk judgments, and are consistent with the stylized facts regarding risk judgments listed in the introduction. For more details regarding the flexibility provided by the two-attribute structure for perceived risk see Jia *et al.* [38]; for details on the empirical validity of the assumptions behind the models, see Butler *et al.* [44].

In particular, these measures of perceived risk show a clear relationship between financial measures of risk and psychological measures of risk. They can also be incorporated into preference models in a natural way, on the basis of a trade-off between perceived risk and expected value. This shows an intuitively appealing connection between perceived risk and preference.

This development uses the expected value of a lottery as the reference point regarding the measures of perceived risk. The expected value is a convenient and probabilistically appealing reference point [2], which makes our risk models mathematically tractable and practically usable. There are other possible reference points that might be considered, such as an aspiration level, a reference lottery, or some other external reference point, such as zero. It would be interesting to consider these alternative reference points in our measures of perceived risk in future research.

Acknowledgment

This article summarizes a stream of research on perceived risk. In particular, we have incorporated materials that appeared previously in Jia J, Dyer JS, Butler JC. Measures of perceived risk. *Management Science* 1999 **45**: 519–532.

References

- [1] Jia, J. & Dyer, J.S. (1995). *Risk-Value Theory*, Working Paper, Graduate School of Business, University of Texas at Austin, Texas.
- [2] Jia, J. & Dyer, J.S. (1996). A standard measure of risk and risk-value models, *Management Science* **42**, 1961–1705.
- [3] Dyer, J.S., & Jia, J. (1998). Preference conditions for utility models: a risk-value perspective, *Annals of Operations Research* **80**, 167–182.
- [4] Pollatsek, A. & Tversky, A. (1970). A theory of risk, *Journal of Mathematical Psychology* **7**, 540–553.
- [5] Coombs, C.H. (1969). *Portfolio Theory: A Theory of Risky Decision Making*, Centre National de la Recherche Scientifique, La Decision.
- [6] Coombs, C.H. & Meyer, D.E. (1969). Risk preference in coin-toss games, *Journal of Mathematical Psychology* **6**, 514–527.
- [7] Coombs, C.H. & Huang, L. (1970). Tests of a portfolio theory of risk preference, *Journal of Experimental Psychology* **85**, 23–29.
- [8] Coombs, C.H. (1975). Portfolio theory and the measurement of risk, in *Human Judgment and Decision Processes*, M.F. Kaplan & S. Schwartz, eds, Academic Press, New York, pp. 63–85.
- [9] Keller, L.R., Sarin, R.K. & Weber, M. (1986). Empirical investigation of some properties of the perceived riskiness of gambles, *Organizational Behavior and Human Decision Process* **38**, 114–130.
- [10] Weber, E.U. & Bottom, W.P. (1990). An empirical evaluation of the transitivity, monotonicity, accounting, and conjoint axioms for perceived risk, *Organizational Behavior and Human Decision Process* **45**, 253–275.
- [11] Coombs, C.H. & Huang, L. (1970). Polynomial psychophysics of risk, *Journal of Mathematical Psychology* **7**, 317–338.
- [12] Coombs, C.H. & Lehner, P.E. (1981). An evaluation of two alternative models for a theory of risk: part 1, *Journal of Experimental Psychology, Human Perception and Performance* **7**, 1110–1123.
- [13] Brachinger, H.W. & Weber, M. (1997). Risk as a primitive: a survey of measures of perceived risk, *OR Spektrum* **19**, 235–250.
- [14] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1997). Thinking coherently, *Risk* **10**, 68–71.
- [15] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**, 203–228.
- [16] Acerbi, C. (2004). Coherent representations of subjective risk-aversion, in *Risk Measures for the 21st Century*, G. Szego, ed, John Wiley & Sons, pp. 147–206.
- [17] Coombs, C.H. & Pruitt, D.G. (1960). Components of risk in decision making: probability and variance preferences, *Journal of Experimental Psychology* **60**, 265–277.
- [18] Barron, F.H. (1976). Polynomial psychophysics of risk for selected business faculty, *Acta Psychologica* **40**, 127–137.
- [19] Coombs, C.H. & Lehner, P.E. (1984). Conjoint design and analysis of the bilinear model: an application to

- judgments of risk, *Journal of Mathematical Psychology* **28**, 1–42.
- [20] Kahneman, D.H. & Tversky, A. (1979). Prospect theory: an analysis of decision under risk, *Econometrica* **47**, 263–290.
- [21] Coombs, C.H. & Bowen, J.N. (1971). A test of VE-theories of risk and the effect of the central limit theorem, *Acta Psychologica* **35**, 15–28.
- [22] Fishburn, P.C. (1988). Foundations of risk measurement, *Encyclopedia of Statistical Sciences*, John Wiley & Sons, New York, pp. 148–152, Vol. 8.
- [23] Coombs, C.H. & Bowen, J.N. (1971). Additivity of risk in portfolios, *Perception & Psychophysics* **10**, 43–46.
- [24] Nygren, T. (1977). The relationship between the perceived risk and attractiveness of gambles: a multidimensional analysis, *Applied Psychological Measurement* **1**, 565–579.
- [25] Rotar, I.V. & Sholomitsky, A.G. (1994). On the Pollatsek-Tversky theorem on risk, *Journal of Mathematical Psychology* **38**, 322–334.
- [26] Luce, R.D. (1980). Several possible measures of risk, *Theory and Decision* **12**, 217–228; Correction, (1980), **13**, 381.
- [27] Sarin, R.K. (1987). Some extensions of Luce's measures of risk, *Theory and Decision* **22**, 25–141.
- [28] Huang, L.C. (1971). *The Expected Risk Function*, Michigan Mathematical Psychology Program Report 71-6, University of Michigan, Ann Arbor.
- [29] Luce, R.D. & Weber, E.U. (1986). An axiomatic theory of conjoint, expected risk, *Journal of Mathematical Psychology* **30**, 188–205.
- [30] Machina, M. (1982). Expected utility analysis without the independence axiom, *Econometrica* **50**, 277–323.
- [31] Weber, E.U. (1988). A descriptive measure of risk, *Acta Psychologica* **69**, 185–203.
- [32] Fishburn, P.C. (1984). Foundations of risk measurement, I, risk as probable loss, *Management Science* **30**, 296–406.
- [33] Fishburn, P.C. (1982). Foundations of risk measurement, II, effects of gains on risk, *Journal of Mathematical Psychology* **25**, 226–242.
- [34] Payne, W.J. (1973). Alternative approaches to decision making under risk: moments versus risk dimensions, *Psychological Bulletin* **80**, 439–453.
- [35] von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.
- [36] Fishburn, P.C. (1970). *Utility Theory for Decision Making*, John Wiley & Sons, New York.
- [37] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley & Sons, New York.
- [38] Jia, J., Dyer, J.S. & Butler, J.C. (1999). Measures of perceived risk, *Management Science* **45**, 519–532.
- [39] Bell, D.E. (1988). One-switch utility functions and a measure of risk, *Management Science* **34**, 1416–1424.
- [40] Bell, D.E. (1995). Risk, return, and utility, *Management Science* **41**, 23–30.
- [41] Sarin, R.K. & Weber, M. (1993). Risk-value models, *European Journal of Operational Research* **70**, 135–149.
- [42] Bell, D.E. (1985). Disappointment in decision making under uncertainty, *Operations Research* **33**, 1–27.
- [43] Jia, J., Dyer, J.S. & Butler, J.C. (2001). Generalized disappointment models, *Journal of Risk and Uncertainty* **22**, 59–78.
- [44] Butler, J., Dyer, J. & Jia, J. (2005). An empirical investigation of the assumptions of risk-value models, *Journal of Risk and Uncertainty* **30**, 133–156.

Related Articles

Operational Risk Modeling

Risk Attitude

Subjective Expected Utility

Utility Function

JIANMIN JIA, JAMES S. DYER AND
JOHN C. BUTLER

Basic Concepts of Insurance

Insurance is an economic institution that allows the transfer of financial risk from an individual to a pooled group of risks by means of a two-party contract. The insured party obtains a specified amount of coverage against an uncertain event (e.g., an earthquake or flood) for a smaller but certain payment (the premium). Insurers may offer fixed, specified coverage or replacement coverage, which takes into account the increased cost of putting the structure back to its original condition.^a

Most insurance policies have some form of deductible, which means that the insured party must cover the first portion of their loss. For example, a 10% deductible on a \$100 000 earthquake policy means that the insurer is responsible for property damage that exceeds \$10 000 up to some prespecified maximum amount, the coverage limit.

Losses and Claims

The insurance business, like any other business, has its own vocabulary. A *policyholder* is a person who has purchased insurance. The term *loss* is used to denote the payment that the insurer makes to the policyholder for the damage covered under the policy. It is also used to mean the aggregate of all payments in one event. Thus, we can say that there was a “loss” under the policy, meaning that the policyholder received a payment from the insurer. We may also say that the industry “lost” \$12.5 billion dollars in the Northridge earthquake.

A *claim* means that the policyholder is seeking to recover payments from the insurer for damage under the policy. A claim does not result in a *loss* if the amount of damage is below the deductible, or subject to a policy exclusion, but there still are expenses in investigating the claim. Even though there is a distinction between a claim and a loss, the terms are often used interchangeably to mean that an insured event occurred, or with reference to the prospect of having to pay out money.

The Law of Large Numbers

Insurance markets can exist because of *the law of large numbers* which states that for a series of independent and identically distributed random variables (such as automobile insurance claims), the variance of the average amount of a claim payment decreases as the number of claims increases. Consider the following gambling example. If you go to Las Vegas and place a bet on roulette, you are expected to lose a little more than 5 cents every time you bet \$1. But each time you bet, you either win or lose whole dollars. If you bet ten times, your average return is your net winnings and losses divided by ten. According to the law of large numbers, the average return converges to a loss of 5 cents per bet. The larger the number of bets, the closer the average loss per bet is to 5 cents.

Fire is an example of a risk that satisfies the law of large numbers since its losses are normally independent of one another.^b To illustrate this, suppose that an insurer wants to determine the accuracy of the fire loss for a group of identical homes valued at \$100 000, each of which has a 1/1000 annual chance of being completely destroyed by fire. If only one fire occurs in each home, the expected annual loss for each home would be \$100 (i.e., $1/1000 \times \$100\,000$).

If the insurer issued only a single policy, then a variance of approximately \$100 would be associated with its expected annual loss.^c

As the number of issued policies, n , increases the variance of the expected annual loss or mean decreases in proportion to n . Thus, if $n = 10$, the variance of the mean is approximately \$10. When $n = 100$ the variance decreases to \$1, and with $n = 1000$ the variance is \$0.10. It should thus be clear that it is not necessary to issue a large number of policies to reduce the variability of expected annual losses to a very small number if the risks are independent.

However, natural hazards – such as earthquakes, floods, hurricanes, and conflagrations such as the Oakland fire of 1991 – create problems for insurers because the risks affected by these events are not independent. They are thus classified as *catastrophic risks*. If a severe earthquake occurs in Los Angeles, there is a high probability that many structures would be damaged or destroyed at the same time. Therefore, the variance associated with an individual loss is actually the variance of all of the losses that occur from the specific disaster. Because of this high

variance, it takes an extraordinarily long history of past disasters to estimate the average loss with any degree of predictability. This is why seismologists and risk assessors would like to have databases of earthquakes, hurricanes, or other similar disasters over 100- to 500-year periods. With the relatively short period of recorded history, the average loss cannot be estimated with any reasonable degree of accuracy.

One way that insurers reduce the magnitude of their catastrophic losses is by employing high deductibles, where the policyholder pays a fixed amount of the loss (e.g., the first \$1000) or a percentage of the total coverage (e.g., the first 10% of a \$100 000 policy). The use of coinsurance, whereby the insurer pays a fraction of any loss that occurs, produces an effect similar to a deductible. Another way of limiting potential losses is for the insurer to place caps on the maximum amount of coverage on any given piece of property.

An additional option is for the insurer to buy reinsurance. For example, a company might purchase a reinsurance contract that covers any aggregate insured losses from a single disaster that exceeds \$50 million up to a maximum of \$100 million. Such an excess-of-loss contract could be translated as follows: the insurer would pay for the first \$50 million of losses, the reinsurer the next \$50 million, and the insurer the remaining amount if total insured losses exceeding \$100 million. An alternative contract would be for the insurer and reinsurer to share the loss above \$50 million, prorated according to some predetermined percentage.

End Notes

- ^a. This paper is based on material in Kunreuther [1].
- ^b. The Oakland fire of October 20, 1991 is a notable exception with 1941 single-unit dwellings totally damaged and 2069 partially destroyed with a total insured loss of \$1.7 billion. The October 1996 wild-fires that destroyed a number of homes in the Los Angeles area at a cost of many millions of dollars is another example of a nonindependent set of events.
- ^c. The variance for a single loss L with probability p is $Lp(1 - p)$. If $L = \$100\,000$ and $p = 1/1000$, then $Lp(1 - p) = \$100\,000 (1/1000)(999/1000)$, or \$99.90.

Reference

- [1] Kunreuther, H. (1998). Insurability conditions and the supply of coverage, in *Paying the Price: The Status and Role of Insurance Against Natural Disasters in the United States*, H. Kunreuther & R. Roth Sr, eds, Joseph Henry Press, Washington, DC, Chapter 2.

Related Articles

Insurance Pricing/Nonlife

Large Insurance Losses Distributions

Pricing of Life Insurance Liabilities

HOWARD KUNREUTHER

Bayes' Theorem and Updating of Belief

Basic Concepts

In statistics, we are concerned with the deductions that can be made from one or more observations about a hypothesis. This could involve a choice between a number of discrete alternatives, which includes hypothesis testing as a particular case, or choosing which value of an unknown quantity is best supported by the evidence, which is estimation theory. In *Bayesian statistics* (see **Bayesian Statistics in Quantitative Risk Assessment**), all such problems are dealt with by the use of Bayes' theorem.

Bayes' theorem is named after the Reverend Thomas Bayes (1702–1761), who studied how to compute a distribution for the parameter of a binomial distribution (to use modern terminology). His work, which makes use of a particular case of the theorem, was published posthumously in 1763 [1]. Pierre-Simon, Marquis de Laplace replicated and extended these results and was the first person to state the theorem in a way that is easily recognizable today (see [2]).

An example that illustrates the techniques involved comes from the consideration of a simplified version of the situation just before the time of the British national referendum as to whether the United Kingdom should remain part of the European Economic Community (EEC) that was held in 1975. Suppose that at that date, which was shortly after an election that the Labour Party had won, the proportion of the electorate supporting Labour (L) stood at 52%, while the proportion supporting the Conservatives (C) stood at 48% (it being assumed for simplicity that support for all other parties was negligible, although this was far from being the case). There were many opinion polls taken at the time, so we can take it as known that 55% of Labour supporters and 85% of Conservative voters intended to vote "Yes" (Y) and the remainder intended to vote "No" (N). Suppose that knowing all this you met someone at the time who said that she intended to vote "Yes", and you were interested in knowing which political party she supported. Using the notation $\Pr(A|B)$ to mean the probability of A , given B , we can summarize the information available by saying that

$$\Pr(L) = 0.52, \Pr(Y|L) = 0.55$$

$$\Pr(C) = 0.48, \Pr(Y|C) = 0.85 \quad (1)$$

(In view of present-day disenchantment with the EEC on many sides these figures may surprise many who were not around at the time, but they are in fact not wholly unrealistic.) What we want to do is to reverse the conditioning and find expressions for $\Pr(L|Y)$ and $\Pr(C|Y)$, and it is exactly this sort of problem that Bayes' theorem is designed to deal with.

The development of the theorem results from the definition of conditional probability. Given two events A and B , we define the intersection $A \cap B$ of them as happening when both of them occur, so in the above example $L \cap Y$ occurs when we find someone who is a Labour voter who also intends to vote "Yes" in the referendum. The formal definition of the conditional probability of A , given B , is then

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \quad (2)$$

In some circumstances the probability of an event can be approximated by the relative frequency with which it occurs in a large number of trials, so if $\#(A)$ denotes the number of occurrences of A among $\#(T)$ trials,

$$\Pr(A) = \text{"lim"} \#(A)/\#(T) \quad (3)$$

as $\#(T)$ tends to infinity. We can then see that

$$\begin{aligned} \frac{\Pr(A \cap B)}{\Pr(B)} &= \frac{\text{"lim"} \#(A \cap B)/\#(T)}{\text{"lim"} \#(B)/\#(T)} \\ &= \text{"lim"} \#(A \cap B)/\#(B) \end{aligned} \quad (4)$$

which is the relative frequency with which A occurs if we restrict ourselves solely to those occasions on which B occurs, and thus it seems reasonably interpretable as the probability of A , given B . It should be clearly understood that there is nothing controversial about Bayes' theorem as such. It is frequently used by probabilists and statisticians, whether or not they are Bayesians. The distinctive feature of Bayesian statistics is the application of the theorem in a wider range of circumstances than is usual in classical statistics. In particular, Bayesian statisticians are always willing to talk of the probability of a hypothesis, both unconditionally (its *prior probability*) and given some evidence (its *posterior probability*), whereas other statisticians will only talk of the probability of a hypothesis in restricted circumstances.

Next observe that if two events A and B are disjoint, so that they both cannot occur together, or symbolically their intersection is null, denoted $A \cap B = \emptyset$, then the probability that one or the other occurs is the sum of their separate probabilities, that is,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) \quad (5)$$

More generally, if there are a number of events A_i , which are disjoint, that is, such that no two of which can occur simultaneously, then the probability that one or the other of them occurs is the sum of their separate probabilities, that is,

$$\Pr(\cup A_i) = \sum \Pr(A_i) \quad (6)$$

Further if these events are exhaustive, so that it is necessarily the case that one of them occurs, then

$$1 = \Pr(\cup A_i) = \sum \Pr(A_i) \quad (7)$$

Now consider any one event E and a set of disjoint events H_i , which are such that, one or the other of them is bound to occur. As the events H_i are disjoint, so are the events $E \cap H_i$, and hence (since if E occurs one of the events $E \cap H_i$ must occur)

$$\begin{aligned} \Pr(E) &= \Pr(E \cap H_i) \\ &= \sum \Pr(E \cap H_i) \end{aligned} \quad (8)$$

Now using the definition of conditional probability we see that

$$\Pr(E) = \sum \Pr(E|H_i) \Pr(H_i) \quad (9)$$

This result is sometimes called the *generalized addition law* or the *law of the extension of the conversation*.

Bayes' Theorem

We can now come to Bayes' theorem. With E and H_i as above, we can use the definitions of $\Pr(H_i|E)$ and $\Pr(E|H_i)$ to see that

$$\begin{aligned} \Pr(H_i|E) \Pr(E) &= \Pr(E \cap H_i) \\ &= \Pr(H_i) \Pr(E|H_i) \end{aligned} \quad (10)$$

from which it follows that

$$\Pr(H_i|E) \propto \Pr(H_i) \Pr(E|H_i) \quad (11)$$

This is one form of *Bayes' theorem*. We note that it can be used to find the probability of a hypothesis,

given some evidence from the probability of the same hypothesis before the evidence was adduced, and the probability of the evidence given that hypothesis. In words it can be remembered as

Posterior probability $\propto$

$$\text{Prior probability} \times \text{Likelihood} \quad (12)$$

where the word *likelihood* is applied to the conditional probability of E given H_i considered as a function of H_i for a fixed E . The reason for using a different word is that, so considered, its properties are different in that, for example, it is not necessarily the case that if we add $\Pr(E|H_i)$ over an exclusive and exhaustive collection of events the sum is unity – in an extreme case the value of $\Pr(E|H_i)$ can even be the same for all i so that the sum might be less than one or even diverge to infinity.

If we go back to the original political example, we see that we can deduce that

$$\Pr(L|Y) \propto \Pr(Y|L) \Pr(L),$$

$$\Pr(C|Y) \propto \Pr(Y|C) \Pr(C) \quad (13)$$

so that

$$\Pr(L|Y) \propto (0.55)(0.52) = 0.286,$$

$$\Pr(C|Y) \propto (0.85)(0.48) = 0.408 \quad (14)$$

Since $\Pr(L|Y) + \Pr(C|Y) = 1$ as we have supposed all involved are either Labour or Conservative supporters, we can deduce that

$$\begin{aligned} \Pr(L|Y) &= \frac{0.286}{(0.286 + 0.408)} \\ &= 0.41 \end{aligned} \quad (15)$$

(Note that while it is sensible to retain additional decimal places in intermediate calculations, it is not good practice to quote the end result with more precision than the original data.)

It is sometimes useful to think of Bayes' theorem in a form which employs a sign of equality rather than one of proportionality as

$$\Pr(H_i|E) = \frac{\Pr(H_i) \Pr(E|H_i)}{\sum_j \Pr(H_j) \Pr(E|H_j)} \quad (16)$$

in which case the denominator

$$\sum_j \Pr(H_j) \Pr(E|H_j) \quad (17)$$

is referred to as the *normalizing factor*, but, as we saw above, it normally suffices to remember it in the form with a proportionality sign.

We defined likelihood above as the conditional probability of E , given H_i considered as a function of H_i for a fixed E . It is, however, worth noting that since we can always use Bayes' theorem with a proportionality sign, it makes no difference if we multiply the likelihood by any constant factor. For this reason, we can generalize the definition of likelihood by saying that it is any constant multiple of the said conditional probability.

Bayes' Theorem for Random Variables

There is also a version of Bayes' theorem for random variables. If we have two random variables X and Y with joint probability density function $f_{X,Y}(x, y)$, then we can define the conditional density of Y given X as

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} \quad (18)$$

where the marginal density $f_X(x)$ can of course be evaluated as

$$f_X(x) = \int f_{X,Y}(x, y) dy \quad (19)$$

Similarly we can define the conditional density of X , given Y . Very similar arguments to those used above give us *Bayes' theorem for random variables* in the form

$$f_{Y|X}(y|x) \propto f_Y(y) f_{X|Y}(x|y) \quad (20)$$

Just as in the discrete case we think of $f_Y(y)$ as the prior density for X while $f_{Y|X}(y|x)$ is the posterior density for Y , given X , and $f_{X|Y}(x|y)$ is the likelihood. As in the discrete case, we can redefine the likelihood to be any constant multiple of this or indeed any multiple that depends on x but not on y . In this case the normalizing factor takes the form

$$\begin{aligned} f_X(x) &= \int f_{X,Y}(x, y) dy \\ &= \int f_Y(y) f_{X|Y}(x|y) dy \end{aligned} \quad (21)$$

A somewhat artificial example of the use of this formula in the continuous case is as follows. Suppose Y is the time before the first occurrence of a radioactive decay, which is measured by an instrument, but that, because there is a delay built into the mechanism, the decay is recorded as having

taken place at a time $X > Y$. We actually observe that the value of X is, say, x , but would like to say what we can about the value of Y on the basis of this knowledge. We might, for example, have

$$f_Y(y) = \exp(-y) \quad (22)$$

$$f_{X|Y}(x|y) = \exp\{-k(x - y)\} \quad (23)$$

for $0 < y < \infty$ and $y < x < \infty$. Then the required conditional density is given by

$$f_{Y|X}(y|x) \propto f_Y(y) f_{X|Y}(x|y) \quad (24)$$

$$\propto \exp\{-(k - 1)y\} \quad (25)$$

for $0 < y < x$. Often we will find that it is enough to get a result up to a constant of proportionality, but if we need the constant it is very easy to find it because we know that the integral (or the sum in the discrete case) must be one. Thus in this case

$$f_{Y|X}(y|x) = \frac{(k - 1) \exp\{-(k - 1)y\}}{\exp\{-(k - 1)x\} - 1} \quad (26)$$

for $0 < y < x$.

The example above concerns continuous random variables, but the same results are obtained in the case of discrete random variables if we interpret the density $f_X(x)$ as the probability mass function, that is, the probability $\Pr(X = x)$ that X takes the value x and similarly for the other densities.

We also encounter cases where we have two random variables, one of which is continuous and the other is discrete. All the above definitions and formulae extend in an obvious way to such a case provided we are careful, for example, to use integration for continuous variables but summation for discrete variables. In particular, the formulation

$$f_{Y|X}(y|x) \propto f_Y(y) f_{X|Y}(x|y) \quad (27)$$

is still valid.

It may help to consider an example (again a somewhat artificial one). Suppose k is the number of successes in n Bernoulli trials, so k has a binomial distribution of index n and parameter p , but that the value of p is unknown, your beliefs about it being uniformly distributed over the interval $[0, 1]$ of possible values. Then for $k = 0, 1, \dots, n$

$$f_{K|P}(k|p) = C_k^n p^k (1 - p)^{n-k} \quad (28)$$

$$f_P(p) = 1 \quad (0 \leq p \leq 1) \quad (29)$$

so that

$$\begin{aligned} f_{P|K}(p|k) &\propto f_P(p) f_{K|P}(k|p) \\ &= C_k^n p^k (1-p)^{n-k} \end{aligned} \quad (30)$$

$$\propto p^k (1-p)^{n-k} \quad (31)$$

The constant can be found by integration if it is required. It turns out that, given k , p has a beta distribution $\text{Be}(k+1, n-k+1)$, and that the normalizing constant is the reciprocal of the beta function $B(k+1, n-k+1)$. Thus, this beta distribution should represent your beliefs about p after you have observed k successes in n trials. This example has a special importance in that it is the one that Bayes himself discussed in the article mentioned below.

Sequential Use of Bayes' Theorem

It should be noted that Bayes' theorem is often used sequentially, starting from a prior distribution and using some data to produce a posterior distribution, which takes this data into account and then using this distribution as a prior one for further analysis, which takes into account further data, and so on. As an illustration of this process, in the example about the probability of success in Bernoulli trials discussed above, we can take the posterior density proportional to $p^k(1-p)^{n-k}$ as a prior for use when further observations are available. So if j is the number of successes in further m trials, then using the old posterior as the new prior results in a posterior

$$\begin{aligned} f_{P|J,K}(p|j, k) &\propto f_{P|K}(p|k) f_{J|P}(j|p) \\ &\propto p^k (1-p)^{n-k} C_j^m p^j (1-p)^{m-j} \\ &\propto p^{j+k} (1-p)^{m+n-j-k} \end{aligned} \quad (32)$$

Note that the result is just the same as if we had taken the original, uniform, prior distribution $f_P(p) = 1$ and looked for a posterior resulting from the observation of $j+k$ successes in $m+n$ trials.

It could also happen that you have prior beliefs about the value of the unknown parameter p , which are equivalent to having observed k successes in n trials, although you may not, in fact, have observed such trials. If this is so, and you then observe j successes in m trials, then of course the end result is just as found above, since

$$\begin{aligned} f_{P|J}(p|j) &\propto f_P(p) f_{J|P}(j|p) \\ &= p^k (1-p)^{n-k} C_j^m p^j (1-p)^{m-j} \\ &\propto p^{j+k} (1-p)^{m+n-j-k} \end{aligned} \quad (33)$$

The details differ but the basic methodology is the same when other distributions are involved.

Fuller accounts of the topics discussed above can be found in [3–5] or [6] (listed in approximate order of sophistication and mathematical difficulty).

References

- [1] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society of London* **53**, 370–418.
- [2] Laplace, P.S. (1814). *Théorie Analytique des Probabilités*, 2nd Edition, Courcier, Paris.
- [3] Berry, D.A. (1996). *Statistics: A Bayesian Perspective*, Duxbury, Belmont.
- [4] Lee, P.M. (2004). *Bayesian Statistics: An Introduction*, 3rd Edition, Arnold, London.
- [5] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd Edition, Springer-Verlag, Berlin.
- [6] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*, John Wiley & Sons, New York.

Related Articles

Decision Analysis

Decision Trees

Expert Judgment

PETER M. LEE

Bayesian Analysis and Markov Chain Monte Carlo Simulation

Overview of Main Concepts

Bayesian analysis offers a way of dealing with information conceptually different from all other statistical methods. It provides a method by which observations are used to update estimates of the unknown parameters of a statistical model. With the Bayesian approach, we start with a parametric model that is adequate to describe the phenomenon we wish to analyze. Then, we assume a *prior distribution* for the unknown parameters θ of the model which represents our previous knowledge or belief about the phenomenon before observing any data. After observing some data assumed to be generated by our model we update these assumptions or beliefs. This is done by applying *Bayes' theorem* to obtain a *posterior probability density* for the unknown parameters given by

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta) d\theta} \quad (1)$$

where θ is the vector of unknown parameters governing our model, $p(\theta)$ is the *prior density* function of θ and x is a sample drawn from the “true” underlying distribution with *sampling density* $p(x|\theta)$ that we model. Thus the posterior distribution for θ takes into account both our prior distribution for θ and the observed data x .

A *conjugate prior family* is a class of densities $p(\theta_i)$ which has the feature that given the sampling density $p(x|\theta)$, the posterior density $p(\theta_i|x)$ also belongs to the same class. The name arises because we say that the prior $p(\theta_i)$ is *conjugate* to the sampling density considered as a *likelihood* function $p(x|\theta)$ for θ given x . The concept of conjugate prior as well as the term was introduced by Raiffa and Schlaifer [1].

After obtaining a posterior distribution for the parameters θ we can compute various quantities of interest such as integrals of the form

$$\int f(y)g(y;\theta)p(\theta|x) dy d\theta \quad (2)$$

where f is some arbitrary function and g is the probability density function describing a related parametric model. In general, because we are not assuming independence between each of the individual parameters, this integral is difficult to compute especially if there are many parameters. This is the situation in which *Markov chain Monte Carlo (MCMC)* (see **Reliability Demonstration; Imprecise Reliability**) simulation is most commonly used.

The distinguishing feature of MCMC is that the random samples of the integrand in equation (1) are *correlated*, whereas in conventional Monte Carlo methods such samples are *statistically independent*. The goal of MCMC methods is to construct an ergodic Markov chain (see **Repair, Inspection, and Replacement Models**) that converges quickly to its stationary distribution which is the required posterior density or some functional thereof such as equation (2).

One can broadly categorize the use of MCMC methods as Bayesian or non-Bayesian. *Non-Bayesian* MCMC methods are used to compute quantities that depend on a distribution from a statistical model that is *nonparametric*. In a *Bayesian* application, we consider a *parametric* model for the problem of interest. We assume some prior distribution on the parameters and try to compute quantities of interest that involve the posterior distributions. This approach remains suitable if the data is sparse, for example, in extreme value applications [2] (see **Statistics for Environmental Toxicity; Extreme Value Theory in Finance; Mathematics of Risk and Reliability: A Select History; Multiattribute Modeling**).

There are many different types of MCMC algorithms. The two most basic and widely used algorithms are the *Metropolis–Hastings algorithm* and the *Gibbs sampler* (see **Bayesian Statistics in Quantitative Risk Assessment**) which will be reviewed subsequently.

Metropolis–Hastings Algorithm

The Metropolis–Hastings algorithm [3–5] has been used extensively in physics but was little known to others until Müller [6] and Tierney [7] expounded its value to statisticians. This algorithm is extremely powerful and versatile and has been included in a list of “top 10 algorithms” [8] and even claimed to be most likely the most powerful algorithm of all time [9].

The Metropolis–Hastings algorithm can draw samples from any *target* probability density π for the uncertain parameters θ requiring only that this density can be calculated at θ . The algorithm makes use of a *proposal density* $q(\theta^t, \zeta)$, which depends on the current state of the chain θ^t to generate each new proposed parameter sample ζ . The proposal ζ is “accepted” as the next state of the chain ($\theta^{t+1} := \zeta$) with *acceptance probability* $\alpha(\theta^t, \zeta)$ and “rejected” otherwise. It is the specification of this probability α that allows us to generate a Markov chain with the desired target stationary density π . The Metropolis–Hastings algorithm can thus be seen as a generalized form of *acceptance/rejection sampling* with values drawn from approximate distributions, which are “corrected” in order that they behave asymptotically as random observations from the target distribution.

The algorithm in step-by-step form is as follows:

1. Given the current position of our Markov chain θ^t , generate a new value ζ from the proposal density q (see below).
2. Compute the *acceptance probability*

$$\alpha(\theta^t, \zeta) := \min \left(1, \frac{\pi(\zeta)q(\zeta, \theta^t)}{\pi(\theta^t)q(\theta^t, \zeta)} \right) \quad (3)$$

where π is the density of the target distribution.

3. With probability $\alpha(\theta^t, \zeta)$, set $\theta^{t+1} := \zeta$, else set $\theta^{t+1} := \theta^t$.
4. Return to step 1.

This algorithm generates a discrete time ergodic Markov chain $(\theta^t)_{t \geq 0}$ with stationary distribution Π corresponding to π , i.e., as $t \rightarrow \infty$

$$P(\theta^t \in B) \rightarrow \Pi(B) \quad (4)$$

for all suitably (Borel) measurable sets $B \in \mathbb{R}^n$.

Some important points to note are given in [5]:

- We need to specify a starting point θ^0 , which may be chosen at random (and often is). Preferably θ^0 should coincide with a mode of the density π .
- We should also specify a *burn-in period* to allow the chain to reach equilibrium. By this we mean that we discard the first n values of the chain in order to reduce the possibility of bias caused by the choice of the starting value θ^0 .
- The *proposal distribution* should be a distribution that is easy to sample from. It is also desirable to

choose its density q to be “close” or “similar” to the target density π , as this will increase the acceptance rate and increase the efficiency of the algorithm.

- We only need to know the *target density function* π up to proportionality – that is, we do not need to know its normalizing constant, since this cancels in the calculation (3) of the *acceptance function* α .

The choice of the burn-in period still remains somewhat of an art, but is currently an active area of research. One can simply use the “eyeballing technique” which merely involves inspecting visual outputs of the chain to see whether or not it has reached equilibrium.

When the proposal density is *symmetric*, i.e. $q(\theta^t, \zeta) = q(\zeta, \theta^t)$ (the original Metropolis algorithm), the computation of the acceptance function α is significantly faster. In this case from equation (3) a proposal ζ is accepted with probability $\alpha = \pi(\zeta)/\pi(\theta^t)$, i.e. its likelihood $\pi(\zeta)$ relative to that of $\pi(\theta^t)$ (as originally suggested by Ulam for acceptance/rejection sampling).

Random Walk Metropolis

If $q(\theta^t, \zeta) := f(|\theta - \zeta|)$ for some density f and norm $|\cdot|$, then this case is called a *random walk chain* (see **Statistical Arbitrage**) because the proposed states are drawn according to the process following $\zeta = \theta^t + \mathbf{v}$, where $\mathbf{v} \sim F$, the distribution corresponding to f . Note that, since this proposal density q is symmetric, the acceptance function is of the simple Metropolis form described above. Common choices for q are the multivariate normal, multivariate t or the uniform distribution on the unit sphere.

If $q(\theta^t, \zeta) := q(\zeta)$, then the candidate observation is drawn *independently* of the current state of the chain. Note, however, that the state of the chain θ^{t+1} at $t + 1$ does depend on the previous state θ^t because the acceptance function $\alpha(\theta^t, \cdot)$ depends on θ^t .

In the random walk chain we only need to specify the *spread* of q , i.e. a maximum for $|\theta - \zeta|$ at a single step. In the independence sampler we need to specify the spread and the location of q .

Choosing the spread of q is also something of an art. If the spread is large, then many of the candidates will be far from the current value. They

will, therefore, have a low probability of being accepted, and the chain may remain stuck at a particular value for many iterations. This can be especially problematic for multimodal distributions; some of its modes may then not be explored properly by the chain. On the other hand, if the spread is small the chain will take longer to traverse the support of the density and low probability regions will be under sampled. The research reported in [10] suggests an optimal acceptance rate of around 0.25 for the random walk chain. In the case of the independence sampler it is important [11] to ensure that the tails of q dominate those of π , otherwise the chain may get stuck in the tails of the target density. This requirement is similar to that in importance sampling.

Multiple-Block Updates

When the number of dimensions is large it can be difficult to choose the proposal density q so that the algorithm converges sufficiently rapidly. In such cases, it is helpful to break up the space into smaller blocks and to construct a Markov chain for each of these smaller blocks [4]. Suppose that we split θ into two blocks (θ_1, θ_2) and let $q_1(\theta_1^t | \theta_2^t, \zeta_1)$ and $q_2(\theta_2^t | \theta_1^t, \zeta_2)$ be the proposal densities for each block. We then break each iteration of the Metropolis–Hastings algorithm into two steps and at each step we update the corresponding block. To update block 1 we use the acceptance function given by

$$\alpha(\theta_1^t | \theta_2^t, \zeta_1) := \min \left(1, \frac{\pi(\zeta_1 | \theta_2^t) q_1(\zeta_1 | \theta_2^t, \theta_1^t)}{\pi(\theta_1^t | \theta_2^t) q_1(\theta_1^t | \theta_2^t, \zeta_1)} \right) \quad (5)$$

and to update block 2 we use

$$\alpha(\theta_2^t | \theta_1^t, \zeta_2) := \min \left(1, \frac{\pi(\zeta_2 | \theta_1^t) q_2(\zeta_2 | \theta_1^t, \theta_2^t)}{\pi(\theta_2^t | \theta_1^t) q_2(\theta_2^t | \theta_1^t, \zeta_2)} \right) \quad (6)$$

If each of the blocks consists of just a single variable, then the resulting algorithm is commonly called the *single-update* Metropolis–Hastings algorithm. Suppose in the single-update algorithm it turns out that each of the marginals of the target distribution $\pi(\theta_i | \theta_{\sim i})$ can be directly sampled from. Then we would naturally choose $q(\theta_i | \theta_{\sim i}) := \pi(\theta_i | \theta_{\sim i})$ since all candidates ζ will then be accepted with probability one. This special case uses the well-known *Gibbs sampler* method [11].

Gibbs Sampler

Gibbs sampling is applicable, in general, when the joint parameter distribution is not known explicitly but the conditional distribution of each parameter, given the others, is known. Let $P(\theta) = P(\theta_1, \dots, \theta_k)$ denote the joint parameter distribution and let $p(\theta_i | \theta_{\sim i})$ denote the conditional density for the i th component θ_i given the other $k - 1$ components, where $\theta_{\sim i} := \{\theta_j : j \neq i\}$ for $i = 1, \dots, k$. Although we do not know how to sample directly from P we do know how to sample directly from each $p(\theta_i | \theta_{\sim i})$. The algorithm begins by picking the arbitrary starting value $\theta^0 = (\theta_1^0, \dots, \theta_k^0)$. It then samples randomly from the conditional densities $p(\theta_i | \theta_{\sim i})$ for $i = 1, \dots, k$ successively as follows:

Sample θ_1^1 from $p(\theta_1 | \theta_2^0, \theta_3^0, \dots, \theta_k^0)$

Sample θ_2^1 from $p(\theta_2 | \theta_1^1, \theta_3^0, \dots, \theta_k^0)$

...

Sample θ_k^1 from $p(\theta_k | \theta_1^1, \theta_2^1, \dots, \theta_{k-1}^1)$

This completes a transition from θ^0 to θ^1 and eventually generates a sample path $\theta^0, \theta^1, \dots, \theta^t, \dots$ of a Markov chain whose stationary distribution is P .

In many cases we can use the Gibbs sampler, which is significantly faster to compute than the more general Metropolis–Hastings algorithm. In order to use Gibbs, however, we must know how to directly sample from the conditional posterior distributions for each parameter, i.e. $p(\theta_i | \theta_{\sim i}, x)$, where x represents the data to time t .

Use of MCMC in Capital Allocation for Operational Risk

Because of the lack of reported data on operational losses (see **Individual Risk Models; Extreme Value Theory in Finance; Simulation in Risk Management**) Bayesian MCMC simulation is well suited for the quantification of operational risk and operational risk capital allocation. In [12], a framework for the evaluation of *extreme* operational losses has been developed, which assumes that market and credit risks may be managed separately, but jointly imposes a *value at risk* (VaR) (see **Equity-Linked Life Insurance; Risk Measures and Economic Capital for (Re)insurers; Credit Scoring via Altman Z-Score**) limit u_{VaR} on these risks.

It is assumed that losses beyond the u_{VaR} level belong to the *operational risk* category. In most

cases, owing to the overlapping of risk types, a detailed analysis of operational loss data is required to support the assumption that the u_{VaR} level approximately equals the *unexpected loss threshold*. This approach to capital allocation for operational risk, which takes into account large but rare operational losses, is naturally based on *extreme value theory* (EVT) [13, 14] and focuses on tail events and modeling the worst-case losses as characterized by loss maxima over regular observation periods.

According to regulatory requirements [15], operational risk capital calculation requires two distributions – a *severity* distribution of loss values (see **Individual Risk Models; Insurance Pricing/Nonlife**) and a *frequency* distribution of loss occurrences. In the approach described here a unified resulting asymptotic model known as the *peaks over threshold* (POT) model [16–18] (see **Extreme Value Theory in Finance; Extreme Values in Reliability**) is applied. It is based on an asymptotic theory of extremes and a point process representation of exceedances over a threshold given to specify the POT model. The following assumption is made.

Given an *i.i.d.* sequence of random losses $X_1, \dots, X_n$ drawn from some distribution we are interested in the distribution of the *excess* $\mathbf{Y} := \mathbf{X} - u$ over the threshold u . The distribution of excesses is given by the conditional distribution function in terms of the tail of the underlying distribution function F as

$$F_u(y) := P(\mathbf{X} - u \leq y | \mathbf{X} > u) = \frac{F(u+y) - F(u)}{1 - F(u)} \quad \text{for } 0 \leq y \leq \infty \quad (7)$$

The limiting distribution $G_{\xi, \beta}(y)$ of excesses as $u \rightarrow \infty$ is known as the *generalized Pareto distribution* (GPD) with *shape* parameter ξ and *scale* parameter β given by

$$G_{\xi, \beta}(y) = \begin{cases} 1 - \left(1 + \xi \frac{y}{\beta}\right)^{-1/\xi} & \xi \neq 0 \\ \text{where } y \in [0, \xi] & \xi \geq 0 \\ \text{or } y \in [0, -\beta/\xi] & \xi < 0 \\ 1 - \exp\left(-\frac{y}{\beta}\right) & \xi = 0 \end{cases} \quad (8)$$

The identification of an appropriate threshold u is again somewhat of an art and requires a data analysis based on a knowledge of EVT [13, 14, 19].

The *capital provision* for operational risk over the unexpected loss threshold u is given in [2] as

$$\lambda_u E(\mathbf{X} - u | \mathbf{X} > u) = \lambda_u \frac{\beta_u + \xi u}{1 - \xi} \quad (9)$$

where $E(\mathbf{X} - u | \mathbf{X} > u) = \frac{\beta_u + \xi u}{1 - \xi}$ is the *expectation* of excesses over the threshold u (which is defined for $\xi \leq 1$ and must be replaced by the *median* for $\xi > 1$), $\beta_u := \sigma + \xi(u - \mu)$ and the exceedances form a Poisson point process with *intensity*

$$\lambda_u := \left(1 + \xi \frac{(u - \mu)}{\sigma}\right)^{-1/\xi} \quad (10)$$

usually measured in days per annum.

The accuracy of our model depends on accurate estimates of the ξ , μ , σ , and β parameters. To address this, hierarchical Bayesian MCMC simulation (see **Bayesian Statistics in Quantitative Risk Assessment**) is used to determine the parameter estimates of interest through intensive computation. The empirical estimation efficiency of this method when back-tested on large data sets is surprisingly good.

Hierarchical Bayesian parameter estimation considers the parameters to be random variables possessing a joint probability density function. The prior density $f_{\theta|\psi}$ of the random parameter vector θ is parametric with a vector of random *hyperparameters* ψ and is conjugate prior to the sampling density $f_{\mathbf{X}|\theta}$ so that the calculated posterior density $f_{\theta|\mathbf{X}_1, \dots, \mathbf{X}_n, \psi} := f_{\theta|\psi+}$ is of the same form with the new hyperparameters $\psi+$ determined by ψ and the observations $X_1, \dots, X_n$. In the hierarchical Bayesian model, the *hyper-hyper parameters* φ are chosen to generate a vague prior due to the lack of a prior distribution for the hyperparameters before excess loss data is seen. Hence, we can decompose the posterior parameter density $f_{\theta|\mathbf{X}, \psi}$ with the observations X and the initial hyper-hyper parameters φ as

$$\begin{aligned} f_{\theta|\mathbf{X}, \psi} &\propto f_{\mathbf{X}|\theta}(X|\theta) f_{\theta|\psi}(\theta|\psi) f_{\psi}(\psi|\varphi) \\ &\propto f_{\mathbf{X}|\theta}(X|\theta) f_{\psi|\theta}(\psi|\theta, \varphi) \\ &\propto f_{\mathbf{X}|\theta}(X|\theta) f_{\psi}(\psi|\varphi+) \end{aligned} \quad (11)$$

Here the Bayesian update of the prior parameter density $f_{\theta} \propto f_{\theta|\psi} f_{\psi}$ is performed in two stages. First by updating the hyper-hyper parameters φ to $\varphi+$ conditional on θ and then evaluating the

corresponding posterior density for this θ given the observations X .

Hierarchical Bayesian MCMC simulation for the parameters is based on the Metropolis–Hasting algorithm described briefly above and in detail in [20]. The idea is that the state of the chain for the parameter vector $\theta := \{\mu_j, \log \sigma_j, \xi_j : j = 1, 2, \dots, J\}$ converges to a stationary distribution which is the Bayesian posterior parameter distribution $f_{\theta|X, \psi}$ given the loss data x and a vector ψ of *hyperparameters* $\{m_\mu, s_\mu^2, m_{\log \sigma}, s_{\log \sigma}^2, m_\xi, s_\xi^2\}$. The hyperparameters are sampled from a conjugate prior *gamma-normal* (GM) distribution and are used to link the parameters $\{\mu_j, \sigma_j, \xi_j : j = 1, 2, \dots, J\}$ of each individual risk [2].

The aim of the model is to estimate the parameters of interest $\{\mu_j, \sigma_j, \xi_j : j = 1, 2, \dots, J\}$ conditional on both the data X and the hyperparameters $\{m_\mu, s_\mu^2, m_{\log \sigma}, s_{\log \sigma}^2, m_\xi, s_\xi^2\}$. The *posterior* distributions of the parameters are normally distributed:

$$\begin{aligned} \mu_j &\sim N(m_\mu, s_\mu^2), \log \sigma_j \sim N(m_{\log \sigma}, s_{\log \sigma}^2) \\ \text{and } \xi_j &\sim N(m_\xi, s_\xi^2) \end{aligned} \quad (12)$$

A schematic summary of the loss data, parameters, and hyperparameters is given in Table 1.

Illustrative Example

The data is assumed to represent the operational losses of a bank attributable to three different business units. The data starts on January 1, 1980 and ends on December 31, 1990. The time span is calculated in years, hence the parameters will also be measured on a yearly basis. The data has been generated from the Danish insurance claims data [19]

(see Figures 1 and 2) by two independent random multiplicative factors to obtain the three sets of loss data summarized in Table 2.

A typical analysis of such data includes time series plots, log histogram plots, sample mean excess plots, QQ plots for extreme value analysis against the GPD, Hill estimate plots of the shape parameter, and plots of the empirical distribution functions. All these tests have been performed for the three data sets to conclude that data are heavy tailed and that the POT model is valid.

Inputs for the MCMC model:

Threshold $u = 30$

Initial parameters:

μ_1	μ_2	μ_3	$\log \sigma_1$	$\log \sigma_2$	$\log \sigma_3$	ξ_1	ξ_2	ξ_3
20	21	22	3	3.2	2.8	0.5	0.4	0.7

The tables below (Tables 3–5) are a summary of the posterior mean estimates of the parameter values β_i and λ_i based on the MCMC posterior distribution mean parameter values.

The plots in Figure 3 below for the results of 2000 simulation loops show that convergence has been reached for the marginal posterior distributions of all parameters for Unit 1 and that the estimates of these parameters are distributed approximately normally. (Those of σ are thus approximately lognormal.) Similar results hold for the other two units.

The capital provision for operational losses is calculated using equation (9). The probability of such losses is given by the choice of threshold u for extreme operational losses. This threshold must be obtained from an analysis of the historical

Table 1 Bayesian hierarchical model

Data x	Type 1	Type 2				Type J	
Business Unit 1	x_{11}	x_{12}	.	.	.	x_{1J}	
Business Unit 2	x_{21}	x_{22}	.	.	.	x_{2J}	
	.	.	.	.	.	.	
	.	.	.	.	.	.	
Business Unit n	$x_{n,1}$	$x_{n,2}$	.	.	.	$x_{n,J}$	
<i>Parameters</i>	θ					<i>Hyperparameters ψ</i>	
Mean (μ)	μ_1	μ_2	.	.	.	μ_J	Mean – m_μ Variance – s_μ^2
Scale ($\log \sigma$)	$\log \sigma_1$	$\log \sigma_2$	.	.	.	$\log \sigma_J$	Mean – $m_{\log \sigma}$ Variance – $s_{\log \sigma}^2$
Shape (ξ)	ξ_1	ξ_2	.	.	.	ξ_J	Mean – m_ξ Variance – s_ξ^2

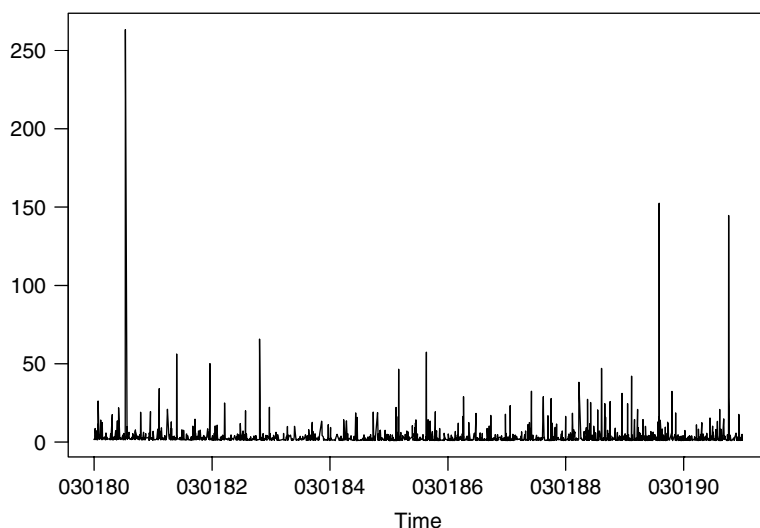


Figure 1 Time series of log “Danish” data X_1

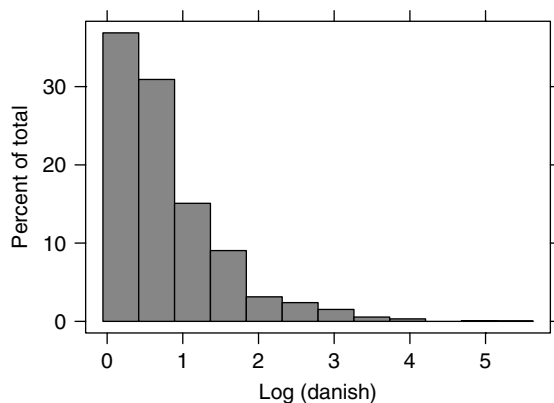


Figure 2 Histogram of log “Danish” data

Table 2 Summary statistics for data

Data	X_1 (Danish)	X_2	X_3
Minimum	1.000	0.8	1.200
First quarter	1.321	1.057	1.585
Mean	3.385	2.708	4.062
Median	1.778	1.423	2.134
Third quarter	2.967	2.374	3.560
Maximum	263.250	210.600	315.900
N	2167	2167	2167
Standard deviation	8.507	6.806	10.209

Table 3 Posterior parameter estimates for Unit 1

For $j = 1$ (Unit 1)

Code	Mean (μ_1)	Mean ($\log \sigma_1$)	Mean (ξ_1)	β_1	λ_1	Expected excess
1000 loops	37.34	3.14	0.77	18.46	1.41	180.70
2000 loops	36.89	3.13	0.80	18.35	1.39	211.75

The number of exceedances above the threshold is 15.

Table 4 Posterior parameter estimates for Unit 2

For $j = 2$ (Unit 2)

Code	Mean (μ_2)	Mean ($\log \sigma_2$)	Mean (ξ_2)	β_2	λ_2	Expected excess
1000 loops	36.41	3.16	0.77	19.04	1.34	186.22
2000 loops	35.76	3.13	0.8	18.5	1.30	218.40

The number of exceedances above the threshold is 11.

Table 5 Posterior parameter estimates for Unit 3

For $j = 3$ (Unit 3)

Code	Mean (μ_3)	Mean ($\log \sigma_3$)	Mean (ξ_3)	β_3	λ_3	Expected excess
1000 loops	39.55	3.05	0.79	14.21	1.71	180.52
2000 loops	39.23	3.03	0.82	13.83	1.70	213.50

The number of exceedances above the threshold is 24.

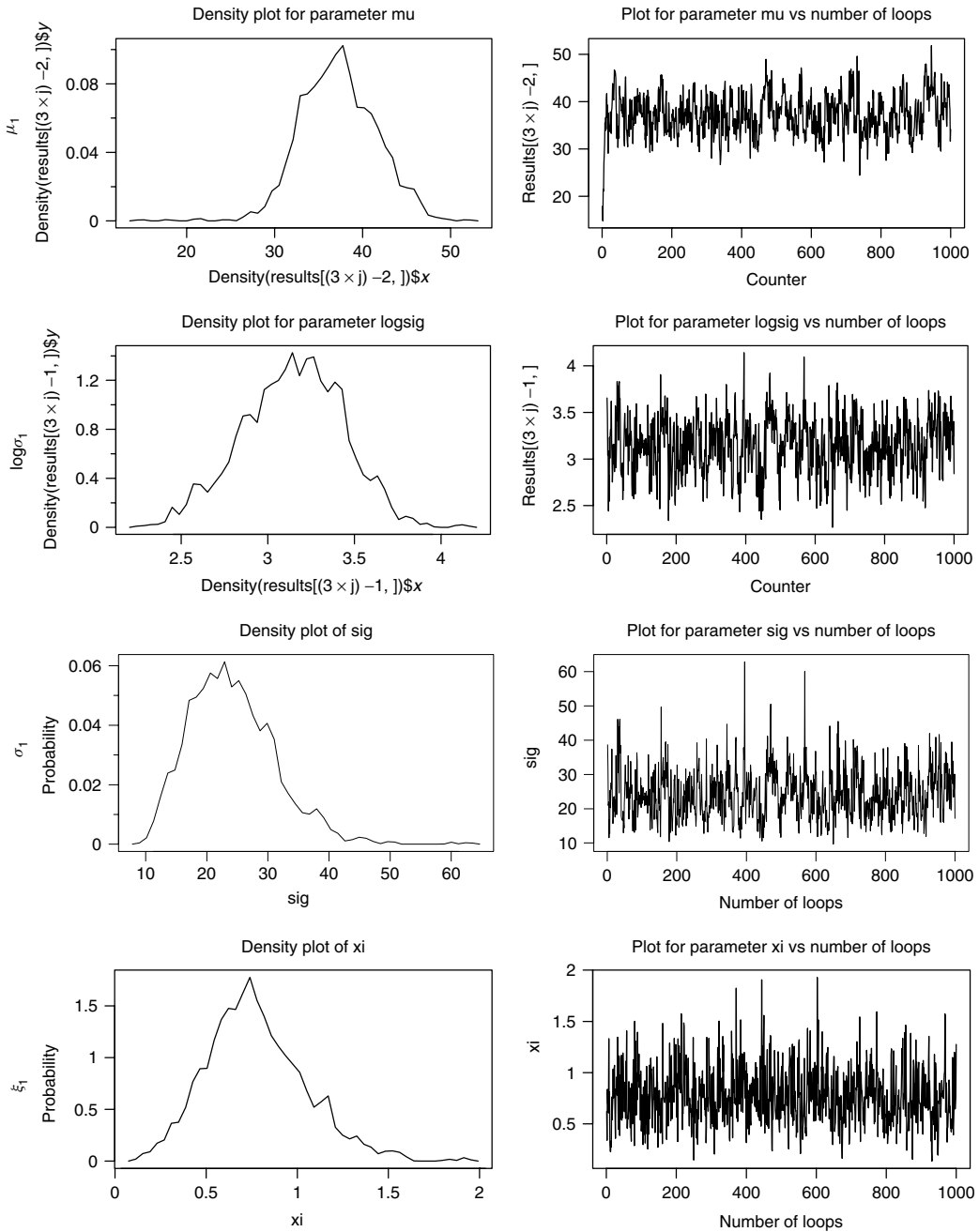


Figure 3 Simulation results of MCMC for 2000 iterations

operational loss data and should agree or exceed the threshold level u_{VaR} of unexpected losses due to market and credit risk. The probability π of crossing the combined market and credit risk threshold u_{VaR} is

chosen according to the usual *VaR* risk management procedures. The level of losses u due to operational risks is exceeded with probability $\rho \leq \pi$, so that $u \geq u_{\text{VaR}}$. The probability of exceeding u depends

on the shape of the tail of the loss distribution but, in general, is very much smaller than π .

Assuming that three types of losses are the bank business unit losses from operational risk over a period of 11 years the bank should hedge its operational risk for these units by putting aside 944.60 units of capital ($1.39 \times 211.75 + 1.30 \times 218.40 + 1.70 \times 213.50$) for any 1-year period (see Tables 3–5).

Although in this illustrative example unexpected losses above the combined VaR level (30 units of capital) occur with probability 2.5% per annum, unexpected operational risk losses will exceed this capital sum with probability less than 0.5%. In practice, lower tail probabilities might be chosen, but similar or higher probability ratios would be obtained.

Note that in this example the loss data for each business unit was generated as independent and the total capital figure takes the resulting diversification effect into account. On actual loss data, the dependencies in the realized data are taken into account by the method, and the diversification effect of the result can be analyzed by estimating each unit separately and adding the individual capital figures (which conservatively treats losses across units as perfectly correlated) [2]. Although the results presented here are based on very large (2167) original sample sizes, the simulation experiments on actual banking data reported in [2] verify the high accuracy of MCMC Bayesian hierarchical methods for exceedance sample sizes as low as 10 and 25, as in this example.

Conclusion

In this chapter, we have introduced MCMC concepts and techniques and showed how to apply them to the estimation of a Bayesian hierarchical model of interdependent extreme operational risks. This model employs the POT model of EVT to generate both *frequency* and *severity* statistics for the extreme operational losses of interdependent business units, which are of interest at the board level of a financial institution. These are obtained respectively in terms of Poisson exceedences of an unexpected loss level for other risks and the GPD.

The model leads to annual business unit capital allocations for unexpected extreme risks that take account of the statistical interdependencies of individual business unit losses.

The concepts discussed in this chapter are illustrated by an artificially created example involving

three business units but actual banking studies are described in [2] and in forthcoming work relating to internally collected operational loss data.

References

- [1] Raiffa, H. & Schlaifer, R. (1961). *Applied Statistical Decision Theory*, Harvard University Press.
- [2] Medova, E.A. & Kyriacou, M.N. (2002). Extremes in operational risk measurement, in *Risk Management: Value At Risk And Beyond*, M.A.H. Dempster, ed, Cambridge University Press, pp. 247–274.
- [3] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**(1), 1087–1092.
- [4] Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**(1), 97–109.
- [5] Chib, S. & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm, *The American Statistician* **49**(4), 327–335.
- [6] Müller, P. (1993). A generic approach to posterior integration and Gibbs sampling, Technical Report, Purdue University.
- [7] Tierney, L. (1994). Markov chains for exploring posterior distributions, *Annals of Statistics* **22**, 1701–1762.
- [8] Dongarra, J. & Sullivan, F. (2000). The top 10 algorithms, *Computing in Science and Engineering* **2**(1), 22–23.
- [9] Beichl, I. & Sullivan, F. (2000). The Metropolis algorithm, *Computing in Science and Engineering* **2**(1), 65–69.
- [10] Roberts, G., Gelman, A. & Gilks, W. (1994). Weak convergence and optimal scaling of random walk Metropolis algorithms, Technical Report, University of Cambridge.
- [11] Casella, G. & Edward, I.G. (1992). Explaining the Gibbs sampler, *The American Statistician* **46**, 167–117.
- [12] Medova, E.A. (2001). Operational risk capital allocation and integration of risks, *Advances in Operational Risk: Firmwide Issues for Financial Institutions*, Risk Books, pp. 115–127.
- [13] Embrechts, P., Kluppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events*, Springer, Berlin.
- [14] Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics*, John Wiley & Sons, New York.
- [15] Bank of International Settlements (2001). *The Bank of International Settlements, Basel, New Basel Capital Accord*, Press Release.
- [16] Leadbetter, M., Lindgreen, G. & Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer, Berlin.
- [17] Leadbetter, M. (1991). On a basis for “Peaks over Threshold” modeling, *Statistics and Probability Letters* **12**, 357–362.

- [18] Smith, R. (2001). Measuring risk with extreme value theory, Chapter 8, in *Risk Management: Value at Risk and Beyond*, M.A.H. Dempster, ed, Cambridge University Press.
- [19] Castillo, E. (1988). *Extreme Value Theory in Engineering*, Academic Press, Orlando.
- [20] Smith, A. & Roberts, G. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society* **B55**, 3–23.

ELENA MEDOVA

Bayesian Statistics in Quantitative Risk Assessment

Benefits of the Bayesian Approach in Risk Assessment

Bayesian inference has a number of advantages in modern data analysis and modeling, including applications in different forms of risk assessment. This is true in terms of the flexibility and applicability of Bayesian techniques as a result of the development of Markov chain Monte Carlo (MCMC) computational methods. By virtue of the relevance of Bayesian inferences and methods, it also holds to the types of data and problems tackled by modern scientific research, including those encountered in quantitative risk assessment where full expression of uncertainty surrounding assessment parameters or outcomes is desirable [1]. Useful introductory readings include those by Gelman *et al.* [2], Congdon [3], and Carlin and Louis [4], while more advanced theoretical approaches include Gamerman and Lopes [5], Bernardo and Smith [6], and Chen *et al.* [7].

Bayesian methods are being increasingly applied in risk assessment as they allow incorporation of subjective belief (*see Subjective Expected Utility*) or expert opinion in the form of prior probability distributions [8], as well as probabilistic approaches rather than deterministic exposure assessments [9].

Various forms of risk assessment have been tackled from a Bayesian perspective, and examples of Bayesian approaches can be found in journals such as *Environmental Toxicology*, *Journal of Exposure Science and Environmental Epidemiology*, *Human and Ecological Risk Assessment*, and *Stochastic Environmental Research and Risk Assessment*. Environmental risk assessment (*see Environmental Health Risk Assessment*) typically involves evaluating the potential impact of an exogenous hazard (e.g., air pollution or soil contamination, food poisoning, flood, or other natural disaster) on human populations or ecological systems, or one associated with a possible planned intervention (e.g., assessing the ecological impact of major airport or road schemes). Examples of Bayesian perspectives in such areas are included in [10–13].

Another type of assessment involves exceedance analysis, e.g., obtaining the probability that a 100-year storm will take place on any given year or the return period of a certain size event. Flood exceedance probabilities, calculated from peak annual stream or river discharge measurements, are important in determining the risk associated with the construction of buildings on a flood plain and also determining the size of dam structures [14, 15]. Finally, one may mention risk assessment associated with industrial and engineering applications where the issue is the risk of accidents, component wear out, or system failures [16] (*see Systems Reliability*).

Frameworks for risk assessment may include identifying hazards, quantifying risks or exposures, and evaluating adverse health effects under varying or critical exposure levels (including dose–response modeling; *see Dose–Response Analysis*). Risk assessment may involve a comparison of the potential health impacts of various risk-management interventions (to be included in cost-effectiveness estimates) or a prioritization between different risks.

As mentioned earlier, it is important that risk assessment recognizes the uncertainties in any evaluation, costing or prioritization, and a Bayesian approach is adapted to this, especially in more complex hierarchical models. For instance, assessment of a disease–exposure relationship may be complicated when exposure measurements are subject to measurement error – such that a full model should include a component relating unknown “true” exposure to the observed exposure. Bayesian analysis allows for full expression of parameter uncertainty propagated across linked models – see Draper [17], with Manson [18] providing a flood risk application. Hence, interval estimates fully allowing for uncertainty may be wider than under frequentist approaches to the same problem.

Much of the data from ecological, environmental, and health research has a complex structure, involving hierarchical clustering of subjects, crossed classifications of subjects, or repeated measures on subjects. Furthermore, exposures to pollutants or other forms of risk are typically unevenly distributed both temporally and geographically [19]. The Bayesian approach naturally adapts itself to hierarchically or spatiotemporally correlated effects *via* conditionally specified priors, with the first stage specifying the likelihood of the data given unknown randomly distributed cluster effects, the second stage specifying

the density of the population of cluster effects, and the third stage providing priors on the population parameters. Examples are provided by Friesen *et al.* [20] who consider random-effects modeling of job exposures in historical epidemiological studies, and Symanski *et al.* [21] who consider clustering of exposures by worker, job groups, building, and plant in a random-effects model (see **Occupational Risks; Occupational Cohort Studies**).

A further distinctive feature of much quantitative risk assessment is the availability of substantial accumulated knowledge about important unknowns. In health risk assessments, the primary concern is often the risk of various cancers, though reproductive and developmental outcomes also figure importantly; there is often clinical knowledge about exposure–disease mechanisms for such outcomes and such knowledge, possibly discounted, can be summarized in prior distributions that can be used in risk analysis.

In the next section, we review some potential benefits of a Bayesian approach in risk assessment and other applications. This is followed by a review of the principle of updating from prior to posterior and of considerations in selecting priors. Some of the issues involved in MCMC sampling are then discussed, such as alternative algorithms and convergence, followed by a discussion of some of the most common ways of measuring model fit and checking model adequacy. The chapter concludes with illustrative applications using the WINBUGS program.

Conceptual and Computational Advantages

Bayesian inference arguably has a number of more general conceptual advantages. However, in the past, Bayesian analysis was impeded by the complex numerical integrations needed. Recently developed computer intensive sampling methods of estimation have revolutionized the application of Bayesian methods and enhanced its potential in terms of inferences.

The Bayesian learning process involves modifying or updating prior probability statements about the parameters, held before observing the data, to updated or posterior knowledge that combines both prior knowledge and the data at hand. This provides a natural way of learning from data so that beliefs or evidence can be cumulatively updated – this is

not possible in the same way in classical analysis. While pre-MCMC Bayesian analysis often relied on conjugate analysis where prior and posterior densities for θ have the same form [22], the advent of MCMC has considerably widened the potential for realistic statements about prior evidence. For example, accumulated evidence may suggest that environmental risks or residue concentrations are positively skewed rather than normally distributed.

The Bayes' method provides interval estimates on parameters that are more in line with commonsense interpretations: a Bayesian 95% credible interval containing the true parameter with 95% probability, whereas classical interpretation of the 95% confidence interval is as the range of values containing the true parameter in 95% of repeated samples.

Using modern sampling methods, Bayesian methods can be applied to complex random-effects models that are difficult to fit using classical methods, and allow inferences without assumptions of asymptotic normality that underlie classical estimation methods. MCMC methods provide a full density profile of a parameter so that any clear nonnormality is apparent, and permit a range of hypotheses about the parameters to be assessed using the sample information.

Prior to Posterior Updating

Bayesian models are typically concerned with inferences on a parameter set $\theta = (\theta_1, \dots, \theta_d)$ of dimension d , which may include “fixed effect” parameters (such as regression coefficients in a normal errors linear regression), random effects as in multilevel or spatial models, hierarchical parameters, unobserved states (in time series models), or even missing observations [23]. For example in a normal linear regression with p predictors, the unknowns would have the dimension $p + 2$, namely, a constant β_0 , regression coefficients for the predictors $\beta_1, \dots, \beta_p$, and the variance term σ^2 .

Expressing prior subject matter knowledge about parameters is an important aspect of the Bayesian modeling process. Such knowledge is summarized in a set of densities $\{p(\theta_1), p(\theta_2), \dots, p(\theta_d)\}$, which are collectively denoted as $p(\theta)$. The likelihood of the observations may be denoted as $p(y|\theta)$. The updated knowledge about parameters conditional on the prior and the data is then contained in a posterior density $p(\theta|y)$. From Bayes' theorem (see **Bayes' Theorem**

and Updating of Belief), one has

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \quad (1)$$

where the denominator on the right side is known as the *marginal likelihood* $p(y)$, and is a normalizing constant to ensure that $p(\theta|y)$ is a proper density (integrates to 1). So, one can express the Bayes' inference process as

$$p(\theta|y) \propto p(y|\theta)p(\theta) \quad (2)$$

since posterior knowledge is a function of the data and prior beliefs. The relative impact of the prior and data on updated beliefs depends on how “informative” the prior is – how precise the density $p(\theta)$ is – and the sample size of the data. A large sample would tend to outweigh even an informative prior. Note that certain priors (e.g., that a regression parameter only take positive values) do have an enduring impact on posterior inferences.

In hierarchical models involving latent data Z (e.g., random effects, missing observations), the joint density has the form $p(y, \theta, Z) = p(y|\theta, Z)p(\theta, Z) = p(y|\theta, Z)p(Z|\theta)p(\theta)$. The posterior density for both Z and θ is

$$p(Z, \theta|y) = \frac{p(y|\theta, Z)p(Z|\theta)p(\theta)}{p(y)} \quad (3)$$

Choice of Prior Density

Choice of an appropriate prior density, whether in mathematical terms or in terms of being subjectively justified by the application is an important issue in the Bayesian approach – which tends to militate against menu driven “off the shelf” application. There are also mathematical reasons to prefer some sorts of prior to others, though the advent of MCMC methods and a range of rejection sampling methods has reduced the need for restricting models to conjugate priors. Thus a β density for a binomial probability is conjugate with a binomial likelihood since the posterior has the same (β) density form as the prior. There may be questions of sensitivity to the choice of prior especially in certain forms of model – examples are the priors used in hierarchical models and in discrete mixture models.

It may be possible to base the prior density for θ on cumulative evidence *via* meta-analysis of existing studies (*see Meta-Analysis in Nonclinical Risk*

Assessment), or *via* elicitation techniques aimed at developing informative priors. This is well established in assessment of engineering risk and reliability, where systematic elicitation approaches such as maximum-entropy priors are used [24, 25]. Simple approximations include the histogram method, which divides the domain of θ into a set of bins and elicits prior probabilities that θ is located in each bin; then $p(\theta)$ may be represented as a discrete prior or converted to a smooth density. Prior elicitation may be aided if a prior is reparameterized in the form of a mean and variance; for example, β priors $\text{Beta}(a, b)$ for probabilities can be expressed as $\text{Beta}(m\tau, (1-m)\tau)$ where m is an estimate of the mean probability and τ is the estimated precision (degree of confidence) of that prior mean.

If a set of existing studies is available providing evidence on the likely density of a parameter, these may be used in the form of preliminary meta-analysis to set up an informative prior for the current study. However, there may be limits to the applicability of existing studies to the current data, and so pooled information from previous studies may be downweighted. For example, the precision of the pooled estimate from previous studies may be scaled downwards, with the scale possibly an extra unknown. When a maximum-likelihood analysis is simple to apply, one option is to downweight the variance–covariance matrix of the maximum likelihood estimate (MLE) [26]. More comprehensive ways of downweighting historical/prior evidence, such as power prior models, have been proposed [27, 28].

Especially in hierarchical models, the form of the second stage prior $p(Z|\theta)$ amounts to a hypothesis about the nature of the random effects. Thus a hierarchical model for small area mortality models may include spatially structured random effects, exchangeable random effects with no spatial pattern, or both, as under the convolution prior of Besag *et al.* [29]. A prior specifying the errors as spatially correlated is likely to be a working model assumption, rather than a true cumulation of knowledge, and one may have several models being compared, with different forms assumed for $p(Z|\theta)$.

In many situations, existing knowledge may be difficult to summarize or elicit in the form of an “informative prior” and to express prior ignorance one may use “default” or “noninformative” priors, and this is generally not problematic for fixed effects (such as linear regression coefficients). Since the

classical maximum-likelihood estimate is obtained without considering priors on the parameters, a possible heuristic is that a noninformative prior leads to a Bayesian posterior estimate close to the maximum-likelihood estimate. It might appear that a maximum-likelihood analysis would, therefore, necessarily be approximated by flat or improper priors, such that a parameter is uniformly distributed between $-\infty$ and $+\infty$, or between 0 and $+\infty$ for a positive parameter. However, such priors may actually be unexpectedly informative about different parameter values [30, 31]. An intermediate option might be a diffuse but proper prior such as a uniform prior with a large but known range or a normal prior with very large variance.

For variance parameters, choice of a noninformative prior is more problematic as improper priors may induce improper posteriors that prevent MCMC convergence, as in a normal hierarchical model with subjects j nested in clusters i ,

$$y_{ij} \sim N(\theta_i, \sigma^2) \quad (4)$$

$$\theta_i \sim N(\mu, \tau^2) \quad (5)$$

The prior $p(\mu, \tau) = 1/\tau$ results in a improper posterior [32]. Flat priors on any particular scale will not be flat on another scale (examples include flat priors on σ (confined to positive values), σ^2 , $\log(\sigma)$, or $\log(\sigma^2)$). Just proper priors (e.g., a γ on $1/\sigma^2$ with small scale and shape parameters) do not necessarily avoid these problems and in fact may favor particular values despite being supposedly only weakly informative. Such priors may cause identifiability problems and impede MCMC convergence [33]. Choice of suitable priors for variances in hierarchical models is an active research area [34–36].

Analytically based rules for deriving noninformative priors include reference priors [37] and Jeffreys' prior

$$p(\theta) \propto |I(\theta)|^{0.5} \quad (6)$$

where $I(\theta)$ is the information matrix, namely, $I(\theta) = -E \left\{ \frac{\partial^2 \ell(\theta)}{\partial \ell(\theta_i) \partial \ell(\theta_j)} \right\}$ where $l(\theta) = \log(L(\theta|y))$ is the log-likelihood. Unlike uniform priors, Jeffreys prior is invariant under transformation of scale.

To assess sensitivity to prior assumptions, the analysis may be repeated for a limited range of alternative priors [38], possibly following the principle of Spiegelhalter *et al.* [39] in providing a range of viewpoints; for example, a prior on a treatment effect in

a clinical trial might be neutral, sceptical, or enthusiastic. Formal approaches to prior robustness may be based on mixture or “contamination” priors. For instance, one might assume a two group mixture with larger probability $1 - q$ on the “main” prior $p_1(\theta)$, and a smaller probability such as $q = 0.2$ on a contaminating density $p_2(\theta)$, which may be any density [40]. One might consider the contaminating prior to be a flat reference prior, or one allowing for shifts in the main prior's assumed parameter values [41, 42].

MCMC Sampling and Inferences from Posterior Densities

Bayesian inference has become intimately linked to sampling-based estimation methods that focus on obtaining the entire posterior density of a parameter. A variety of adaptations of Monte Carlo methods have been proposed to sample from posterior densities. Standard Monte Carlo methods generate independent simulations $u^{(1)}, u^{(2)}, \dots, u^{(T)}$ from a target density $\pi(u)$, so that $E_\pi[g(u)] = \int g(u)\pi(u) du$ is estimated as

$$\bar{g} = \frac{1}{T} \sum_{t=1}^T g(u^{(t)}) \quad (7)$$

and $\bar{g}$ tends to $E_\pi[g(u)]$ as $T \rightarrow \infty$. By contrast, independent sampling from the posterior density $\pi(\theta) = p(\theta|y)$ is not usually feasible but dependent samples $\theta^{(t)}$ can be used provided the sampling satisfactorily covers the support of $p(\theta|y)$ [43]. So MCMC methods generate dependent draws *via* Markov chains defined by the assumption

$$p(\theta^{(t)}|\theta^{(1)}, \dots, \theta^{(t-1)}) = p(\theta^{(t)}|\theta^{(t-1)}) \quad (8)$$

so that only the preceding state is relevant to the future state. Sampling from such a Markov chain converges to the stationary distribution required, $\pi(\theta)$, if additional requirements (irreducibility, aperiodicity, and positive recurrence) on the chain are satisfied.

There is no limit to the number of samples T of θ that may be taken from a posterior density $p(\theta|y)$. Such sampling generates estimates of density characteristics (moments, quantiles, 95% credible intervals), and can be used to provide probabilities on hypotheses relating to the parameters [44]. For example, the 95% credible interval may be estimated using the 0.025 and 0.975 quantiles of the sampled

output $\{\theta_k^{(t)}, t = B + 1, \dots, T\}$, where B is a prior sequence of iterations to ensure convergence.

Monte Carlo posterior summaries typically include posterior means and variances of the parameters, obtainable as moment estimates from the MCMC output. This is equivalent to estimating the integrals

$$E(\theta_k|y) = \int \theta_k p(\theta|y) d\theta \quad (9)$$

$$\begin{aligned} \text{Var}(\theta_k|y) &= \int \theta_k^2 p(\theta|y) d\theta - [E(\theta_k|y)]^2 \\ &= E(\theta_k^2|y) - [E(\theta_k|y)]^2 \end{aligned} \quad (10)$$

One may also use the MCMC output to obtain posterior means, variances, and credible intervals for functions $\Delta = \Delta(\theta)$ of the parameters [45]. These are estimates of the integrals

$$E[\Delta(\theta)|y] = \int \Delta(\theta) p(\theta|y) d\theta \quad (11)$$

$$\begin{aligned} \text{Var}[\Delta(\theta)|y] &= \int \Delta^2 p(\theta|y) d\theta - [E(\Delta|y)]^2 \\ &= E(\Delta^2|y) - [E(\Delta|y)]^2 \end{aligned} \quad (12)$$

For $\Delta(\theta)$ its posterior mean is obtained by calculating $\Delta^{(t)}$ at every MCMC iteration from the sampled values $\theta^{(t)}$. The theoretical justification for such estimates is provided by the MCMC version of the law of large numbers [46], namely, that

$$\sum_{t=1}^T \Delta(\theta^{(t)})/T \longrightarrow E_{\pi}[\Delta(\theta)] \quad (13)$$

provided that the expectation of $\Delta(\theta)$ under $\pi(\theta) = p(\theta|y)$, denoted $E_{\pi}[\Delta(\theta)]$, exists.

Posterior probability estimates may also be made using the samples from an MCMC run. These might relate to the probability that θ_k exceeds a threshold b , and provide an estimate of the integral $\Pr(\theta_k > b|y) = \int_b^{\infty} p(\theta_k|y) d\theta_k$. This would be estimated by the proportion of iterations where $\theta_k^{(t)}$ exceeded b , namely,

$$\hat{Pr}(\theta_k > b|y) = \sum_{t=1}^T 1(\theta_k^{(t)} > b)/T \quad (14)$$

where $1(A)$ is an indicator function that takes the value 1 when A is true, or 0 otherwise. Thus in a disease mapping application, one might wish to

obtain the probability that an area's smoothed relative mortality risk θ_k exceeds one, and so count iterations where this condition holds.

Summarizing Inferences from a Bayesian Analysis

Posterior summaries of parameters or related theoretical quantities typically include posterior means and variances of individual parameters, as well as selected percentiles, e.g., the 5th and 95th or the 2.5th and 97.5th. It is useful to present the posterior median as well as the posterior mean to provide at least an informal check on possible asymmetry in the posterior density of a parameter.

From the percentile tail points, the associated equal-tail credible interval for a parameter can be obtained (e.g., the 90% credible interval is obtained if the summary includes the 5th and 95th percentiles from the MCMC sampling output). Another form of credible interval is the $100(1 - \alpha)\%$ highest probability density (HPD) interval, defined such that the density for every point inside the interval exceeds that for every point outside the interval, and also defined to be the shortest possible $100(1 - \alpha)\%$ credible interval.

An overall depiction of the marginal posterior density of an individual parameter may be provided by a histogram or kernel smoothing density estimate [47], based on the full set of posterior samples for individual parameters from the MCMC run subsequent to convergence. A bivariate kernel density may be estimated to summarize the joint posterior densities of two parameters. Such plots will demonstrate whether posterior normality is present or instead features such as skewness or multiple modes are present. As an example of a clearly nonnormal posterior, Gelfand *et al.* [48] present a kernel density estimate for a variance ratio.

Summary presentations or graphical checks on collections of parameters (e.g., a set of random effects) may also be relevant – for example, school effects in a multilevel analysis of exam results. One may use relevant plots (e.g., $Q-Q$ or histogram plots) to assess assumptions of normality typically made in the priors about such effects; this would involve plotting the posterior means or medians of all the effects [49].

The remaining practical questions include establishing an MCMC sampling scheme (see section titled “MCMC Sampling Algorithms”) and that convergence to a steady state has been obtained for practical

purposes (see section titled “Assessing MCMC Convergence”). The sampling output can also be used to provide model fit and model checking criteria (see section titled “Model Fit and Predictions from the Model”). For example, one may sample predictions from the model in the form of replicate data $y_{\text{new}}^{(t)}$ and using these one can check whether a model’s predictions are consistent with the observed data. Predictive replicates are obtained by sampling $\theta^{(t)}$ and then sampling y_{new} from the likelihood model $p(y_{\text{new}}|\theta^{(t)})$.

MCMC Sampling Algorithms

The Metropolis–Hastings algorithm (M–H algorithm) is the prototype for MCMC schemes that simulate a Markov chain $\theta^{(t)}$ for a d -dimensional parameter θ , with $p(\theta|y)$ as the stationary distribution. Following Hastings [50], candidate values θ^* are sampled only according to the improvement they make in $p(\theta|y)$, and if they satisfy a comparison (the Hastings correction) involving a candidate generating density $g(\theta^*|\theta^{(t)})$. This density should ideally be a close approximation for $p(\theta|y)$ but as in the selection of an importance distribution in importance sampling, g should have fatter tails than $p(\theta|y)$ for the algorithm to be efficient.

The M–H algorithm produces a Markov chain that is reversible with respect to $p(\theta|y)$, and so has $p(\theta|y)$ as its stationary distribution [51]. Specifically, the chain is updated from the current value $\theta^{(t)}$ to a potential new value θ^* with probability

$$\alpha(\theta^*|\theta^{(t)}) = \min \left(1, \frac{p(\theta^*|y)g(\theta^{(t)}|\theta^*)}{p(\theta^{(t)}|y)g(\theta^*|\theta^{(t)})} \right) \quad (15)$$

where $g(\theta^*|\theta^{(t)})$ is the ordinate of g for the candidate value, and $g(\theta^{(t)}|\theta^*)$ is the density associated with moving back from θ^* to the current value. If the proposed new value θ^* is accepted, then $\theta^{(t+1)} = \theta^*$; but if it is rejected, the next state is the same as the current state, i.e., $\theta^{(t+1)} = \theta^{(t)}$. An equivalent way of stating the mechanism is

$$\begin{aligned} \theta^{(t+1)} &= \theta^* & \text{if } u \leq \frac{p(\theta^*|y)g(\theta^{(t)}|\theta^*)}{p(\theta^{(t)}|y)g(\theta^*|\theta^{(t)})} \\ \theta^{(t+1)} &= \theta^{(t)} & \text{otherwise} \end{aligned} \quad (16)$$

where u is a draw from a $U(0,1)$ density [52].

The M–H scheme, therefore, has a transition kernel that allows for the possibility of not moving, namely,

$$\begin{aligned} K(\theta^{(t)}, \theta^{(t+1)}) &= g(\theta^{(t+1)}|\theta^{(t)}) \alpha(\theta^{(t+1)}|\theta^{(t)}) + 1(\theta^{(t+1)} = \theta^{(t)}) \\ &\quad \times \int \left[1 - \int \alpha(u|\theta^{(t)}) \right] \times g(u|\theta^{(t)}) du \end{aligned} \quad (17)$$

The target density $p(\theta|y)$ appears in the form of a ratio so it is not necessary to know any normalizing constants, and from equation (2), one can use the scheme

$$\alpha(\theta^*|\theta^{(t)}) = \min \left(1, \frac{p(y|\theta^*)p(\theta^*)g(\theta^{(t)}|\theta^*)}{p(y|\theta^{(t)})p(\theta^{(t)})g(\theta^*|\theta^{(t)})} \right) \quad (18)$$

involving likelihood values $p(y|\theta)$ and ordinates of prior densities $p(\theta)$.

For symmetric proposal densities, with $g(\theta^*|\theta^{(t)}) = g(\theta^{(t)}|\theta^*)$, such as a normal centered at $\theta^{(t)}$, one obtains the simpler Metropolis update scheme [53], whereby

$$\alpha(\theta^*|\theta^{(t)}) = \min \left[1, \frac{p(\theta^*|y)}{p(\theta^{(t)}|y)} \right] \quad (19)$$

Another option is independence sampling, when $g(\theta^*|\theta^{(t)})$ is independent of the current value $\theta^{(t)}$, that is $g(\theta^*|\theta^{(t)}) = g(\theta^*)$. If the proposal density has the form $g(\theta^*|\theta^{(t)}) = g(\theta^{(t)} - \theta^*)$, then a random walk Metropolis scheme is obtained. For example, if $d = 1$, then a univariate normal random walk takes samples $W^{(t)} \sim N(0, 1)$ and a proposal $\theta^* = \theta^{(t)} + \sigma W^{(t)}$, where σ determines the size of the jump (and the acceptance rate). A uniform random walk samples $U^{(t)} \sim U(-1, 1)$ and scales this to form a proposal $\theta^* = \theta^{(t)} + \kappa U^{(t)}$.

The rate at which proposals generated by g are accepted under equations (18) or (19) depends on how close θ^* is to $\theta^{(t)}$, how well the proposal density matches the shape of the target density $p(\theta|y)$, and the variance of the proposal density. For a normal proposal density $g = N(\theta, \Sigma)$, a higher acceptance rate is obtained by reducing the diagonal elements in Σ , but then the posterior density will take longer to explore. A high acceptance rate is not necessarily desirable as autocorrelation in sampled values will be high since the chain is moving in a restricted

portion of the posterior density. A low acceptance rate has the same problem, since the chain is getting locked at particular values. One possibility is to use a variance or dispersion matrix estimate V from a maximum-likelihood analysis, and scale it by a constant $c > 1$, so that the proposal density variance is $\Sigma = cV$ [54]. Performance also tends to be improved if parameters are transformed to the full range of positive and negative values $(-\infty, \infty)$, thus lessening the difficulty of sampling from skewed posterior densities.

In problems where d is large, θ is typically divided into D blocks or components $\theta = (\theta_1, \dots, \theta_D)$, and componentwise updating is applied. Let $\theta_{[j]} = (\theta_1, \theta_2, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_D)$ denote components apart from θ_j , and $\theta_j^{(t)}$ be the current value of θ_j . Suppose that at step j of iteration $t + 1$ the preceding $j - 1$ parameter blocks have already been updated via M-H algorithms, while $\theta_j, \theta_{j+1}, \dots, \theta_D$ are still to be updated. Let the partially updated form of $\theta_{[j]}$ be denoted

$$\theta_{[j]}^{(t,t+1)} = (\theta_1^{(t+1)}, \theta_2^{(t+1)}, \dots, \theta_{j-1}^{(t+1)}, \theta_{j+1}^{(t)}, \dots, \theta_D^{(t)}) \quad (20)$$

The candidate θ_j^* for replacing $\theta_j^{(t)}$ is generated from the j th proposal density, denoted $g(\theta_j^* | \theta_j^{(t)}, \theta_{[j]}^{(t,t+1)})$. Now acceptance of a candidate value involves assessing whether there is improvement or not in the full conditional densities $p(\theta_j^{(t)} | y, \theta_{[j]}^{(t,t+1)})$ that specify the density of θ_j conditional on other parameters $\theta_{[j]}$. The candidate value θ_j^* is accepted with probability

$$\begin{aligned} & \alpha(\theta_j^* | \theta_j^{(t)}, \theta_{[j]}^{(t,t+1)}) \\ &= \min \left[1, \frac{p(\theta_j^* | \theta_{[j]}^{(t,t+1)}, y) g(\theta_j^{(t)} | \theta_j^*, \theta_{[j]}^{(t,t+1)})}{p(\theta_j^{(t)} | \theta_{[j]}^{(t,t+1)}, y) g(\theta_j^* | \theta_j^{(t)}, \theta_{[j]}^{(t,t+1)})} \right] \end{aligned} \quad (21)$$

Gibbs Sampling

The Gibbs sampler [55–57] is a particular componentwise algorithm in which the candidate density $g(\theta^* | \theta)$ for values θ_j^* to update $\theta_j^{(t)}$ equals the full conditional $p(\theta_j^* | \theta_{[j]})$ so that proposals are accepted with probability 1. Gibbs sampling updates involve

successive parameter updating, which when completed forms the transition from $\theta^{(t)}$ to $\theta^{(t+1)}$:

$$\theta_1^{(t+1)} \sim f_1(\theta_1 | \theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_D^{(t)}) \quad (22)$$

$$\theta_2^{(t+1)} \sim f_2(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_D^{(t)}) \quad (23)$$

$$\theta_D^{(t+1)} \sim f_D(\theta_D | \theta_1^{(t+1)}, \theta_3^{(t+1)}, \dots, \theta_{D-1}^{(t+1)}) \quad (24)$$

Full conditional densities can be obtained by abstracting out from the full posterior density (proportional to likelihood times prior), those elements in the likelihood or prior that include θ_j and treating other components as constants [58]. Ideally, these are standard densities (e.g., gamma, normal) that are easy to simulate from, but nonstandard conditional densities can be sampled using adaptive rejection sampling [59] or the ratio method [60].

Assessing MCMC Convergence

There is no guarantee that sampling from an MCMC algorithm will converge to the posterior distribution, despite obtaining a high number of iterations. Slow convergence will be seen in trace plots that wander or exhibit short-term trends, rather than fluctuating rapidly around a stable mean. Failure to converge is typically evident in a subset of the model parameters; for example, fixed regression effects in a general linear mixed model may show convergence while the parameters relating to the random components may not. Often, measures of overall fit (e.g., model deviance) converge while certain component parameters do not.

Problems of convergence in MCMC sampling may reflect problems in model identifiability: either formal nonidentification as in multiple random-effects models or poor empirical identifiability due to “overfitting”, namely, fitting an overly complex model to a small sample. Choice of diffuse priors tends to increase the chance that models are poorly identified, especially in complex hierarchical models or small samples [61]. Elicitation of more informative priors or application of parameter constraints may assist identification and convergence.

Another source of poor convergence is suboptimal parameterization or data form, and so improved convergence will be obtained under alternative parameterization. For example, convergence is

improved by centering independent variables in regression applications [62]. Similarly, convergence in random-effects models may be lessened by a centered hierarchical prior [63]. Consider two way nested data, with $j = 1, \dots, m$ repetitions over subjects $i = 1, \dots, n$

$$y_{ij} = \mu + \alpha_i + u_{ij} \quad (25)$$

with $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $u_{ij} \sim N(0, \sigma_u^2)$. The centered version has $y_{ij} \sim N(\kappa_i, \sigma_u^2)$ and $\kappa_i \sim N(\mu, \sigma_\alpha^2)$. For three way nested data, the standard model form

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + u_{ijk} \quad (26)$$

with $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $\beta_{ij} \sim N(0, \sigma_\beta^2)$, whereas the centered version is $y_{ijk} \sim N(\zeta_{ij}, \sigma_u^2)$, $\zeta_{ij} \sim N(\kappa_i, \sigma_\beta^2)$, and $\kappa_i \sim N(\mu, \sigma_\alpha^2)$. Papaspiliopoulos *et al.* [64] compare the effects of centered and noncentered hierarchical model parameterizations on convergence in MCMC sampling.

Many practitioners prefer to use two or more parallel chains with diverse starting values to ensure full coverage of the sample space of the parameters [65]. Running multiple chains often assists in diagnosing poor identifiability of models. This is illustrated most clearly when identifiability constraints are missing from a model, such as in discrete mixture models that are subject to “label switching” during MCMC updating [66]. Single runs may be adequate for straightforward problems, or as a preliminary to obtain inputs to multiple chains. Convergence for multiple chains may be assessed using Gelman–Rubin scale reduction factors (often referred to simply as *G–R statistics* or *G–R factors*) that measure the convergence of the between chain variance in the sampled parameter values to the variance of all the chains combined. The ratio goes to 1 if all chains are sampling identical distributions. Parameter samples from poorly identified models will show wide divergence in the sample paths between different chains and variability of sampled parameter values between chains will considerably exceed the variability within any one chain. To measure variability of samples $\theta_j^{(t)}$ within the j th chain ($j = 1, \dots, J$) over T iterations after a burn-in of B iterations, define

$$w_j = \sum_{t=B+1}^{B+T} \frac{(\theta_j^{(t)} - \bar{\theta}_j)^2}{(T-1)} \quad (27)$$

with variability within chains V_w then defined as the average of the w_j .

Ideally, the burn-in period is a short initial set of samples where the effect of the initial parameter values tails off; during the burn-in phase, trace plots of parameters will show clear monotonic trends as they reach the region of the posterior.

Between chain variance is measured by

$$V_B = \frac{T}{J-1} \sum_{j=1}^J (\bar{\theta}_j - \bar{\theta})^2 \quad (28)$$

where $\bar{\theta}$ is the average of the $\bar{\theta}_j$. The potential scale reduction factor compares a pooled estimator of $\text{Var}(\theta)$, given by $V_P = V_B/T + TV_W/(T-1)$ with the within sample estimate V_W . Specifically, the potential scale reduction factor (PSRF) is $(V_P/V_W)^{0.5}$ with values under 1.2 indicating convergence.

Another multiple chain convergence statistic is due to Brooks and Gelman [67] and known as the *Brooks–Gelman–Rubin (BGR) statistic*. This is a ratio of parameter interval lengths, where for chain j the length of the $100(1-\alpha)\%$ interval for a parameter is the gap between 0.5α and $(1-0.5\alpha)$ points from T simulated values $\theta^{(t)}$. This provides J within-chain interval lengths, with mean I_W . For the pooled output of TJ samples, the same $100(1-\alpha)\%$ interval I_P is also obtained. Then the ratio I_P/I_W should converge to 1 if there is convergent mixing over different chains. Brooks and Gelman also propose a multivariate version of the original G–R scale reduction factor.

Since parameter samples obtained by MCMC methods are dependent, there will be correlations at lags 1, 5, etc., the size of which depends (*inter alia*) on the form of parameterization, the complexity of the model, and the form of MCMC sampling used (e.g., block or univariate sampling). Nonvanishing autocorrelations at high lags mean that less information about the posterior distribution is provided by each iterate and a higher sample size T is necessary to cover the parameter space. Analysis of autocorrelation in sequences of MCMC samples amounts to an application of time series methods, in regard to issues such as assessing stationarity in an autocorrelated sequence. Autocorrelation at lags 1, 2, and so on may be assessed from the full set of sampled values $\theta^{(t)}, \theta^{(t+1)}, \theta^{(t+2)}, \dots$, or from subsamples K steps apart $\theta^{(t)}, \theta^{(t+K)}, \theta^{(t+2K)}, \dots$, etc. If the chains are mixing satisfactorily, then the autocorrelations in the one step apart iterates $\theta^{(t)}$ will fade to zero as the lag increases (e.g., at lag 10 or 20).

Model Fit and Predictions from the Model

Methods to assess models must be able to choose among different models, though if a few models are closely competing, an alternative strategy is to average over models. However, choosing the best fitting model is not an end to model criticism since even the best model may not adequately reproduce the data.

Consider model choice first. Let m be a multinomial model index, with m taking values between 0 and K . Formal Bayes' model choice is based on prior model probabilities $P(m = k)$ and marginal likelihoods $P(y|m = k)$, which are the normalizing constants in the Bayes' formula,

$$P(\theta_k|y, m = k) = \frac{P(y|\theta_k, m = k)P(\theta_k|m = k)}{P(y|m = k)} \quad (29)$$

The marginal likelihood can be written as

$$P(y|m = k) = \frac{P(y|\theta_k, m = k)\pi(\theta_k|m = k)}{P(\theta_k|y, m = k)} \quad (30)$$

or following a log transform as

$$\begin{aligned} \log[P(y|m = k)] &= \log[P(y|\theta_k, m = k)] \\ &\quad + \log[P(\theta_k|m = k)] \\ &\quad - \log[P(\theta_k|y, m = k)] \end{aligned} \quad (31)$$

The term $\log[P(\theta_k|m = k)] - \log[P(\theta_k|y, m = k)]$ is a penalty favoring parsimonious models, whereas a more complex model virtually always leads to a higher log-likelihood $\log[P(y|\theta_k, m = k)]$. The marginal likelihood is thus the probability of the data y given a model, and is obtained by averaging over the priors assigned to the parameters in that model, since

$$P(y|m = k) = \int P(y|\theta_k, m = k)\pi(\theta_k|m = k) d\theta_k \quad (32)$$

Suppose choice is confined to two possible models, $m = 0$ and 1. One may fit them separately and consider their relative fit in terms of summary statistics, such as the marginal likelihood. Alternatively, using model search techniques such as reversible jump MCMC [62], one may search over models k as well as over parameter values $\{\theta_k|m = k\}$. The best

model is chosen on the basis of posterior probabilities on each model. Under the first scenario, such probabilities involve the ratio of marginal likelihoods $P(y|m = 1)/P(y|m = 0)$, also termed the *Bayes' factor* and denoted B_{10} , and the ratio of the prior probabilities, $Q_1 = P(m = 1)/P(m = 0)$. The posterior probability of a model is obtained from its prior probability and the marginal likelihood *via* the formula $P(m = k|y) = P(m = k)P(y|m = k)/P(y)$. It follows that

$$\begin{aligned} P(m = 1|y)/P(m = 0|y) &= P(y|m = 1)P(m = 1)/[P(y|m = 0)P(m = 0)], \\ &= B_{10}[P(m = 1)/P(m = 0)] \end{aligned} \quad (33)$$

namely, that the posterior odds on model 1 equal the Bayes' factor times the prior odds on model 1. Under the second scenario, the proportion of samples when model k is chosen is equal to $P(m = k|y)$ when prior model probabilities are equal. There is no necessary constraint in these model comparisons that models 0 and 1 are nested with respect to one another – an assumption often necessarily made in classical tests of goodness of model fit.

If the posterior model probability of a particular model among $K + 1$ options is above some critical threshold (e.g., $P(m = k|y) > 0.01$), then model averaging may be carried out [68]. The average density of a parameter function Δ is obtained as

$$P(\Delta|y) = \sum_{k=0}^{K} P(\Delta_k|m = k, y)P(m = k|y) \quad (34)$$

Model averaging can be done in other ways, such as when regression variable selection is used, or when time series switching models are applied [69].

The marginal likelihood can be difficult to obtain in complex models and other “informal” techniques are often used. A widely used approach is analogous to the Akaike Information Criterion (AIC) in classical statistics [70], namely, the deviance information criterion (DIC) of Spiegelhalter *et al.* [71]. In developing this criterion, Spiegelhalter *et al.* propose an estimate for the total number of effective parameters or model dimension, denoted d_e , generally less than the nominal number of parameters in hierarchical random-effects models. Let $L^{(t)} = \log[P(y|\theta^{(t)})]$ denote the log-likelihood obtained at the t th iteration in an MCMC sequence, with $D^{(t)} = -2L^{(t)}$

then being the deviance at that iteration. Then $d_e = E(D|y, \theta) - D(\bar{\theta}|y)$, where $E(D|y, \theta)$ is estimated by the mean $\bar{D}$ of the sampled deviances $D^{(t)}$, and $D(\bar{\theta}|y)$ is the deviance at the posterior mean $\bar{\theta}$ of the parameters. The DIC is then

$$D(\bar{\theta}|y) + 2d_e \quad (35)$$

with the model having the lowest DIC being preferred. Effective parameter estimates in practice include aspects of a model such as the precision of its parameters and predictions.

While a DIC or posterior model probability might favor one model over another it still remains a possibility that none of the models being considered reproduces the data effectively. Model checking involves assessing whether predictions from the model reproduce the main features of the data and particular unusual features of the data as well such as skewness or bimodality. The most widely applied Bayesian model checking procedures are based on replicate data y_{rep} sampled from the model. These also constitute a model prediction under the Bayesian approach and often figure centrally in uncertainty statements in risk assessment applications; examples include the prediction of risk of infectious disease over a region using risk mapping methods based on existing disease point patterns [72].

In the Bayesian method, the information about θ is contained in the posterior density $p(\theta|y)$ and so prediction is correspondingly based on averaging predictions $p(y_{\text{rep}}|y, \theta)$ over this posterior density. Generally $p(y_{\text{rep}}|y, \theta) = p(y_{\text{rep}}|\theta)$, namely, that predictions are independent of the observations given θ , so that the posterior predictive density is obtained as

$$p(y_{\text{rep}}|y) = \int p(y_{\text{rep}}|\theta)p(\theta|y) d\theta \quad (36)$$

A model checking procedure based on the posterior predictive density $p(y_{\text{rep}}|y)$, involves a discrepancy measure $D(y_{\text{obs}}; \theta)$, such as the deviance or χ^2 [73]. The realized value of the discrepancy $D(y_{\text{obs}}; \theta)$ is located within its reference distribution by a tail probability analogous to a classical p value:

$$p_b(y_{\text{obs}}) = P_R[D(y_{\text{rep}}; \theta) > D(y_{\text{obs}}; \theta) | y_{\text{obs}}] \quad (37)$$

In practice, this involves calculating $D(y_{\text{rep}}^{(t)}, \theta^{(t)})$ and $D(y_{\text{obs}}, \theta^{(t)})$ in an MCMC run of length T and then calculating the proportion of samples for

which $D(y_{\text{rep}}^{(t)}, \theta^{(t)})$ exceeds $D(y_{\text{obs}}, \theta^{(t)})$. Systematic differences in distributional characteristics (e.g., in percents of extreme values or in ratios of variances to means) between replicate and actual data indicate possible limitations in the model(s). Specifically, values of p_b around 0.5 indicate a model consistent with the actual data, whereas extreme values (close to 0 or 1) suggest inconsistencies between model predictions and actual data.

Another model checking procedure based on replicate data is suggested by Gelfand [74] and involves checking for all sample cases $i = 1, \dots, n$ whether observed y are within 95% intervals of y_{rep} . This procedure may also assist in pointing to possible overfitting, e.g., if all (i.e., 100%) of the observed y are within 95% intervals of y_{rep} .

Illustrative Applications in WINBUGS

We consider two worked examples where principles of risk assessment are involved. The analysis uses the freeware WINBUGS package (now being developed under the acronym OPENBUGS) which has a syntax based on the S+ and R packages. Discussion on how to use WINBUGS appears in [75, 76], and [77, Appendix B].

Ground Ozone Exceedances

The first application involves an air quality time series (*see Air Pollution Risk*). Elevated concentrations of ground level ozone are harmful to human health and ecosystems. In humans, concentrations of $200 \mu\text{g m}^{-3}$ (over 8 h) cause irritation to the eyes and nose and very high levels (over $1000 \mu\text{g m}^{-3}$) may inhibit lung function through inflammatory response. Ground level (tropospheric) ozone also may contribute to acid rain formation and damage some man-made materials such as elastomers and fabrics.

In the United Kingdom, an Air Quality Strategy objective has been set for ground level ozone, namely, that the maximum daily concentration (measured via 8-h running means through the day) should not exceed $100 \mu\text{g m}^{-3}$ for more than 10 days yr^{-1} . Hence an annual record for a given recording station will consist of $365 \times 24 = 8760$ readings of 8-h running means. We consider such data through the calendar year 2005 for air quality monitoring site in England, namely, Westminster in Central London

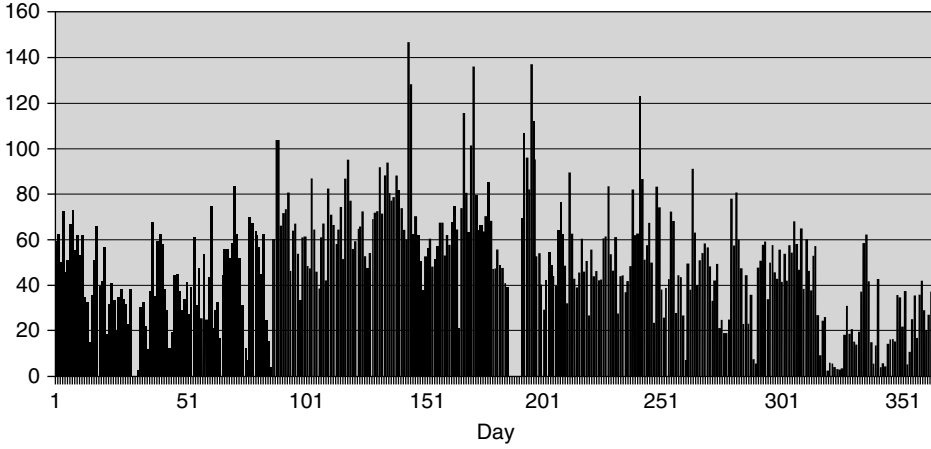


Figure 1 Maximum daily readings, Westminster, UK in 2005

(see <http://www.londonair.org.uk/>). Figure 1 plots the daily maximum readings.

Such series tend to have missing values; in England, the Department for Environment, Food, and Rural Affairs set a minimum level of 75% data capture for station records to be included in official returns, though the series under consideration has only 5% missing values. Figure 1 shows a spell of missing values in the early summer months (May, June) when ozone readings are at their highest. We assume missingness at random [78], so that inferences can proceed without modeling the missingness probability mechanism.

We wish to assess the likely range of days with exceedance, namely, days with maximum 8-h running mean exceeding $100 \mu\text{g m}^{-3}$. A predicted exceedance total is required for the actual data since there are missing observations; hence the observed value of 10 exceedances, based on a data capture rate of $8364/8760 = 0.955$, is a minimum. The data can be regarded in several ways: in terms of hour effects repeated each day through the year, individual day effects, month effects, etc.

Preliminary analysis of the data structure suggests both nonconstant means and variances, and while such heterogeneity can be modeled by day specific or even parameters specific to each of the 8760 readings, a relatively parsimonious option assumed here is for month and hour specific effects to define changing means μ_t and log precisions ω_t at the level of readings. For readings over $t = 1, \dots, 8760$ points, let $m = m_t$ ($m = 1, \dots, 12$) and $h = h_t$ ($h =$

$1, \dots, 24$) denote the associated month and hour. The data are then assumed to be truncated normal with

$$y_t \sim N(\mu_t, \exp(-\omega_t))I(0, \infty) \quad (38)$$

$$\mu_t = \alpha_{1m} + \alpha_{2h} \quad (39)$$

$$\omega_t = \gamma_{1m} + \gamma_{2h} \quad (40)$$

where all γ and α parameters are random. To identify the level of the hour effects $\{\alpha_{2h}, \gamma_{2h}\}$ a corner constraint is used, with $\alpha_{21} = \gamma_{21} = 0$. Note that the models for μ_t and ω_t omit an intercept, so that the α_{1m} and γ_{1m} series are identified without a corner constraint being needed. Normal first-order random walks are assumed for the four series $\{\alpha_{1m}, \gamma_{1m}, \alpha_{2h}, \gamma_{2h}\}$, with γ priors on the precisions. See below for the code.

For future planning what is of relevance is the predicted exceedance total over the whole year adjusting for the 4.5% data loss. This has mean 10.6 with 95% interval (10, 12). Also relevant to assessing uncertainty in the annual exceedance total is the posterior predictive density $p(y_{\text{rep}}|y)$ where in the present analysis y includes both observed and missing readings, $y = (y_{\text{obs}}, y_{\text{mis}})$.

A two chain run of 10 000 iterations with diverse starting values is made, with convergence apparent early (within 500 iterations). Model checks using the method of Gelfand [79] are satisfactory. The expected mean exceedance total from the posterior predictive density $p(y_{\text{rep}}|y)$ is 13.5, with a 95% interval from 7 to 21 days. Figure 2 contains the posterior density

```

model {for (t in 1:8760) {exom[t] <- exp(omeg[t])
y[t] ~ dnorm(mu[t],exom[t]) I(0,); yrep[t] ~ dnorm(mu[t],exom[t]) I(0,)
mu[t] <- alph1[month[t]] +alph2[hour[t]]
omeg[t] <- gam1[month[t]] +gam2[hour[t]]}

for (m in 2:12) {alph1[m] ~ dnorm(alph1[m-1],tau[1])
gam1[m] ~ dnorm(gam1[m-1],tau[2])}
for (h in 2:24) {alph2[h] ~ dnorm(alph2[h-1],tau[3])
gam2[h] ~ dnorm(gam2[h-1],tau[4])}
alph1[1] ~ dnorm(0,0.001); gam1[1] ~ dnorm(0,0.001)
alph2[1] <- 0; gam2[1] <- 0
for (j in 1:4) {tau[j] ~ dgamma(1,0.001)}
# exceedances
for (t in 1:8760) {exobs[day[t],hour[t]] <- step(y[t]-100);
exnew[day[t],hour[t]] <- step(yrep[t]-100)}
for (d in 1:365) {exday1[d] <- step(sum(exnew[d,1:24])-1);
exday2[d] <- step(sum(exobs[d,1:24])-1)}
Extot[1] <- sum(exday1[]); Extot[2] <- sum(exday2[])}
```

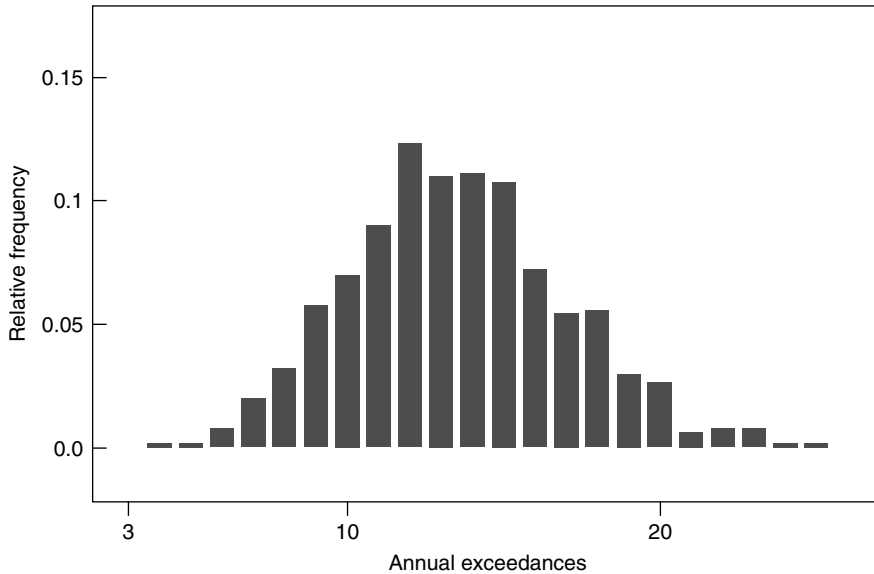


Figure 2 Posterior predictive density for exceedances

of the predicted number of exceedances. One might go on to investigate questions such as the chance of more than 15 exceedances in a year or five or more in any particular month. Figure 3 plots the posterior means and 95% credible intervals for the month effects α_{1m} in the level of ozone. This makes the strong monthly variations clear. It may be noted finally that other approaches to analyzing air quality are possible; for example, Raftery [80] adopts a

Bayesian model which considers days between ozone exceedances as a Poisson process.

Arsenic in Drinking Water and the Associated Cancer Risk

A 1999 study by the National Academy of Sciences [81] found that arsenic in drinking water causes bladder, lung, and skin cancer, and may also cause

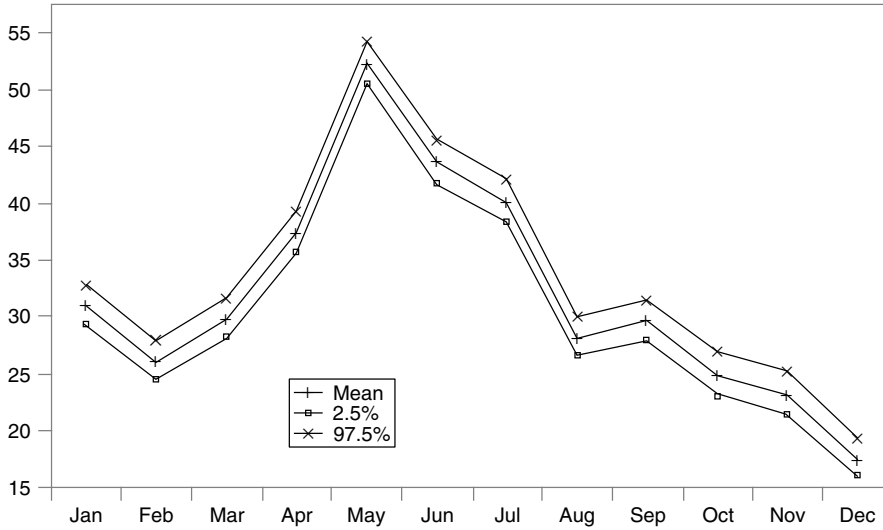


Figure 3 Plot of α_{1m} parameters

kidney and liver cancer. It estimated that daily ingestion of water containing $50 \mu\text{g l}^{-1}$ of arsenic would add about 1% to a person's lifetime risk of dying from cancer. Morales *et al.* [82] undertook an analysis of cancer risk in relation to a US congress requirement on the US Environmental Protection Agency to revise its standard for arsenic in drinking water below the $50 \mu\text{g l}^{-1}$ level which was then in place. They made a risk assessment for cancers of the bladder, liver, and lung resulting from differing exposure level to arsenic in water. This was based on accumulated cancer mortality data from 42 villages in South West Taiwan relying on well water subject to arsenic contamination. The comparison population is from the same region but with access to uncontaminated water. Relevant data appears in Table A10 of the NRC report and in Table 2 of Morales *et al.*

The data consist of cancer deaths for 13 adult age bands (20–24, 25–29, ..., 80–85) in the 42 villages and the background population with zero exposure – so there are $43 \times 13 = 559$ data cells. The 43 populations are denoted $v = 1, \dots, 43$ and the age-groups are denoted $t = 1, \dots, 13$. The analysis by Morales *et al.* uses Poisson regression of cancer deaths y_{vt} over 1973–1986 in relation to population years P_{vt} exposed to risk. The predictors are centered ages of each band (i.e., 22.5 for the first band, up to 82.5 for the last) and the exposure level, x_v , in each population in micrograms per liter. These

range up to over $900 \mu\text{g l}^{-1}$ in one population. Ages are transformed to be centered and standardized, with a_t denoting transformed age. There are 38 different exposure levels since some villages had equal exposures. Thus $y_{vt} \sim \text{Po}(h_{vt} P_{vt})$, and we follow one of the model forms used by Morales *et al.* assuming independent effects of age and exposure, namely, a quadratic exponential age effect and a linear dose effect. Let $x = x_v$ be the exposure in population v then the hazard (mortality) rate in population v at age t is

$$h_{xt} = (1 + \gamma x)(\exp(\beta_1 + \beta_2 a_t + \beta_3 a_t^2)) \quad (41)$$

Let q_t be life table death rates (for 5 year age-groups t), which Morales *et al.* take from a life table for the United States in 1996. Then the lifetime risk LR_x at exposure x of mortality from a particular cancer or group of cancers over the course of a typical lifetime can be calculated by integrating the hazard h_{xt} over t , i.e., over the typical lifetime. So

$$\begin{aligned} LR_x &= \int_0^\infty h_{xt} S_t dt \\ &\approx 1 - \sum_t q_t \exp\left(-5 \sum_{r \leq t} h_{xr}\right) \end{aligned} \quad (42)$$

where S_t is the probability of surviving till age t .

Here we consider female bladder cancer and consider the predicted risk and its 95% credible intervals at exposures $x^* = \{0, 5, 10, \dots, 500\} \mu\text{g l}^{-1}$.

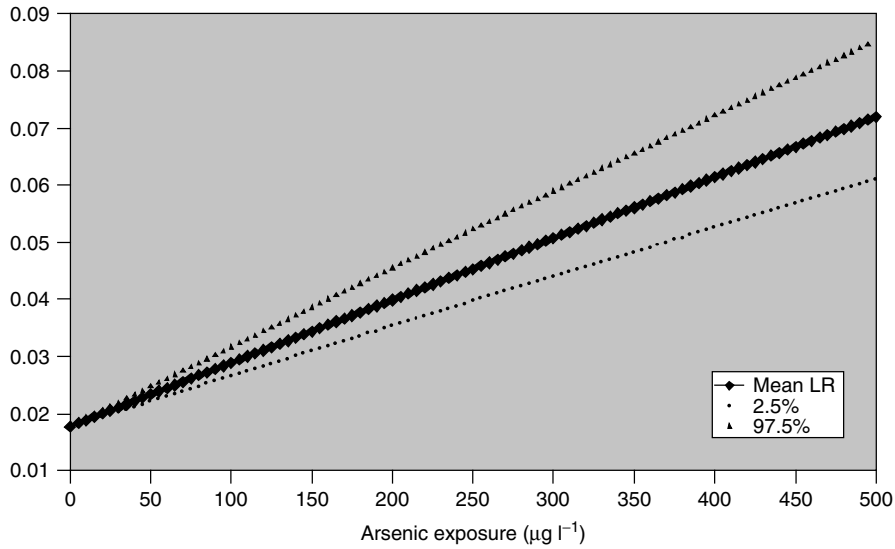


Figure 4 Lifetime bladder cancer mortality risk (F)

```
model { for (i in 1:559) {y[i] ~ dpois(mu[i]); mu[i] <- h[i]*P[i]
                                h[i] <- h1[age[i]]*h2[expos[i]]}

for (t in 1:13) {h1[t] <-
exp(beta[1]+beta[2]*a[t]+beta[3] *pow(a[t],2))}
for (x in 1:38) {h2[x] <- 1+gam*e[x]}
# cum hazard
for (x in 1:101) {H[x,1] <- haz[x,1]
                  estar[x] <- (x-1)*5
# hazards by age and exposures 0,5,10,.500
for (t in 1:13) {haz[x,t] <- h1[t]*(1+gam*estar[x])}
for (t in 2:13) {H[x,t] <- H[x,t-1]+haz[x,t]}
for (x in 1:101) {LR[x] <- 1-sum(r[x,]); RLR[x] <- LR[x]/LR[1]
# check if excess risk over 10%
                  p10[x] <- step(RLR[x]-1.1)
for (t in 1:13) {r[x,t] <- q[t+1]*exp(-5*H[x,t])}}
# priors
gam ~ dnorm(0,0.001); beta[1] ~ dnorm(-10,0.001);
beta[2] ~ dnorm(0,0.001); beta[3] ~ dnorm(0,0.001)}
```

Figure 4 plots the lifetime risk at exposures up to $500 \mu\text{g l}^{-1}$. The excess risk relative to the zero exposure group and the probability (for various levels of x) that the excess risk exceeds 10% are also derived. Analysis is by the WINBUGS code (see above):

The parameter γ has a mean (95% interval) of 0.032 (0.024, 0.041) and so it is clear that the risk of cancer increases with arsenic exposure. We find a 36% posterior probability that a $15 \mu\text{g l}^{-1}$ exposure leads to 10% excess risk and a

96% probability that a $20 \mu\text{g l}^{-1}$ exposure leads to 10% excess risk. Exposures of $25 \mu\text{g l}^{-1}$ and above have a 100% probability of leading to 10% excess risk.

References

- [1] Morgan, M., Henrion, M. & Small, M. (1992). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press.

- [2] Gelman, A., Carlin, J., Stern, H. & Rubin, D. (2004). *Bayesian Data Analysis*, 2nd Edition, CRC: Chapman & Hall.
- [3] Congdon, P. (2006). *Bayesian Statistical Modelling*, 2nd Edition, John Wiley & Sons.
- [4] Carlin, B. & Louis, T. (2000). *Bayes and empirical Bayes methods for data analysis*, 2nd Edition, Chapman and Hall-CRC Press Boca Raton.
- [5] Gamerman, D. & Lopes, H. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd Edition, CRC/Chapman & Hall.
- [6] Bernardo, J. & Smith, A. (2001). *Bayesian Theory*, 2nd Edition, John Wiley & Sons.
- [7] Chen, M., Shao, Q. & Ibrahim, J. (2001). *Monte Carlo Methods in Bayesian Computation*, Springer.
- [8] Fox, D. (2006). Statistical issues in ecological risk assessment, *Human and Ecological Risk Assessment* **12**, 120–129.
- [9] Paulo, M., van der Voet, H., Jansen, M., ter Braak, C. & van Klaveren, J. (2005). Risk assessment of dietary exposure to pesticides using a Bayesian method, *Pest Management Science* **61**, 759–766.
- [10] Barbolini, M., Cappabianca, F. & Savi, F. (2004). Risk assessment in avalanche-prone areas, *Annals of Glaciology* **38**, 115–122.
- [11] Bates, S., Cullen, A. & Raftery, A. (2003). Bayesian uncertainty assessment in multicompartiment deterministic simulation models for environmental risk assessment, *Environmetrics* **44**, 355–371.
- [12] Qin, X., Ivan, J., Ravishanker, N. & Liu, J. (2005). Hierarchical Bayesian estimation of safety performance functions for two-lane highways using Markov Chain Monte Carlo modeling, *Journal of Transportation Engineering* **131**, 345–351.
- [13] Messner, M., Chappell, C. & Okhuysen, P. (2001). Risk assessment for cryptosporidium: a hierarchical Bayesian analysis of human dose response data, *Water Research* **35**, 3934–3940.
- [14] Lye, L. (1990). Bayes estimate of the probability of exceedance of annual floods, *Stochastic Environmental Research and Risk Assessment* **4**, 55–64.
- [15] Van Gelder, P. (1996). A Bayesian analysis of extreme water levels along the Dutch coast using flood historical data, *Stochastic Hydraulics* **7**, 243–251.
- [16] Siu, N. & Kelly, D. (1998). Bayesian parameter estimation in probabilistic risk assessment, *Reliability Engineering and System Safety* **62**, 89–116.
- [17] Draper, D. (1995). Assessment and propagation of model uncertainty, *Journal of the Royal Statistical Society Series B* **57**, 45–97.
- [18] Manson, S. (2002). Decision making and uncertainty: Bayesian analysis of potential flood heights, *Geographical Analysis* **34**, 112–129.
- [19] Jarup, L. (2004). Health and environment information systems for exposure and disease mapping, and risk assessment, *Environmental Health Perspectives* **112**, 995–997.
- [20] Friesen, M., MacNab, Y., Marion, S., Demers, P., Davies, H. & Teschke, K. (2006). Mixed models and empirical Bayes estimation for retrospective exposure assessment of dust exposures in Canadian sawmills, *Annals of Occupational Hygiene* **50**, 281–288.
- [21] Symanski, E., Sallsten, G. & Chan, W. (2001). Heterogeneity in sources of exposure variability among groups of workers exposed to inorganic mercury, *Annals of Occupational Hygiene* **45**, 677–687.
- [22] George, E., Makov, U. & Smith, A. (1993). Conjugate likelihood distributions, *Scandinavian Journal of Statistics* **20**, 147–156.
- [23] Gelman, A. & Rubin, D. (1996). Markov chain Monte Carlo methods in biostatistics, *Statistical Methods in Medical Research* **5**, 339–355.
- [24] Hodge, R., Evans, M., Marshall, J., Quigley, J. & Walls, L. (2001). Eliciting engineering knowledge about reliability during design-lessons learnt from implementation, *Quality and Reliability Engineering International* **17**, 169–179.
- [25] Siu, N. & Kelly, D. (1998). Bayesian parameter estimation in probabilistic risk assessment, *Reliability Engineering and System Safety* **62**, 89–116.
- [26] Birkes, D. & Dodge, Y. (1993). *Alternative Methods of Regression*, John Wiley & Sons, New York.
- [27] Ibrahim, J. & Chen, M. (2000). Power prior distributions for regression models, *Statistical Science* **15**, 46–60.
- [28] Chen, M.-H., Ibrahim, J. & Shao, Q.-M. (2000). Power prior distributions for generalized linear models, *Journal of Statistical Planning and Inference* **84**, 121–137.
- [29] Besag, J., York, J. & Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics* **43**(1), 1–59.
- [30] Zhu, M. & Lu, A. (2004). The counter-intuitive non-informative prior for the Bernoulli family, *Journal of Statistical Education* **12**, 1–10.
- [31] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis* **1**, 515–533.
- [32] Kass, R. & Wasserman, L. (1996). The selection of prior distributions by formal rules, *Journal of American Statistical Association* **91**, 1343–1370.
- [33] Gelfand, A. & Sahu, S. (1999). Gibbs sampling, identifiability and improper priors in generalized linear mixed models, *Journal of American Statistical Association* **94**, 247–253.
- [34] Daniels, M. (1999). A prior for the variance in hierarchical models, *Canadian Journal of Statistics* **27**, 567–578.
- [35] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis* **1**, 515–533.
- [36] Gustafson, P., Hossain, S. & MacNab, Y. (2006). Conservative priors for hierarchical models, *Canadian Journal of Statistics* **34**, 377–390.
- [37] Berger, J. & Bernardo, J. (1992). On the development of reference priors, in *Bayesian Statistics IV*, J. Bernardo,

- J. Berger, A. Dawid & A. Smith, eds, Oxford University Press, pp. 35–60.
- [38] Gustafson, P. (1996). Robustness considerations in Bayesian analysis, *Statistical Methods in Medical Research* **5**, 357–373.
- [39] Spiegelhalter, D., Freedman, L. & Parmar, M. (1994). Bayesian approaches to randomised trials, *Journal of Royal Statistical Society, Series A* **157**, 357–416.
- [40] Gustafson, P. (1996). Robustness considerations in Bayesian analysis, *Statistical Methods in Medical Research* **5**, 357–373.
- [41] Berger, J. (1990). Robust Bayesian analysis: sensitivity to the prior, *Journal of Statistical Planning and Inference* **25**, 303–328.
- [42] Berger, J., Moreno, E., Pericchi, L., Bayarri, M., Bernardo, J., Cano, J. & Horra, J. (1994). An overview of robust Bayesian analysis, *Test* **3**, 5–124.
- [43] Gilks, W., Richardson, S. & Spiegelhalter, D. (1996). Introducing Markov Chain Monte Carlo, in *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson & D. Spiegelhalter, eds, Chapman & Hall, London, pp. 1–20.
- [44] Smith, A. & Gelfand, A. (1992). Bayesian statistics without tears: a sampling-resampling perspective, *The American Statistician* **46**, 84–88.
- [45] van Dyk, D. (2002). Hierarchical models, data augmentation, and MCMC, in *Statistical Challenges in Modern Astronomy III*, G. Babu & E. Feigelson, eds, Springer, New York, pp. 41–56.
- [46] Tierney, L. (1994). Markov chains for exploring posterior distributions, *Annals of Statistics* **22**, 1701–1728.
- [47] Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall.
- [48] Gelfand, A., Hills, S., Racine-Poon, A. & Smith, A. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling, *Journal of the American Statistical Association* **85**, 972–985.
- [49] Lange, N. & Ryan, L. (1989). Assessing normality in random effects models, *Annals of Statistics* **17**, 624–642.
- [50] Hastings, W. (1970). Monte-Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97–109.
- [51] Roberts, G. & Rosenthal, J. (2004). General state space Markov chains and MCMC algorithms, *Probability Surveys* **1**, 20–71.
- [52] Chib, S. & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm, *American Statistician* **49**, 327–345.
- [53] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**, 1087–1092.
- [54] Roberts, G., Gelman, A. & Gilks, W. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms, *Annals of Applied Probability* **7**, 110–120.
- [55] Gelfand, A. & Smith, A. (1990). Sampling based approaches to calculating marginal densities, *Journal of American Statistical Association* **85**, 398–409.
- [56] Gilks, W., Clayton, D., Spiegelhalter, D., Best, N., McNeil, A., Sharples, L. & Kirby, A. (1993). Modelling complexity: applications of Gibbs sampling in medicine, *Journal of Royal Statistical Society B* **55**, 39–52.
- [57] Casella, G. & George, E. (1992). Explaining the Gibbs sampler, *American Statistician* **46**, 167–174.
- [58] Gilks, W. (1996). Full conditional distributions, in *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson & D. Spiegelhalter, eds, Chapman & Hall, London, pp. 75–88.
- [59] Gilks, W. & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling, *Applied Statistics* **41**, 337–348.
- [60] Wakefield, J., Gelfand, A. & Smith, A. (1991). Efficient generation of random variates via the ratio-of-uniforms method, *Statistics and Computing* **1**, 129–133.
- [61] Gelfand, A. & Sahu, S. (1999). Gibbs sampling, identifiability and improper priors in generalized linear mixed models, *Journal of American Statistical Association* **94**, 247–253.
- [62] Zuur, G., Garthwaite, P. & Fryer, R. (2002). Practical use of MCMC methods: lessons from a case study, *Biometrical Journal* **44**, 433–455.
- [63] Gelfand, A., Sahu, S. & Carlin, B. (1995). Efficient parameterization for normal linear mixed effects models, *Biometrika* **82**, 479–488.
- [64] Papaspiliopoulos, O., Roberts, G. & Skold, M. (2003). Non-centered parameterisations for hierarchical models and data augmentation, in *Bayesian Statistics 7*, J. Bernardo, S. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith & M. West, eds, Oxford University Press, pp. 307–326.
- [65] Gelman, A. & Rubin, D. (1996). Markov chain Monte Carlo methods in biostatistics, *Statistical Methods in Medical Research* **5**, 339–355.
- [66] Fruhwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models, *Journal of American Statistical Association* **96**, 194–209.
- [67] Brooks, S. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations, *Journal of Computational and Graphical Statistics* **7**, 434–456.
- [68] Hoeting, J., Madigan, D., Raftery, A. & Volinsky, C. (1999). Bayesian model averaging, *Statistical Science* **14**, 382–401.
- [69] Lu, Z. & Berliner, L. (1999). Markov switching time series models with application to a daily runoff series, *Water Resources Research* **35**, 523–534.
- [70] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**(6), 716–723.
- [71] Spiegelhalter, D., Best, N., Carlin, B. & van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of Royal Statistical Society, Series B* **64**, 583–639.
- [72] Jiruše, M., Machek, J., Beneš, V. & Zeman, P. (2004). A Bayesian estimate of the risk of tick-borne diseases, *Applications of Mathematics* **49**, 389–404.

- [73] Gelman, A., Meng, X. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies, *Statistica Sinica* **6**, 733–807.
- [74] Gelfand, A. (1996). Model determination using sampling-based methods, in *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson & D. Spiegelhalter, eds, Chapman & Hall, London, pp. 145–161.
- [75] Fryback, D., Stout, N. & Rosenberg, M. (2001). An elementary introduction to Bayesian computing using WinBUGS, *International Journal of Technology Assessment in Health Care* **17**, 96–113.
- [76] Scollnik, D. (2001). Actuarial modeling with MCMC and BUGS, *North American Actuarial Journal* **5**, 96–124.
- [77] Woodworth, G. (2004). *Biostatistics: A Bayesian Introduction*, John Wiley & Sons.
- [78] Rubin, D. (1976). Inference and missing data, *Biometrika* **63**, 581–592.
- [79] Gelfand, A. (1996). Model determination using sampling-based methods, in *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson & D. Spiegelhalter, eds, Chapman & Hall, London, pp. 145–161.
- [80] Raftery, A.E. (1989). Are ozone exceedance rates decreasing? *Statistical Science* **4**, 378–381.
- [81] National Research Council (1999). *Arsenic in Drinking Water*, National Academies Press, <http://www.nap.edu/books/0309063337/html>.
- [82] Morales, K., Ryan, L., Kuo, T., Wu, M. & Chen, C. (2000). Risk of internal cancers from arsenic in drinking water, *Environmental Health Perspectives* **108**, 655–661.

Further Reading

- Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82**, 711–732.

PETER CONGDON

Behavioral Decision Studies

Probability judgments are an important part of our daily lives. Considering the conclusions we draw from observations, the arguments we make in discussions, and the decisions we make, accurate probability judgments often determine the success of our actions and decisions. Probability judgments are at the heart of many of our core life decisions: What is the probability of getting a grant? What is the probability of winning a law suit? What are the chances of successfully implementing a software system or launching a new product line?

The common denominator of many studies in psychology is the assumption that human probability judgments are often biased [1]. People, for instance, overestimate low probabilities [2], think bad things will happen to others but not to themselves [3], or simply rely on a few salient observations instead of many, representative ones [4]. In this review, we question this view and present evidence of humans' remarkable capacities in judging probabilities.

From an abstract point of view, probability judgments often take the form of one of the following questions: What is the probability that an object X is part of class Y ? What is the probability that a particular event E is the consequence of a process P ? Or, more generally: what is the probability that event E will occur? We discuss the cognitive mechanisms that allow people to answer these questions and show that people often do so fairly adaptively. We start with people's judgments for single risks, continue with probability judgments for combined events, and then discuss how and when people use base-rate information. We close by presenting evidence that imperfect judgments can, nevertheless, be adaptive.

Judging Single Risks

When asked to estimate the frequency of a risk such as a car accident or a flood, people make a frequency judgment for the occurrence of that particular event. The question is, can people judge risk frequencies correctly, or do biases occur? In one classic study, participants estimated the mortality rates of various risks (in the United States), including floods, tornadoes, and homicide [2]. The results reveal two different

phenomena (Figure 1): First, people overestimate rare risks, such as death from botulism or smallpox vaccination, and they underestimate frequent risks, such as cancer or strokes. Overestimation of rare and underestimation of frequent risks can be seen as incidences of biased risk judgments. Second (and this is in contrast to the first point), people judge risk frequencies rather accurately. This claim is supported by a correlation of $r = 0.74$ between objective risks and people's judged risk frequencies [5]. Following the convention in the social sciences that correlations larger than $r = 0.50$ are considered strong, one can conclude that people perceive risk frequencies very accurately. That is, they can predict a rank order of risks very well. Taken together, systematic misperception of risk frequencies and accurate risk perception go hand in hand and are not contradictory. Several explanations have been offered for the two phenomena. The most prominent are regression to the mean [6] and the availability heuristic [4].

Regression to the mean is a common phenomenon in the social sciences and refers to the fact that extreme values (i.e., both low and high ones) tend to the mean [6, 7]. *Floor* and *ceiling* effects have been offered as one explanation for regression to the mean. A floor effect occurs when data cannot fall short of a particular value, and a ceiling effect occurs when data cannot surpass a particular value. Floor effects can explain why rare risks are overestimated, because a floor of zero occurrences constitutes a lower bound. A rare event, thus, is more likely to be over- than underestimated. Floor effects can explain why people in Lichtenstein *et al.* study [2] overestimated rare risks (Figure 1). Ceiling effects are more problematic here, because frequency estimates have no definite upper bound. That is, whereas floor effects might explain the overestimation of rare risks, ceiling effects cannot explain the underestimation of frequent risks.

A more fruitful explanation of regression to the mean asserts that the phenomenon is a necessary consequence as soon as two variables do not correlate perfectly [8]. Suppose a person knows nothing about the frequencies of the various risks. If this is the case, the line depicting this person's risk judgments would be completely flat (i.e., slope = 0), because all risks will be judged equally probable. The more the person knows, the steeper the curve will be. In an uncertain world, however, the person will not learn perfectly and the curve's slope will remain lower than 1. As long as the correlation between objective risks and

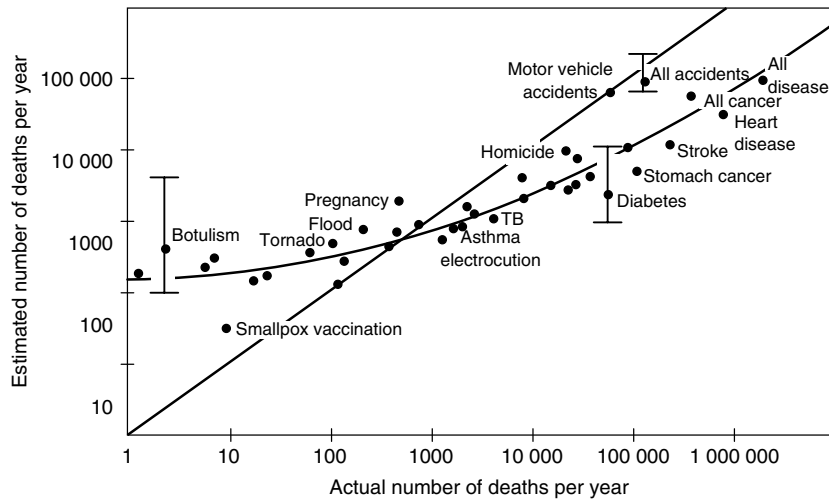


Figure 1 Relationship between estimated and actual number of deaths per year for 41 causes of death. The curved line is the best-fitting quadratic regression line. From “Judged frequency of lethal events” [Reproduced from [2]. © American Psychological Association, 1978.]

the person’s risk judgments is imperfect (i.e., $r < 1$), regression to the mean must occur.

In summary, regression to the mean is one explanation for people’s “accurate and biased” judgments of risk frequencies. Although floor effects can explain the overestimation of rare risks, ceiling effects cannot explain the underestimation of frequent ones. The imperfect correlation between objective and judged risk frequencies, thus, best explains regression to the mean.

Another explanation for people’s biased and accurate risk perceptions ([2]; Figure 1) is referred to as the availability heuristic [2]. One might judge the frequency of a class or the probability of an event by the ease with which examples or occurrences can be brought to mind. For example, one may assess the probability of a motor vehicle accident by recalling such occurrences among one’s acquaintances, friends, or family members. From an abstract perspective, the availability heuristic is used to answer questions of the following type: How many objects *X* of the type *Y* are there? How common is event *E* under condition *C*? In general, the availability heuristic is quite useful for judging risk frequencies – as supported by the high correlation between objective and judged risk frequencies in Figure 1.

When is a risk cognitively available and when is it not? Or, to put it differently, which factors

influence the availability of risks? Factors such as (a) disproportionate exposure, (b) imaginability, and (c) ease of retrieval have been suggested to determine the availability of risk [2, 9].

Disproportionate exposure means that the newspapers, for instance, report disproportionately more often about spectacular risks, such as tornados, than about unspectacular risks, such as strokes or stomach cancer. Disproportionate media exposure, then, triggers availability, which causes people’s overestimation of risks [2].

Imaginability means that events that are easy to imagine are cognitively more available than events that are difficult to imagine. For instance, most people can better imagine a tornado than stomach cancer. Consequently, a tornado’s higher availability could explain its biased frequency judgment [2].

Ease of retrieval refers to the ease with which relevant instances can be brought to mind, rather than the sheer number of recalled instances that determines risk judgments [9]. Imagine one group of people is asked to write down three things that increase the risk of cancer. Another group is asked to write down eight things. Then both groups are asked how likely it is to die from cancer. Which group will give the higher probability estimate? The availability heuristic would predict that the group that wrote down eight causes would give a higher probability

estimate than the group that listed only three. Yet this is not always the case: whereas it is easy to retrieve three causes, it is difficult to retrieve eight causes. People in the three-causes group may think that because it is so easy to retrieve three causes, cancer seems rather likely. People in the eight-causes group may think that because it is so difficult to retrieve eight causes, cancer seems rather unlikely. Paradoxically, despite more evidence (eight rather than three causes), people give lower risk estimates. Similar effects have been found in other domains [10, 11]. In sum, both regression to the mean and the availability heuristic explain people's "accurate and biased" risk judgments well [5].

Judging Risks for Combined Events

Up to now, we have covered judgments of single rather than combined events. We now turn to risk judgments for combined events, which can be either conjunctive or disjunctive. We start with risk judgments for conjunctive events.

Judging Risks for Conjunctive Events

Conjunction effects have been debated extensively in the context of probabilistic judgments. For illustrative purposes, consider the following experimental task [12]:

A health survey was conducted in a representative sample of adult males in British Columbia of all ages and occupations. Mr. F. was included in the sample. He was selected by chance from the list of participants. Which of the following statements is more probable? (Check one)

Mr. F. has had one or more heart attacks.

Mr. F. has had one or more heart attacks, and he is over 55 years old.

According to probability theory, the conjunction of two events cannot exceed the probability of either single event, that is, $p(A \& B) \leq p(A), p(B)$. Yet 58% of statistically naive respondents did not find the correct answer in this task. Rather, a conjunction error occurred; that is, they thought the conjunction – Mr. F. has had one or more heart attacks and he is over 55 years – to be more likely than the single event.

The reason for this error is attributed to the representativeness heuristic [12]. That is, older people

(in the experimental task those who are older than 55 years) are more likely to have already encountered one or more heart attacks than people who are younger than 56. To put it differently: heart attack seems to be more representative for older people.

What is available in people's minds is the co-occurrence of the two events. As a consequence, when asked to evaluate whether a person has had one or more heart attacks and is also over 55 years, people often overestimate the co-occurrence of the events. Indeed, to evaluate the co-occurrence of the events more accurately, a person has to think of four groups: (a) people who have had one or more heart attacks and are over 55 years, (b) people who have had one or more heart attacks and are younger than 56, (c) people who have never had a heart attack and are over 55 years, and (d) people who have never had a heart attack and are younger than 56 years. People, however, tend to focus on the group in which both events occurred (a) and neglect those groups that fall into one of the other three combinations.

Is it the representativeness heuristic that causes the conjunction fallacy? Evidence has emerged that in natural language, terms such as "and" (in the second statement of the experimental task) are semantically ambiguous [13, 14]. That is, in a fair test, the phrase "and" must have the same meaning in probability theory as it does in natural language. In probability theory "and" refers to an intersection only, but in natural language "and" refers to either a union or an intersection. For instance, the sentence "We invited friends and colleagues to the party" refers to a union, not to an intersection, of friends and colleagues. There is agreement that (a) the semantic ambiguity of "and" is a factor that contributes to conjunction effects, and (b) certain conditions can override the representativeness heuristic [14].

Judging Risks for Disjunctive Events

Conjunctive events refer to the intersection of two or more events. Disjunctive events refer to the union of two or more events. The disjunction of two or more events can be either implicit or explicit. Imagine, for instance, you are asked to judge your likelihood of dying from cancer (implicit disjunction). Suppose further you are asked to judge your likelihood of dying from skin cancer *or* colorectal cancer, *or* stomach cancer (explicit disjunction). Obviously, there are more than three kinds of cancer. Whereas the implicit

disjunction (cancer) covers all kinds of cancers, the explicit disjunction does not. Consequently, normative theory predicts higher likelihood estimates for the implicit than for the explicit disjunction.

Yet this is not the case. Researchers have found that people's likelihood estimates of the implicit disjunction (cancer) are *lower* than the likelihood estimates of the explicit disjunction (skin cancer, etc.). This phenomenon is called *subadditivity* [15], and it has been found for both students and experts [16]. Psychologically, the explicit disjunction reminds people of the different cancers one can die of, making them cognitively more accessible. Because of this cognitive accessibility, the probability estimate is higher for the explicit than the implicit disjunction.

Cognitive accessibility, however, is not the only mechanism that can explain subadditivity. Imagine an experiment in which Group 1 receives the following question:

What is the probability that you will die from skin cancer *or* stomach cancer?

Group 2 receives the following two questions:

What is the probability that you will die of skin cancer?

What is the probability that you will die of stomach cancer?

Whereas Group 1 receives the disjunction, Group 2 receives the constituents. Does the sum of the constituents' probabilities (Group 2) equal the probability judgment of the disjunction (Group 1)? Research indicates subadditivity is at work again. That is, the probability estimate for the disjunction (Group 1) falls short of the sum of the separate probability estimates for the constituents (Group 2; [17]). Accessibility, however, cannot explain subadditivity here, because for both groups the constituents were mentioned explicitly. The mechanism at work must be a different one that is commonly referred to as anchoring and adjustment [18].

Initially, both groups proceed in the same way by separately judging the probability of each kind of cancer. Group 1, in addition, has to combine these probability judgments (anchors), which are, importantly, lower than the overall probability of the disjunction. Participants may not fully detach from these lower probability estimates and arrive at an estimate for the disjunction that is subadditive. That is, anchoring

(initial probability estimates) and insufficient adjustment (sticking to these anchors) may explain subadditivity. In sum, disjunctive events are subadditive, and (a) cognitive accessibility and (b) anchoring and adjustment can explain this phenomenon.

Do People Ignore Base Rates?

An important question is whether people are insensitive to prior probabilities of outcomes – in other words, do they ignore base-rate information? Consider the following experiment [19]:

A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

- (a) 85% of the cabs in the city are Green and 15% are Blue.
- (b) A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

What is the probability that the cab involved in the accident was Blue rather than Green?

Using Bayes' theorem (*see Bayes' Theorem and Updating of Belief*), one can calculate the correct posterior probability $p(H|D)$, where H and $\neg H$ denote the two possible outcomes (i.e., Blue or Green) and D denotes the reliability of the witness (i.e., 80% in the above experiment). If the values are given in the standard probability format (as in the experiment), the posterior probability can be calculated by solving the following equation:

$$\begin{aligned} p(H|D) &= \frac{p(H)p(D|H)}{[p(H)p(D|H) + p(\neg H)p(D|\neg H)]} \\ &= \frac{(0.15)(0.80)}{[(0.15)(0.80) + (0.85)(0.20)]} \end{aligned} \quad (1)$$

The result is $p(H|D) = 0.41$. Hence, despite the report of the witness, the hit-and-run cab is more likely to be Green than Blue. The reason for this result is that the base rate (0.15) is rather low. When asked to judge the probability that the cab involved in the accident was Blue rather than Green (i.e., the posterior probability $p(H|D)$), people typically gave median responses of $p = 0.80$. This probability judgment resembles the reliability of the witness but ignores base-rate information. According

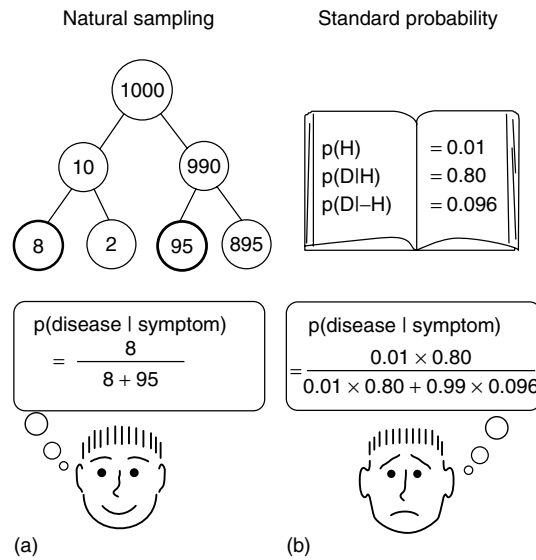


Figure 2 Bayesian inference and information representation. Natural sampling of frequencies (a) and standard probability format (b). From “How to improve Bayesian reasoning without instruction: Frequency formats” [Reproduced from [21]. © American Psychological Association, 1995.]

to the “heuristics-and-biases” program, Bayes’ theorem is no longer expected to describe the workings of the mind, and people’s poor performance has been attributed to the mind’s limited information-processing abilities [20].

However, it was not long before the first experiments were conducted to challenge the view of the mind as a poor Bayesian performer [21, 22]. If frequency information is presented in absolute frequencies rather than in abstract probabilities, people’s judgments often come up to the predictions of Bayes’ theorem. Consider the following task [21]:

Imagine an old, experienced physician in an illiterate society. She has no books or statistical surveys and therefore must rely solely on her experience. Her people have been afflicted by a previously unknown and severe disease. Fortunately, the physician has discovered a symptom that signals the disease, although not with certainty. In her lifetime, she has seen 1000 people, 10 of whom had the disease. Of those 10, 8 showed the symptom; of the 990 not afflicted, 95 did. Now a new patient appears. He has the symptom. What is the probability that he actually has the disease?

Presented in this way, that is, in absolute frequencies rather than in abstract probabilities, Bayes’ theorem reduces to a simple natural sampling tree, as shown

in Figure 2(a). All the physician needs is the number of cases that had both the symptom and the disease (8) and the number of cases that had the symptom but no disease (95) and their sum ($8 + 95$). Presented in the standard probability format, in contrast, Figure 2(b) shows that the calculation of the correct answer is much more intricate.

The findings from a series of experiments clearly show that people can perform Bayesian reasoning if information is presented in a frequency format [21, 23]. Specifically, when presented with a frequency format, 46% of the participants considered base rates and answered in line with the correct Bayesian solution; presented with a standard probability format, this proportion dropped to only 16%. In sum, information representation affects cognitive algorithms. Frequency formats made many participants follow Bayes’ theorem without any teaching or instruction. When information is presented appropriately, people do consider base rates more often.

Are Judgments Drawn from Small Samples Valid?

In the real-world people have to rely on a limited rather than an unlimited number of observations. This

number of observations, that is, the sample size, has been the focus of much research in psychology [1]. Normatively, the laws of statistics prescribe that conclusions drawn from a small sample are less reliable than conclusions drawn from a large sample. Do people consider the size of a sample when they judge the likelihood of events?

The answer seems to be no – at least suggested by one stream of research [24]. Consider the following demonstration: Empirically, in the whole population the mean height of men is 6 ft. From this population, samples of the sizes 10, 100, and 1000 are drawn. Normatively, the mean of a small sample is more likely to deviate from the population mean than the mean of a large sample. Therefore, people are expected to estimate a higher probability of getting a mean height larger than 6 ft (i.e., the population mean) for the small (10) sample than for the large (1000) sample. Yet in one study, in contradiction of this normative rule, participants estimated *equal* probabilities of getting a mean height larger than 6 ft for all three samples sizes (10, 100, and 1000). Obviously, in this experiment they did not consider sample size. In another experiment, squash players answered the following question [25]:

As you know, a game of squash can be played to either 9 or 15 points. Holding all other rules of the game constant, if A is a better player than B, which scoring system will give B a better chance of winning?

All participants had some knowledge of statistics. In contradiction of the normative rule, most participants said that the scoring system should not make any difference. Considering the argument that an atypical outcome is more likely to occur in a small sample than in a large one, it is obvious that the probability of winning is higher for player B when the game is played to 9 rather than 15 points. From this and similar experiments, it has been concluded that people “neglect sample size” and rely on judgments from samples that are too small. Yet, one might wonder whether conclusions from small samples might have some advantages, too.

Indeed, empirical research has found evidence for the benefits of drawing conclusions from small sample sizes [26]. A small sample of the size of 7 ± 2 , Kareev argued, is most likely to produce a sample correlation that is more extreme than the correlation of the population. On the basis of Miller’s

work on the limited number of items people can consider at one time [27], Kareev argued that capacity limitations help people maximize the chances of early detection of strong and useful relations [26]. The question of when and how this benefit occurs is currently an issue of hot debate [28–30].

Judging Personal Risks: Unrealistic or Adaptive?

Imagine you are asked to estimate your personal risk of confronting a certain event in your life, for instance, getting divorced. In particular, you are requested to judge whether your risk of getting divorced is higher, the same, or lower than that of a comparable person. In experiments such as this, normative theory predicts that half of the participants will evaluate their risk to be above a comparable other person, and the other half below. In contrast to this prediction, most participants give lower rather than higher risk estimates. That is, they think that others, but not themselves, will get divorced and thus they show *unrealistic optimism* [3]. Similar results were found for other negative events, such as committing suicide or dying from cancer.

For positive events, in contrast, probability judgments reverse: most people think that they are more likely to experience positive events than comparable other persons. For example, most people think that their chances of owning a home or winning an award for their work are higher than those of comparable others. People thus show unrealistic optimism again. This research suggests that people are poor judges of their future risks and prospective opportunities. Clearly, underestimating risks and overestimating opportunities can lead to severe consequences. People may forgo medical screening tests or unwisely invest money, time, and effort in projects having little chance of success. High expectations, furthermore, can lead to more disappointment, if these expectations are not met. Modest expectations, in fact, are generally believed to enhance one’s happiness. In sum, misperceiving one’s risks and opportunities may have severe consequences.

Does unrealistic optimism indeed impede human functioning? No. A very large body of research shows that optimistic people are better off. Optimistic people are more motivated than realistic people, especially after a decision is made, and optimism rises after decisions [31]. Such postdecisional optimism makes

sense, because questioning one's decision may only undermine one's energy and motivation, which is needed for implementing the decision.

Optimistic people, moreover, achieve better outcomes than realistic people. That is, they are more successful or earn more money than realistic people [13, 32]. Because of their higher expectations, however, they rarely reach their goals. Missing one's goals is generally believed to cause disappointment. Paradoxically, if optimistic people miss their goals, they are *less* disappointed than realistic people [33, 34], because they interpret their failures less negatively than realistic people do. In sum, compared to realistic people, optimistic people (a) are more motivated after a decision, (b) achieve better outcomes, (c) miss their goals more often because of their higher expectations, but (d) paradoxically, are less dissatisfied if their goals have not been met – in short, optimists are unrealistic and adaptive.

Does Irrelevant Information Affect Probability Judgments?

People often estimate risks by starting from an initial value. Usually, this initial value, or anchor, is adjusted to reach a final risk estimate [18]. Anchoring and adjustment describes this process. In one study, for example, participants were asked to estimate the probability of a war between the United States and the Soviet Union [35]. One-third of the participants were asked to judge whether the probability of such a war was smaller or larger than 90%. The second third of the participants were asked to judge whether the probability was smaller or larger than 1%. The remaining third got no anchor. Then, participants estimated the probability of war between the two countries. As expected, participants with an anchor of 90% gave higher probability estimates than participants with no anchor, who gave higher estimates than participants with an anchor of 1%. Similar anchoring effects were found in other domains, too [36–39].

What conditions cause anchoring effects? Research has identified several [40, 41]: scale compatibility, category compatibility, and extremity of the anchor. *Scale compatibility* means that anchoring effects are stronger when the anchor's metric and people's subsequent estimates are on the same scale. In the example above, for instance, both the initial anchors (90% or 1%) and participants' subsequent probability estimates are on the same (probability)

scale. The anchoring effect would be smaller, say, if participants had to judge whether a certain line was shorter or longer than 1 cm (90 cm) and then state their probability estimate for a war between the United States and the Soviet Union.

Category compatibility means that people show stronger anchoring effects when the anchor's metric and the metric of the subsequent estimate belong to the same category than when they do not [38]. Anchoring effects, would be smaller, for example, if people estimated at first whether the probability of an atomic power plant accident was lower or higher than 90% (1%) and then estimated the probability of a war between the United States and the Soviet Union. Note that both estimates belong to the same scale (probability) but not to the same category (atomic power plant accident *versus* war).

Even *extreme anchors* influence people's judgments [38]. Common wisdom tells us that extreme anchors should be less effective than moderate ones, because people would judge them as implausible. In contrast to this intuition, extreme and implausible anchors produced anchoring effects that were just as large as moderate and plausible ones. That is, even extreme anchors such as 99.9999 or 0.0001% would produce anchoring effects in the war example.

Incentives, however, have influenced anchoring effects very little, if at all [18, 39]. That is, in laboratory experiments anchoring effects were not reduced when people's payoffs depended on the accuracy of their final estimates.

The research reviewed above suggests the pessimistic view that irrelevant numerical values (anchors) bias people's probability judgments. Anchors produce assimilation effects, because people's probability judgments tend toward the initial anchor. A different conclusion emerges, however, if we do not consider the estimated probability values but the subjective impressions of these probability values [42]. For the subjective impressions of probability, a contrast rather than an assimilation effect occurs.

Suppose one group of people is asked whether the probability of war is greater than 1%. Another group is asked whether the probability of war is smaller than 90%. Then both groups estimate the probability of war. Because of anchoring and adjustment, the 1% group will give lower probability estimates than the 90% group; thus an assimilation effect has occurred. If asked, however, whether the estimated probability is low or high, a contrast effect occurs: for the

first group, 1% serves as a reference point and any increase is perceived as high; for the second group, 90% serves as a reference point and any decrease is perceived as low. Paradoxically, then, the 1% group will give lower probability estimates than the 90% group, but the 1% group will perceive their estimates to be higher than the 90% group will perceive theirs. These results match other findings that show, for instance, that Olympic athletes who won the bronze medal are happier than athletes who won the silver medal, because bronze medalists compared themselves to athletes who finished fourth, whereas silver medalists compared themselves to the gold medalists who finished first [43]. In sum, anchoring effects can be paradoxical: an assimilation effect in one direction may be compensated for by a contrast effect in the other direction. A clear bias, thus, is not evident.

Conclusion

Probability judgments play a key role in our coping with our daily lives. The common denominator of much research in decision science is the assumption that human probability judgments are often biased. Although this is correct for some specific instances, in this review we presented evidence that this assumption seems to be too narrow. People, for instance, assess the frequencies of causes of death quite accurately ($r = 0.74$), consider base rates and thus follow Bayes' rule if frequency information is presented in absolute frequencies rather than abstract probabilities, or compensate for assimilated frequency judgments (anchoring and adjustment) by forming subjective impression that reverse their biased judgments. In real life, in which the correct normative solutions are most often not known, heuristics are useful tools for making fast and frugal judgments and decisions [21, 44].

Acknowledgments

This research was supported by the Austrian Science Fund (FWF, Österreichischer Forschungsförderungsfond, grant number P18907-G11).

References

- [1] Kahneman, D., Slovic, P. & Tversky A. (eds) (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.
- [2] Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M. & Combs, B. (1978). Judged frequency of lethal events, *Journal of Experimental Psychology: Human Learning and Memory* **4**, 551–578.
- [3] Weinstein, N. (1980). Unrealistic optimism about future live events, *Journal of Personality and Social Psychology* **39**, 806–820.
- [4] Tversky, A. & Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability, *Cognitive Psychology* **5**, 207–232.
- [5] Hertwig, R., Pachur, T. & Kurzenhäuser, S. (2005). Judgments of risk frequencies: tests of possible cognitive mechanisms, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **31**, 621–642.
- [6] Fiedler, K. (2002). Frequency judgment and retrieval structures: splitting, zooming, and merging the units of the empirical world, in *Frequency Processing and Cognition*, P. Sedlmeier & T. Betsch, eds, Oxford University Press, Oxford, pp. 67–87.
- [7] Sedlmeier, P., Hertwig, R. & Gigerenzer, G. (1998). Are judgments of the positional frequencies of letters systematically biased due to availability? *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**, 754–770.
- [8] Furby, L. (1973). Interpreting regression toward the mean in developmental research, *Developmental Psychology* **8**, 172–179.
- [9] Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H. & Simons, A. (1991). Ease of retrieval as information: another look at the availability heuristic, *Journal of Personality and Social Psychology* **61**, 195–202.
- [10] Rothman, A.J. & Schwarz, N. (1998). Constructing perceptions of vulnerability: personal relevance and the use of experiential information in health judgments, *Personality and Social Psychology Bulletin* **24**, 1053–1064.
- [11] Schwarz, N. & Vaughn, L.A. (2002). The availability heuristic revisited: ease of recall and content of recall as distinct sources of information, in *Heuristics and Biases: The Psychology of Intuitive Judgment*, T. Gilovich, D. Griffin & D. Kahneman, eds, Cambridge University Press, Cambridge, pp. 103–119.
- [12] Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment, *Psychological Review* **90**, 293–315.
- [13] Hertwig, R. & Gigerenzer, G. (1999). The “conjunction fallacy” revisited: how intelligent inferences look like reasoning errors, *Journal of Behavioral Decision Making* **12**, 275–305.
- [14] Mellers, B., Hertwig, R. & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration, *Psychological Science* **12**, 269–275.
- [15] Tversky, A. & Koehler, D.J. (1994). Support theory: a nonextensional representation of subjective probability, *Psychological Review* **101**, 547–567.

- [16] Fischhoff, B., Slovic, P. & Lichtenstein, S. (1978). Fault trees: sensitivity of estimated failure probabilities to problem representation, *Journal of Experimental Psychology: Human Perception and Performance* **4**, 330–344.
- [17] Bar-Hillel, M. (1973). On the subjective probability of compound events, *Organizational Behavior and Human Performance* **9**, 396–406.
- [18] Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases, *Science* **185**, 1124–1131.
- [19] Tversky, A. & Kahneman, D. (1982). Evidential impact of base rates, in *Judgment Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic & A. Tversky, eds, Cambridge University Press, Cambridge, pp. 153–160.
- [20] Lichtenstein, S., Fischhoff, B. & Phillips, L.D. (1982). Calibration of probabilities: the state of the art to 1980, in *Judgment Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic & A. Tversky, eds, Cambridge University Press, Cambridge, pp. 306–334.
- [21] Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats, *Psychological Review* **102**, 684–704.
- [22] Kleiter, G.D. (1994). Natural sampling: rationality without base rates, in *Contributions to Mathematical Psychology, Psychometrics, and Methodology*, G.H. Fischer & D. Laming, eds, Springer, New York, pp. 375–388.
- [23] Hoffrage, U., Lindsey, S., Hertwig, R. & Gigerenzer, G. (2000). Communicating statistical information, *Science* **290**, 2261–2262.
- [24] Kahneman, D. & Tversky, A. (1972). Subjective probability: a judgment of representativeness, *Cognitive Psychology* **3**, 430–454.
- [25] Kahneman, D. & Tversky, A. (1982). On the study of statistical intuitions, *Cognition* **11**, 123–141.
- [26] Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations, *Psychological Review* **107**, 397–402.
- [27] Miller, G.A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information, *Psychological Review* **63**, 81–97.
- [28] Anderson, R.B., Doherty, M.E., Berg, N.D. & Friedrich, J.C. (2005). Sample size and the detection of correlation – a signal detection account: comment on Kareev (2000) and Juslin and Olsson (2005), *Psychological Review* **112**, 268–279.
- [29] Juslin, P. & Olsson, H. (2005). Capacity limitations and the detection of correlations: comment on Kareev (2000), *Psychological Review* **112**, 256–267.
- [30] Kareev, Y. (2005). And yet the small-sample effect does hold: reply to Juslin and Olsson (2005) and Anderson, Doherty, Berg, and Friedrich (2005), *Psychological Review* **112**, 280–285.
- [31] Taylor, S.E. & Gollwitzer, P.M. (1995). The effects of mindset on positive illusions, *Journal of Personality and Social Psychology* **69**, 213–226.
- [32] Sherman, S.J., Skov, R.B., Hervitz, E.F. & Stock, C.B. (1981). The effects of explaining hypothetical future events: from possibility to actuality and beyond, *Journal of Experimental Social Psychology* **17**, 142–158.
- [33] Cooper, A.C. & Artz, K.W. (1995). Determinants of satisfaction for entrepreneurs, *Journal of Business Venturing* **10**, 439–457.
- [34] Klaaren, K.J., Hodges, S.D. & Wilson, T.D. (1994). The role of affective expectations in subjective experience and decision making, *Social Cognition* **12**, 77–101.
- [35] Plous, S. (1987). Thinking the unthinkable: the effects of anchoring on likelihood estimates of nuclear war, *Journal of Applied Social Psychology* **19**, 67–91.
- [36] Chapman, G.B. & Johnson, E.J. (1999). Anchoring, activation, and the construction of values, *Organizational Behavior and Human Decision Processes* **79**, 115–153.
- [37] Englich, B. & Mussweiler, T. (2001). Sentencing under uncertainty: anchoring effects in the courtroom, *Journal of Applied Social Psychology* **31**, 1535–1551.
- [38] Strack, F. & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: mechanisms of selective accessibility, *Journal of Personality and Social Psychology* **73**, 437–446.
- [39] Wilson, T.D., Houston, C., Etling, K.M. & Brekke, N. (1996). A new look at anchoring effects: basic anchoring and its antecedents, *Journal of Experimental Psychology: General* **4**, 387–402.
- [40] Chapman, G.B. & Johnson, E.J. (1994). The limits of anchoring, *Journal of Behavioral Decision Making* **7**, 223–242.
- [41] Chapman, G.B. & Johnson, E.J. (2002). Incorporating the irrelevant: anchors and judgments of belief and value, in *Heuristics and Biases*, T. Gilovich, D. Griffin & D. Kahneman, eds, Cambridge University Press, Cambridge, pp. 120–138.
- [42] Teigen, K.H. (2006). How lower and upper limit estimates influence judgments and evaluations, *Paper Presented at the Meeting of the Society for Judgment and Decision Making*, Houston, November.
- [43] Medvec, V.H., Madey, S.F. & Gilovich, T. (1995). When less is more: counterfactual thinking and satisfaction among Olympic medalists, *Journal of Personality and Social Psychology* **69**, 603–610.
- [44] Brandstätter, E., Gigerenzer, G. & Hertwig, R. (2006). The priority heuristic. Making choices without trade-offs, *Psychological Review* **113**, 409–432.

Related Articles

Expert Judgment

Group Decision

Interpretations of Probability

Risk Attitude

EDUARD BRANDSTÄTTER AND RENÉ RIEDL

Benchmark Analysis

Benchmark analysis is a statistical procedure applied to dose–response data from toxicologic, clinical, or epidemiologic studies that can be useful in setting health guidelines for exposures to toxic substances. Such guidelines include those promulgated by the United States Environmental Protection Agency (USEPA) called *reference doses* (RfDs) for oral exposures, or *reference concentrations* (RfCs) for exposures through the ambient air. The USEPA defines the RfD as an estimate (with uncertainty spanning perhaps an order of magnitude) of a daily exposure to the human population, including sensitive subgroups that is likely to be without an appreciable risk of deleterious effects during a lifetime [1].

Previously, an RfD or RfC was based on a no-observed-adverse-effect-level (NOAEL – the largest dose or exposure at which no adverse health response was observed), which is the largest exposure or dose at which no statistically or biologically significant adverse effect has been identified, and which is also smaller than any dose or exposure at which an adverse effect has been noted. Once a NOAEL is determined, the RfC or RfD is calculated by converting the NOAEL to an equivalent human exposure, and then dividing by uncertainty factors that account for various types of shortcomings in the data base.

There are several potential shortcomings to this NOAEL approach to setting exposure guidelines:

1. The shape of the dose–response curve (*see Dose–Response Analysis*) is not taken into account in determining a NOAEL. Also, the NOAEL must be one of the experimental doses. In an experiment with a steep dose–response, in which a marginally significant response occurred at a particular dose, it may appear highly likely that no response would have been seen at, e.g., a twofold lower dose. However, the NOAEL would be defined as the next lower experimental dose, which could be 10-fold or more lower.
2. All other things being equal, a smaller study should result in a lower exposure guideline because it entails more uncertainty. However, the NOAEL from a smaller study tends to be larger

because a smaller sample size makes statistical significance less likely to be established.

3. Application of the NOAEL approach is problematic in cases in which an adverse response occurred at the smallest experimental dose, so that a NOAEL is not determined.
4. Deciding whether a particular dose is a NOAEL can be controversial. For example, for the data set in Table 1, the response at 10 ppm (9/50) is not statistically different from that in unexposed subjects (5/50) ($p = 0.19$), so that 10 ppm could be considered to be the NOAEL. On the other hand, since the response at 10 ppm is higher than that in unexposed subjects, and appears to be consistent with the dose–response suggested by the increased responses at 15 and 20 ppm, 5 ppm could alternatively be considered to be the NOAEL. How such controversy is resolved can have a large effect upon an exposure guideline.
5. The NOAEL approach is most readily applicable to data from toxicological or clinical studies (*see Toxicity (Adverse Events)*) in which test subjects are distributed into groups whose subjects all have a common exposures. It is not straightforward to apply the approach to epidemiological data, where each subject can have a unique exposure. Also, in some epidemiological studies there is no unexposed group to compare with exposed groups.

The benchmark approach was proposed as an alternative to the NOAEL approach in setting exposure guidelines [2, 3]. The benchmark dose (BMD – defined as the dose that causes a specified increase in the risk, or probability, of an adverse response) is the dose or exposure that corresponds to a specified increase in risk (called the *benchmark risk* or *BMR*, level of increased risk (e.g., 0.1) used in the definition of the BMD) over the risk in an unexposed population. It is estimated by fitting a mathematical dose–response model to data. A statistical lower bound on the BMD (BMDL – statistical lower bound on the BMD) replaces the NOAEL in determining an exposure guideline. The USEPA uses the BMD method for determining RfDs and RfCs for noncancer responses [1]. Also, the USEPA guidelines for cancer risk assessment specify the determination of an AED₁₀, which is equivalent to a BMD corresponding to a BMR of 0.1 [4].

Table 1 Example of BMD calculation from binary data^(a)

Dose (ppm)	Number of animals	Number of responders		
		Observed	Predicted	
			Weibull	Logistic
0	50	5	5.26	4.89
5	50	4	5.29	5.12
10	50	9	6.56	7.12
15	50	15	16.5	16.1
20	50	45	39.6	39.7
BMD ^(b)	—		12.8	12.4
BMDL ^(c)	—		10.1	9.7

^(a) Reproduced from [2]. © John Wiley & Sons, Ltd, 1984

^(b) Maximum – likelihood estimate of BMD using $BMR = 0.1$

^(c) Ninety – five percent of statistical lower bound on BMD computed by the profile likelihood method [5–7]

The BMD approach addresses the shortcomings of the NOAEL mentioned above:

1. The BMDL takes into account the shape of the dose–response through the fitted dose–response model, and the BMDL is not constrained to be one of the experimental doses.
2. Because the BMDL accounts for statistical variation, smaller studies tend to result in smaller BMDLs.
3. The BMDL can be calculated from a study in which an effect was seen at all doses.
4. The determination of a BMDL does not involve the potentially controversial decision of determining whether an effect was present at a particular dose.
5. The BMD approach can be applied to either grouped or ungrouped data, and does not require an unexposed group.

On the other hand, the BMD approach involves issues that are not present in the NOAEL approach:

1. The BMDL may depend upon the particular mathematical dose–response model used.
2. The level of risk (BMR) used to define the BMD must be decided upon.
3. Different modeling approaches, and perhaps different definitions of risk (*see Absolute Risk Reduction*), are required for continuous responses (such as blood pressure which can theoretically assume any value in a range) than for binary

responses (which indicate only whether or not a particular effect was present).

The BMD is defined as a dose that corresponds to a specified increase in risk, whereas the NOAEL is sometimes viewed as a dose that truly had no effect, rather than a dose at which no effect was identified (leaving open the possibility that an effect was present that a study with greater statistical power might have detected). For persons with this latter view of the NOAEL, replacing the NOAEL with the BMDL represents a major paradigm shift.

Performing a BMD Analysis Using Binary Data

Let $P(d)$ represent the probability that a subject exposed to a dose of d has an adverse health outcome. We assume $P(d)$ has a mathematical form that involves unknown parameters that are estimated by fitting the dose–response model, $P(d)$, to data. The fitting of the model to data is generally accomplished using the method of maximum likelihood, in which the unknown parameters are selected so that the probability of observing the experimental outcome is maximized [5].

The BMD can be defined as the dose for which the additional risk equals the predetermined BMR, i.e., as the solution to the equation

$$P(BMD) - P(0) = BMR \quad (1)$$

The maximum-likelihood estimate of the BMD is calculated by replacing the unknown parameters by their maximum-likelihood estimates, and solving this equation for BMD. The profile likelihood method is recommended for calculating a statistical lower bound, BMDL, on the BMD [5–7].

Dose–Response Models for Binary Data

A number of different dose–response models have been proposed for use in BMD calculations [8, 9]. Examples include the Weibull model,

$$P(d) = 1 - \exp[-\alpha - (\beta d)^K], \quad \alpha, \beta \geq 0, K \geq 1 \quad (2)$$

and the logistic model,

$$P(d) = 1 - \frac{1}{[1 + \alpha + (\beta d)^K]}, \quad \alpha, \beta \geq 0, K \geq 1 \quad (3)$$

where α , β , and K are the parameters estimated by fitting the model to data [5, 6, 8, 9]. In each model, α specifies the background response, β gauges the potency of the substance for causing the health effect, and K determines the shape of the dose–response. Both of these models can be expanded to include covariates such as age and gender.

The inclusion of the shape parameter, K , permits these models to be nonlinear. The restriction, $K \geq 1$, is recommended with each model in order to rule out biologically implausible dose–responses that hook up very sharply at very low doses, and which can result in BMDLs that are unrealistically small [10].

Table 1 presents the results of a BMD analysis that applied the Weibull and logistic models (*see Logistic Regression*) to some fabricated toxicological data. The BMR was set equal to 0.1. Although 15 was the lowest dose at which the response was significantly elevated over the response in controls, because the response at a dose of 10 appears to be part of an increasing trend, identification of a NOAEL could be controversial. Both the Weibull and logistic models fit these data adequately. (A χ^2 goodness-of-fit test resulted in a p value >0.4 in both cases.) The BMDLs obtained from the two models were also similar (10.1 using Weibull model and 9.7 using the logistic model).

Figure 1 illustrates the calculation of the BMD maximum-likelihood estimate and the BMDL for the

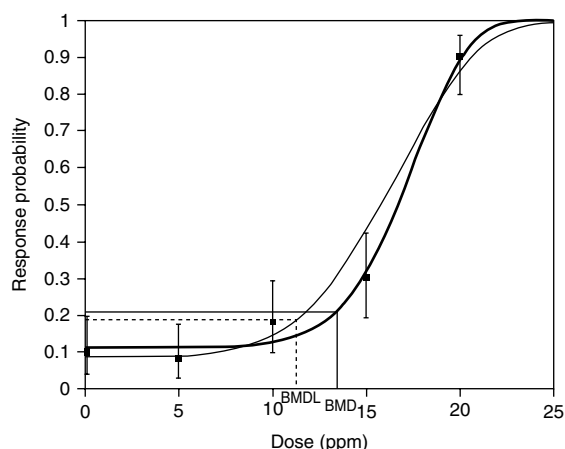


Figure 1 Illustration of the calculation of the BMD from the data in Table 1, using the Weibull model (equation (2))

Weibull model applied to the data in Table 1. The vertical bars in the figure represent 90% confidence intervals for the response probabilities. The darker curve is the maximum-likelihood fit of the Weibull model, $P(d)$. The parameter values obtained from this fit were $\alpha = 0.12$, $\beta = 0.056$, and $K = 7.3$. The BMD estimate of 13.5 ppm is defined by where the upper horizontal line, placed at a response of $P(0) + BMR = 0.11 + 0.1 = 0.21$, intersects the curve $P(d)$. The lighter curve is the Weibull model, $P^*(d)$, that defines the BMDL (i.e., the Weibull curve determined by parameters corresponding to the BMDL derived using the profile likelihood method [5–7]). The $BMDL = 11.3$ ppm is defined by where the lower (dotted) horizontal line, placed at a response of $P^*(0) + BMR = 0.086 + 0.1 = 0.186$, intersects the curve $P^*(d)$.

Selection of the Risk Level (BMR)

Before the BMD method can be applied, the BMR (level of additional risk resulting from exposure at the BMD) must first be specified. In the example presented above, the BMR was set at 0.1, in keeping with an original proposal by Crump that the BMR be in the range 0.01–0.1 [2]. It is important not to make the BMR so small that the BMD is highly model-dependent. The limited size of most toxicological experiments (50 or fewer animals per group) makes it difficult to reliably estimate doses corresponding to risks as low as 0.01, which suggests that a BMR in

the upper end of this range may be more appropriate. Many of the RfCs and RfDs promulgated by the USEPA based on the BMD have utilized $BMR = 0.1$ [1]. Similarly, the USEPA guidelines for cancer risk assessment incorporate calculation of an ED_{10} , which is equivalent to a BMD corresponding to a $BMR = 0.1$ [4]. For different BMD analyses to be comparable, they should be based on a common value for the BMR.

Model Selection

In addition to the Weibull and logistic models discussed herein, many others could be considered. Models for binary data with no more than three parameters (one for each of background, shape, and potency) will often provide comparable BMDs, so long as the BMR is in the range of risks that can be estimated reasonably well by the data. However, this may not be true for data sets that are sparse or do not demonstrate a strong dose–response. A model that fits the data adequately may be preferred over one that does not, although often several models will provide comparable fits. Because of this, in order to standardize the determination of an exposure guideline, it could be useful to specify a default model to be used so long as it fits the data adequately. The RfCs and RfDs promulgated by the USEPA have often been based upon the Weibull model or the special case of the Weibull with $K = 1$ [1].

Performing a BMD Analysis Using Continuous Data

A simple way to calculate a BMD from continuous data is to first convert the data to a binary form by defining a response to be abnormal if it is more extreme than some predetermined limit, y_0 , and then applying the methods discussed above for binary responses. However, possibly more powerful approaches may be obtained by directly modeling the continuous responses. Such modeling is often carried out using a model that assumes that the (possibly transformed) continuous response is normally distributed, with the mean, and possibly the standard deviation, depending upon the dose. Using such a model, perhaps the simplest approach is to define the BMD as the dose corresponding either to a specified absolute or percentage change in the mean response.

However, this approach does not account for normal variation in a response. For example, a 10% change in red blood cell count is within the range of normal variation, whereas a 10% change in temperature would be fatal. A better way to account for normal variation is to define the BMD as the dose that corresponds to a predetermined change in the mean response normalized by the standard deviation in unexposed subjects [11].

Because both binary and continuous data are often available, it would be helpful to have a common definition of the BMD that can be applied to both types of data. Otherwise, it may not be clear how BMDs determined from the two types of data are related. Suppose that responses larger than a predetermined limit, y_0 , are considered to be adverse. For example, if the response were systolic blood pressure and y_0 was selected as 160, then a systolic blood pressure greater than 160 would be considered to be an adverse response. Alternatively, instead of assuming a fixed limit, y_0 , one could specify a fixed probability, $P(0)$, of an adverse response in an unexposed individual (e.g., $P(0) = 0.05$ in keeping with the convention in clinical testing that the normal range includes 95% of the responses of healthy individuals) of a unexposed subject having an adverse response. With either y_0 or $P(0)$ fixed, the probability of an adverse response in a subject with a dose d can then be calculated using a model for the continuous response suggested above without the need to convert the data to binary form [10–12]. Using this form for the probability of an adverse effect, the BMD can then be defined using equation (1) in a manner analogous to that used for binary data.

Software for BMD Analysis

The calculations needed to implement a benchmark analysis are nonstandard and may be difficult to perform using standard statistical software. Several special-purpose software packages are available for performing benchmark calculations [8, 9, 13].

References

- [1] United States Environmental Protection Agency (USEPA) (2006). *Integrated Risk Information Service (IRIS)*, <http://www.epa.gov/iris> (accessed Aug 2006).

- [2] Crump, K.S. (1984). A new method for determining allowable daily intakes, *Fundamental and Applied Toxicology* **4**, 854–871.
- [3] American Industrial Health Council (1993). Evaluating the benchmark dose methodology: AIHC/EPA workshop, *AIHC Journal* **1**, 10–11.
- [4] United States Environmental Protection Agency (USEPA) (2005). *Guidelines for Carcinogen Risk Assessment. Risk Assessment Forum*, U.S. Environmental Protection Agency, Washington, DC, EPA/630/P-03/001F.
- [5] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*, Chapman & Hall, London.
- [6] Crump, K.S. (2002). Benchmark analysis, in *Encyclopedia of Environmetrics*, John Wiley & Sons, West Sussex, pp. 163–170.
- [7] Venzon, D. & Moolgavkar, S. (1988). A method for computing profile-likelihood-based confidence intervals, *Applied Statistics* **37**, 87–94.
- [8] United States Environmental Protection Agency (USEPA) (1998). *Benchmark Dose Software Draft Beta Version 1.1b*, Office of Research and Development, National Center for Environmental Assessment.
- [9] Howe, R.B. (1984). *THRESHW – A Computer Program to Compute a Reference Dose from Quantal Animal Toxicity Data using the Benchmark Dose Method*, The K. S. Crump Group, Ruston.
- [10] Crump, K.S. (2002). Critical issues in benchmark calculations from continuous data, *Critical Reviews in Toxicology* **32**, 133–153.
- [11] Crump, K.S. (1995). Calculation of benchmark doses from continuous data, *Risk Analysis* **15**, 79–89.
- [12] Gaylor, D.W. & Slikker, W. (1990). Risk assessment for neurotoxic effects, *Neurotoxicology* **11**, 211–218.
- [13] Crump, K.S. & Van Landingham, C. (1995). *Bench-C: A Fortran Program to Calculate Benchmark Doses from Continuous Data*, The K. S. Crump Group, Ruston.

KENNY S. CRUMP

Benchmark Dose Estimation

Introduction and Background

A primary objective of quantitative risk analysis is the characterization of risk *via* one or more quantitative estimates of the severity and likelihood of damage to humans or to the environment caused by a hazardous agent [1]. In many cases, however, epidemiological or human data on hazardous substances (*see What are Hazardous Materials?*) are not available or are inadequate for quantitative risk assessment. Thus, to assess the potential risk of an exposure to a hazardous agent, bioassays are conducted on laboratory rodents or other biological systems where the dose levels of the substance(s) are administered at relatively high dose levels (*see Cancer Risk Evaluation from Animal Studies*). A major quantitative component of such studies is statistical characterization of the *dose–response* (*see Dose–Response Analysis*) to the toxic agent and, from this, assessing possible risks to humans and setting acceptable reference doses (RfDs) or reference concentrations (RfCs) *via* extrapolation of the results of such bioassays to low-dose ranges usually encountered in environmental exposures.

Traditionally, risk and RfD estimation has been based heavily on a hypothesis-testing procedure where a no-observed-adverse-effect level (NOAEL) and a lowest-observed-adverse-effect level (LOAEL) are generated using multiple comparisons between the control group and the various dose groups. First, the largest experimental dose that is not significantly different from the zero-dose control group (NOAEL) is determined. If the NOAEL is not well defined, the smallest experimental dose that is significantly different from the zero-dose control (LOAEL) is determined. The NOAEL (LOAEL) is considered as an estimator of the dose threshold of the hazard. Next, if the NOAEL is well defined, the RfD is often defined as NOAEL/100. Here the 100 in the denominator of the RfD is called an *uncertainty* or *safety factor* (*SF*) that accounts for interspecies extrapolation of experimental results from animals to humans. To be more specific, 100 is taken as a factor of 10 for animal to human extrapolation, and then a factor of 10 for individual variation in sensitivity (=100).

In the absence of a NOAEL, the RfD is defined as LOAEL/1000, where another factor of 10 is used in the denominator since a NOAEL is not available. The RfD is then used in the regulatory process for setting acceptable levels of human exposure to potentially toxic substances.

Although the NOAEL/SF approach is easy to understand and has an intuitive appeal, it suffers many drawbacks. Examples of such are as follows: the NOAEL must be one of the experimental doses and is known to decrease as the number of replicates increases; it can fail to estimate the potential risk at the RfD, which is usually greater than zero; information on the dose–response relationship is lost since this approach only gives one single summary measure and does not utilize all of the data; it does not take into account the slope of the dose–response curve nor the variability in the estimate of the dose–response; and it is not straightforward to apply to epidemiological data [2–4].

To overcome some of the drawbacks of the NOAEL/LOAEL approach, Crump devised what is now known as *benchmark dose methodology* [2, 5]. In a risk assessment the benchmark dose (BMD) is defined as the point estimate of the effective dose/concentration that would result in a predetermined increase in risk (known as the *benchmark risk* or *BMR*) over the risk in an unexposed population. A statistical lower confidence limit on the BMD is known as the a lower limit on the benchmark dose (*BMDL*) [2]. In 1995, the United States Environmental Protection Agency (USEPA) adopted the BMD approach in setting RfDs and recommends that target BMRs should be between 1 and 10%. In fact, USEPA has developed a general technical guidance and a computer package, known as *BMDS*, to facilitate the application of the BMD method in risk assessment [4, 6].

To apply the BMD methodology, an appropriate statistical model that quantifies, at least approximately, the relationship between exposure level (dose) and response is fit to the experimental data. From this model, we calculate the dose (BMD) that corresponds to the specified risk (BMR) and a lower $100(1 - \alpha)\%$ confidence limit (BMDL) is derived. Defining risk as the probability of some predefined adverse outcome (such as death, weight loss, birth defect, cancer, or mutation) exhibited in a subject exposed the hazardous agent and assuming risk changes with increasing dose or exposure, d , to

the toxic agent, we write this probability as $R(d)$. To correct for the spontaneous risk of response, $R(0)$, additional risk and extra risk functions are frequently used in risk estimation [7–10]. Additional risk, $R_A(d)$, is defined to be the risk beyond that of the control (background, spontaneous) level; that is

$$R_A(d) = R(d) - R(0) \quad (1)$$

In some instances, additional risk is further corrected into extra risk to account for nonresponse in the unexposed population. Extra risk, $R_E(d)$, is defined as the additional risk among subjects who would not have responded under control conditions:

$$R_E(d) = \frac{R_A(d)}{1 - R(0)} \quad (2)$$

Notice that equation (2) assumes $0 \leq R(0) < 1$.

The BMD for Quantal Response Data

With quantal data, a subject is classified as exhibiting an adverse health effect or not and the proportion of subjects exhibiting the effect is recorded. These data are very common in carcinogenesis, teratogenesis, and mutagenesis studies [7, 11–13]. The usual assumption under this setting is that the number of subjects exhibiting an adverse response is binomially distributed. Formally, let $d_1, d_2, \dots, d_k$ be the k dose groups employed, $N_1, N_2, \dots, N_k; n_1, n_2, \dots, n_k$ be the total and responding subjects in the i th dose group, respectively, and $p_i = p(d_i, \Theta)$, $i = 1, 2, \dots, k$ the probability of response to dose i where Θ is a vector of parameters in the posited dose–response model. To find the maximum-likelihood estimator, $\hat{\Theta}$ of Θ , we can maximize the log-likelihood function $L(\Theta) = \sum_{i=1}^k (n_i \ln(p_i) + (N_i - n_i) \ln(1 - p_i))$. Then for a specified $BMR = R^*$, and using extra risk to account for the risk beyond background risk, the estimated BMD ($\hat{BMD}$) is obtained by solving the following equation:

$$R^* = \frac{p(BMD; \hat{\Theta}) - p(0; \hat{\Theta})}{1 - p(0; \hat{\Theta})} \quad (3)$$

Alternatively, if additional risk is used, the $B\hat{M}D$ is the solution to the equation:

$$R^* = p(BMD; \hat{\Theta}) - p(0; \hat{\Theta}) \quad (4)$$

Note that since $0 < 1 - p(0; \hat{\Theta}) < 1$, the $B\hat{M}D$ obtained from equation (3) is at least as large as that obtained from equation (4).

Various methods (e.g., bootstrap, delta method, and likelihood ratio) exist for calculating the BMDL. Of these, the likelihood-ratio method is recommended as it generally gives better coverage results [12, 14]. Using the likelihood-ratio approach, the BMDL is the minimum BMD such that

$$2[L_{\max} - L(BMD)] \leq \chi^2_{1-(\alpha/2);1} \quad (5)$$

where $L_{\max}$ is the maximum value of $L(\hat{\Theta})$ and $\chi^2_{1-(\alpha/2);1}$ is the 100 $(1 - \frac{\alpha}{2})\%$ percentage point of the χ^2 distribution with 1 degree of freedom.

Many dose–response models exist that can be used to model the risk at a specific dose for quantal data. (For detailed descriptions of some of the popular models see, for example, [6, 15].) Examples of such models include the logistic model (*see Logistic Regression*)

$$p(d) = [1 + \exp(-\alpha - \theta d)]^{-1} \quad (6)$$

and the Weibull model,

$$p(d) = \alpha + (1 - \alpha)[1 - \exp(-\theta d^\alpha)] \quad (7)$$

In many cases, more than one model adequately fits the data at hand. In such cases, the model with best fit should be employed [16]. Besides the adequacy of the model, biological plausibility, consistency between the estimated model and the observed variables, the Akaike’s information criterion (AIC), the magnitude of the difference between the BMD and BMDL are all factors to be considered when choosing the model [8].

In many developmental and reproductive toxicology studies, the binomial model is inappropriate because of the possible correlation of responses between fetuses in the same litter (i.e., litter effect). To incorporate the extra-binomial variation owing to this added litter effect, a generalized binomial model such as the beta binomial is used to model risk [17]. Once

the log-likelihood is maximized, equations (3–5) are used to obtain the BMD and BMDL.

The BMD for Continuous Response Data

When the endpoint of interest is measured on a continuous scale (e.g., altered body weight), it is difficult to define and observe an adverse effect unambiguously [18]. One way to deal with such situations is to consider a continuous response that exceeds a specified cutoff to be abnormal. This dichotomization, in effect, induces a quantal response from continuous data and the methodology outlined above for quantal data can be applied to obtain the BMD/BMDL. Although this is a simple approach, it does not make optimal use of the data. Furthermore, the data may be available only in summary form (e.g., mean response) and such dichotomization cannot be achieved [2]. Another more practical approach is to assume that response, y_{ij} , of the j th subject at dose d_i is normally (or log-normally [19]) distributed, $i = 1, 2, \dots, g; j = 1, 2, \dots, N_i$. Then, under the normal distribution and assuming a constant variance among the g dose groups, the log-likelihood function has the form:

$$L = -\frac{g}{2} \ln(2\pi) - \sum_{i=1}^g \left[\frac{N_i}{2} \ln \sigma^2 + \frac{(N_i - 1)s_i^2}{2\sigma^2} + \frac{N_i(\bar{y}_i - \mu(d_i))^2}{2\sigma^2} \right] \quad (8)$$

where $\mu(d_i)$ is the observed mean for the i th dose group from the assumed model s_i^2 and $\bar{y}_i$ are the usual sample variance and mean, respectively, of the i th dose group. When the variance is not constant among the different dose groups and is allowed to be a power function of the mean i.e., $\sigma_i^2 = \alpha[\mu(d_i)]^\rho$, the log-likelihood takes the form:

$$L = -\sum_{i=1}^g \left[\frac{N_i}{2} \ln \alpha + \frac{N_i \rho}{2} \ln[\mu(d_i)] + H_i \right] \quad (9)$$

where

$$H_i = \frac{A_i}{2\alpha[\mu(d_i)]^\rho} + \frac{B_i}{2\alpha[\mu(d_i)]^{\rho-1}} + \frac{N_i}{2\alpha[\mu(d_i)]^{\rho-2}} \quad (10)$$

with

$$A_i = (N_i - 1)s_i^2 + N_i\bar{y}_i^2 \quad \text{and} \quad B_i = N_i\bar{y}_i \quad (11)$$

For continuous dose–response models, the BMD is defined to be the dose that results in a predefined change in the mean response. The change can be any of the following:

1. An absolute change in the mean response. In this case, the estimated BMD ($\hat{BMD}$) is obtained by solving the equation $|\mu(BMD) - \mu(0)| = R^*$, where R^* is the specified BMR.
2. A change in the mean equal to a specified number of control standard deviations and the $\hat{BMD}$ is attained as the solution to the equation $|\mu(BMD) - \mu(0)| = R^*\hat{\sigma}_0$.
3. A specified fraction of the control group mean and the $\hat{BMD}$ is the solution to the equation $|\mu(BMD) - \mu(0)| = R^*\mu(0)$.
4. A specified value of the mean at the BMD and the $\hat{BMD}$ is the solution to the equation $\mu(BMD) = R^*$.
5. A change equal to a specified fraction of the range of the responses and the $\hat{BMD}$ is the solution to the equation $\frac{\mu(BMD) - \mu(0)}{\mu_{\max} - \mu(0)} = R^*$. Here again, the BMDL can be calculated by appealing to equation (5).

As in the quantal case, several statistical dose–response models exist to model risk for continuous endpoints (see, for example [6, 15]).

More recently, Pan *et al.* devised simultaneous confidence bounds that have applications in multiple inferences for low-dose risk estimation [20]. The developed methodology can be used to construct simultaneous, one-sided, upper confidence limits on risk for quantal or continuous endpoints. By inverting the upper confidence limits on risk, BMDLs can be obtained for any finite number of prespecified BMRs. Throughout the effort, valid $100(1 - \alpha)\%$ inferences can be made by the underlying simultaneity of the method. The simultaneous approach was later applied to quantal and continuous data [7, 21]. Nitcheva *et al.* adjust the simultaneous limits for multiplicity by applying a Bonferroni adjustment for quantal data and Wu *et al.* adjust for the multiplicity problem for continuous response data [13, 22]. Another recent development is the application of the BMD method *via* bootstrapping to longitudinal data arising in neurotoxicity studies [23].

References

- [1] Stern, A. (2002). Risk assessment, quantitative, in *Encyclopedia of Environmetrics*, 1st Edition, A.H. El-Shaarawi & W.W. Piegorsch, eds, John Wiley & Sons, Chichester, pp. 1837–1843.
- [2] Crump, K.S. (1995). Calculation of benchmark doses from continuous data, *Risk Analysis* **15**, 79–89.
- [3] Gaylor, D.W. (1996). Quantalization of continuous data for benchmark dose estimation, *Regulatory Toxicology and Pharmacology* **24**, 246–250.
- [4] U.S. EPA (2000). *Benchmark Dose Technical Guidance Document: External Review*, U.S. Environmental Protection Agency, Washington, DC, <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=22506>, Draft number EPA/630/R-00/001.
- [5] Crump, K.S. (1984). A new method for determining allowable daily intakes, *Fundamental and Applied Toxicology* **4**, 854–871.
- [6] U.S. EPA (2001). *Help Manual for Benchmark Dose Software Version 1.3*, Technical Report number EPA/600/R-00/014F, National Center for Environmental Assessment. U.S. Environmental Protection Agency, Research Triangle Park.
- [7] Al-Saidy, O.M., Piegorsch, W.W., West, R.W. & Nitcheva, D.K. (2003). Confidence bands for low-dose risk estimation with quantal response data, *Biometrics* **59**, 1056–1062.
- [8] DPR MT-1 (2004). *Guidance for Bench Mark Dose (BMD) Approach – Quantal Data*, Medical Toxicology Branch, Department of Pesticide Regulations, Sacramento.
- [9] Gaylor, D.W., Ryan, L., Krewski, D. & Zhu, Y.L. (1998). Procedures for calculating benchmark doses for health risk assessment, *Regulatory Toxicology and Pharmacology* **28**, 150–164.
- [10] Gephart, L.A., Salminen, W.F., Nicolich, M.J. & Pelekis, M. (2001). Evaluation of subchronic toxicity using the benchmark dose approach, *Regulatory Toxicology and Pharmacology* **33**, 37–59.
- [11] Bailer, A.J., Noble, R.B. & Wheeler, M.W. (2005). Model uncertainty and risk estimation for experimental studies of quantal responses, *Risk Analysis* **25**, 291–299.
- [12] Crump, K.S. & Howe, R. (1985). A review of methods for calculating confidence limits in low dose extrapolation, in *Toxicological Risk Assessment, Biological and Statistical Criteria*, D.B. Clayson, D. Krewski & I. Munro, eds, CRC Press, Boca Raton, Vol. 1, pp. 187–203.
- [13] Nitcheva, D.K., Piegorsch, W.W., West, R.W. & Kodell, R.L. (2005). Multiplicity-adjusted inferences in risk assessment: benchmark analysis with quantal response data, *Biometrics* **61**(1), 277–286.
- [14] Moerbeek, M., Piersma, A.H. & Slob, W. (2004). A comparison of three methods for calculating confidence intervals for the benchmark dose, *Risk Analysis* **24**, 31–40.
- [15] Wheeler, M.W. (2005). Benchmark dose estimation using SAS®, in *Proceedings of the Thirtieth Annual SAS® Users Group International Conference*, G.S. Nelson, eds, SAS Institute, Cary, pp. 201–230, <http://www2.sas.com/proceedings/sugi30/201-30.pdf>.
- [16] Crump, K.S. (2002). Benchmark Analysis, in *Encyclopedia of Environmetrics*, 1st Edition, A.H. El-Shaarawi & W.W. Piegorsch, eds, John Wiley & Sons, Chichester, pp. 163–170.
- [17] Chen, J.J. & Kodell, R.L. (1989). Quantitative risk assessment for teratological effects, *Journal of the American Statistical Association* **84**, 966–971.
- [18] Kodell, R.L. & West, R.W. (1993). Upper confidence intervals on excess risk for quantitative responses, *Risk Analysis* **13**, 177–182.
- [19] Gaylor, D.W. & Slikker, W.L. (1990). Risk assessment for neurotoxic effects, *Neurotoxicology* **11**, 211–218.
- [20] Pan, W., Piegorsch, W.W. & West, R.W. (2003). Exact one-sided simultaneous confidence bands via Uusi-paikka's method, *Annals of the Institute of Statistical Mathematics* **55**, 243–250.
- [21] Piegorsch, W.W., West, R.W., Pan, W. & Kodell, R.L. (2005). Low-dose risk estimation via simultaneous statistical inferences, *Journal of the Royal Statistical Society, series C (Applied Statistics)* **54**, 245–258.
- [22] Wu, Y., Piegorsch, W.W., West, R.W., Tang, D., Petkewich, M.O. & Pan, W. (2006). Multiplicity-adjusted inferences in risk assessment: benchmark analysis with continuous response data, *Environmental and Ecological Statistics* **13**, 125–141.
- [23] Zhu, Y., Jia, Z., Wang, W., Gift, J.S., Moser, V.C. & Pierre-Louis, B.J. (2005). Analyses of neurobehavioral screening data: benchmark dose estimation, *Regulatory Toxicology and Pharmacology* **42**, 190–201.

Further Reading

- Allen, B.C., Kavlock, R.J., Kimmel, C.A. & Faustman, E.M. (1994). Dose-response assessment for developmental toxicity. II. Comparison of generic benchmark dose estimates with no observed adverse effect levels, *Fundamental and Applied Toxicology* **23**, 487–495.
- DPR MT-2 (2004). *Guidance for Benchmark Dose (BMD) Approach – Continuous Data*, Medical Toxicology Branch, Department of Pesticide Regulations, California Environmental Protection Agency, Sacramento.
- Gaylor, D.W. & Chen, J.J. (1996). Precision of benchmark dose estimates for continuous (nonquantal) measurements of toxic effects, *Regulatory Toxicology and Pharmacology* **24**, 19–23.
- Morales, K.H., Ibrahim, J.G., Chen, C.-J. & Ryan, L.M. (2006). Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water, *Journal of the American Statistical Association* **101**, 9–17.

- Morales, K.H. & Ryan, L.M. (2005). Benchmark dose estimation based on epidemiologic cohort data, *Environmetrics* **16**, 435–447.
- Piegorsch, W.W., Nitcheva, D.K. & West, R.W. (2006). Excess risk estimation under multistage model misspecification, *Journal of Statistical Computation and Simulation* **76**, 423–430.
- Piegorsch, W.W. & West, R.W. (2005). Benchmark analysis: shopping with proper confidence, *Risk Analysis* **25**, 913–920.
- Piegorsch, W.W., West, R.W., Pan, W. & Kodell, R.L. (2005). Simultaneous confidence bounds for low-dose risk assessment with nonquantal data, *Journal of Biopharmaceutical Statistics* **15**, 17–31.

Related Articles

Air Pollution Risk

Detection Limits

Environmental Risk Assessment of Water Pollution

Low-Dose Extrapolation

OBAID M. AL-SAIDY

Benzene

The intent of this article is to familiarize the reader with the background of regulatory risk assessments of benzene and the epidemiological, exposure, medical, and biochemical issues of this compound, which are important in the risk assessment process, are in many instances unresolved. Benzene is possibly the most highly regulated, highly studied, and recognized human carcinogen in the history of toxicology. This classification has been formalized by organizations such as IARC (International Agency for Research on Cancer), the USEPA (United States Environmental Protection Agency), and the EU (European Union). It is virtually ubiquitous in most terrestrial environments [1] due to its origin in both anthropogenic sources such as motor gasoline as well as its exhaust and cigarette smoke, and naturally occurring sources such as wildfires and volcanism.

Physical/Chemical Properties

Benzene, molecular formula C_6H_6 and molecular weight 78.1, is the simplest of the aromatic hydrocarbons. Benzene is quite volatile with a boiling point of $80.1^\circ C$ and vapor pressure at $26^\circ C$ of 100 mmHg. Liquid benzene in the environment or workplace will evaporate readily, and, therefore, inhalation is a major potential exposure pathway. Benzene is described as having a pleasant, sweet aromatic odor. Although there is a wide range of reported values, the most commonly endorsed odor threshold is 1.5–8 ppm [2]. The liquid is lighter than water (0.874 g ml^{-1} at 25°) and is somewhat water soluble (1.8 g l^{-1}) resulting in its presence in surface or groundwater in contact with benzene-containing products. This has been most commonly seen as the presence of benzene-containing groundwater around leaking underground gasoline tanks. With a flash point of $-11.1^\circ C$ and flammability limits of 1.5–8.0% in air by volume, benzene is classified as a highly flammable substance.

Evolution of the Understanding of Benzene Toxicity

The potential for chronic toxicity of benzene was reported as early as the 1890s [3]. Exposure to large amounts of benzene was known to result in aplastic

anemia and often death. During the first half of the twentieth century, more cases of hematotoxicity following occupational benzene exposure were reported. Typical findings of that period were documented in the publication of a series of case studies by Hamilton [4]. The potential for health impacts from benzene exposure became even more well known with the publication of an extended case series of Turkish shoe and leather workers in the 1970s [5]. Aksoy, a Turkish physician, described hematotoxicity in peripheral blood of these individuals as well as described cases of leukemias. Owing to the presence of benzene in glues used by the workers, benzene concentrations in the workplace were as high as 600 ppm. Vigliani described a series of Italian printers exposed to high concentrations of benzene in the rotogravure industry, who developed hematologic disease including various leukemias [6]. Benzene concentrations were characterized as high but no quantitative measurements were reported, and the authors, therefore, did not attempt to develop a quantitative relationship between exposure and disease.

The recognition that exposure to benzene could be related to the subsequent development of one or more forms of hematopoietic malignancy was established with the publication of a retrospective cohort mortality study of workers employed by Goodyear Corporation in the manufacture of sheets of rubber hydrochloride film known as *Pliofilm* [7]. This cohort has since become referred to informally as the *Pliofilm cohort*. Pliofilm was used extensively in waterproofing and protective coating applications, from the 1930s to the 1960s or early 1970s. Although reliable air monitoring of benzene was largely nonexistent until the 1960s, this cohort has the advantage of no confounding exposures to other hydrocarbon solvents. The cohort has been updated in 1988 and again in 2002. The authors performed a conditional logistic regression analysis in which they related stratified exposures to benzene to Standardized Mortality Ratios (SMRs) for “all leukemias” in order to develop a dose–response for benzene-related leukemogenesis [8].

Subsequent to the publication of the initial Pliofilm cohort study, numerous publications the link between benzene exposure and acute forms of myeloid leukemia was solidified. However, many authors felt that lymphoid forms of hematopoietic malignancy could not be excluded. In 1997 the US National Cancer Institute in collaboration with the

Chinese Academy of Preventive Medicine published the results of a cohort study of over 70 000 Chinese workers in approximately 80 industries who were judged to be exposed to benzene in the workplace [9]. These authors reported an association between benzene exposure and increase in acute myelogenous leukemia (AML) as well as the unexpected association of benzene exposure with an increase in non-Hodgkin lymphoma (NHL). Although other authors had previously reported possible links between benzene and NHL, this was the first study that reported that NHL is related to quantitative estimates of benzene exposure rather than unspecified concentrations of benzene in hydrocarbon solvents or occupation in certain jobs considered to involve benzene exposure.

Since 1980, the Australian petroleum industry has sponsored a prospective cohort study of petroleum refinery and distribution workers in Australia. Referred to as the Health Watch study, it has been conducted by several Australian universities. Similar to US and EU petroleum cohorts, exposures have been low and the updates of the cohort mortality data have not reported consistent increases in any form of leukemia. A nested case-control study published by Glass *et al.* [10] and the update of the cohort in 2000 reported a statistically significant increase in AML after cumulative exposures of less than 8-ppm years. This cumulative exposure is much lower than the range of threshold estimates of 50–200-ppm years estimated from the updated Pliofilm cohort. A significant increase in AML at these low levels is not supported by data from the most recent update [11].

The current consensus seems to be that benzene can reliably be considered to have a causative relationship to acute nonlymphocytic leukemia (ANLL). There is considerable dispute about the relationship between benzene and NHL. This is clouded by the fact that NHL is a composite term for over 20 distinct forms of lymphoid malignancies [12].

Animal Data

Unlike many other carcinogens, there is no accepted animal model for benzene-induced leukemia. Lifetime oncogenicity studies of benzene administered by gavage or by inhalation in mice and rats have shown the presence of solid tumors in tissues such as preputial gland in male mice, Zymbal gland in

rats, and liver tumors in female mice. Hematopoietic malignancies have been observed in rats exposed to benzene, but the hematologic changes seen in rodents have not been directly equated with human leukemias (*see Cancer Risk Evaluation from Animal Studies*).

Although animal data is not currently used as a basis for quantitative risk assessment of benzene, studies on rodents and other species such as nonhuman primates have provided useful information about the pharmacokinetics and metabolism of benzene that can be applied to humans. Although a laboratory animal model of ANLL has not been discovered, the interactions of benzene and its metabolites with genetic material and their effects at the biochemical level on systems such as DNA replication and regulation of hematopoiesis have provided important insights into the potential mechanisms of benzene-induced leukemogenesis in humans [13].

Metabolism

It is accepted that benzene must be metabolized to result in hematotoxicity. The initial step in benzene metabolism is the formation of benzene oxide *via* metabolism by cytochrome P-450 IIE1 (CYP1IE1). This reactive intermediate may rearrange to form phenol (the major pathway) or may undergo ring opening leading to muconaldehyde, an alkylating agent, and ultimately to muconic acid excreted in the urine (a minor pathway). A third minor reaction of benzene oxide occurs with glutathione to form *S*-phenyl glutathione. A second hydroxylation on the phenol pathway generates either catechol or hydroquinone. Hydroquinone can oxidize to benzoquinone, a very strong alkylating agent and generator of chemically reactive oxygen species. Although there has been controversy in the past about the role of the muconic acid pathway, the oxidative route to hydroquinone is generally thought to be the toxicologically relevant pathway [2].

Understanding of the metabolism of benzene has a significant value in the risk assessment process. The initial urinary phenol test recommended by American Conference of Governmental and Industrial Hygienists (ACGIH) and Occupational Safety and Health Agency (OSHA) has been replaced by immunological assays of urinary *S*-phenyl mercapturic acid, a metabolite of *S*-phenyl glutathione. This can be as

much as 3 orders of magnitude more sensitive than former assays and there are no known dietary confounders. Identification of the enzymes involved in metabolism of benzene and in detoxication of its metabolites has led to the discovery of genetic polymorphisms in these enzymes, which may help to explain interindividual variation in benzene hematotoxicity. These polymorphisms have been related to the risk of hematopoietic toxicity after chronic exposures to high concentrations of benzene [14]. The importance of these polymorphisms to potential toxicity at much lower concentrations such as US and EU occupational standards or at urban ambient concentrations is not known.

Regulatory History

Prior to the formation of the ACGIH in 1946, benzene had been the subject of occupational exposure recommendations from various sources. A typical recommendation not to exceed 100 ppm was based upon the known hematotoxicity of benzene including aplastic anemia. This 100 ppm recommendation as an 8-h time-weighted average (TWA) was adopted by ACGIH as its first threshold limit value (TLV) for benzene. Over the years, as more information became available, the TLV was revised downward with new values of 50, 48, 35, 25, and 10 ppm between 1948 and 1970.

OSHA

Reacting to the reports of benzene-associated leukemias by Aksoy and Vigliani, and the results of the Pliofilm study, the US OSHA in 1978 promulgated an emergency standard for benzene of 1 ppm as an 8-h TWA. This was supported by a new quantitative risk assessment of the Pliofilm cohort by Crump and Allen [15]. The OSHA benzene risk assessment by Crump and Allen [15] was based upon data from the Pliofilm cohort. The authors attempted to refine the exposure analysis used by National Institute for Occupational Safety and Health (NIOSH) by assuming that benzene exposure in the workplace would have changed in parallel with changing occupational exposure recommendations. They then worked backward from typical exposures of the 1970s (when the TLV ranged between 10 and 25 ppm) to the 100 ppm recommendation during

the earliest years of the cohort. These estimated exposures were somewhat higher than the values assumed by the NIOSH authors. Crump and Allen followed the practice at that time of basing their risk assessment on mortality from all leukemias. These results were used by OSHA to set a final 8-h TWA standard for benzene of 1 ppm in 1987 [16].

USEPA

USEPA also calculated a unit risk factor for benzene in a risk assessment that was published in USEPA [17]. This estimate was performed by Crump, also the author of the OSHA risk assessment, and the total leukemia mortality outcome of the Pliofilm cohort was used as the primary basis of the estimate. In addition, two other studies of smaller chemical industry cohorts were factored into the risk estimate [18, 19]. The cancer unit risk value estimated by Crump for EPA was $0.026 \text{ mg}^{-1} \text{ M}^{-3}$. This has been updated from a single point estimate to a range of unit risk estimates from 2.6×10^{-6} to $7.8 \times 10^{-6} \mu\text{g}^{-1} \text{ M}^{-3}$ [20].

IARC

Although not a regulatory agency, the IARC is widely regarded as authoritative for classifying potential carcinogenic hazards to humans. In volume 29 of its monograph series, IARC reviewed the available data on benzene and listed it as Group I, a known human carcinogen [21].

ACGIH in the 1990s

Although ACGIH TLVs are recommendations and do not carry the force of law, they are influential in the standard setting processes of many regulatory bodies. The mortality data of the Pliofilm cohort was updated in 1988 by NIOSH. In 1990 the ACGIH proposed a reduction of the then current TLV of 10–0.1 ppm. In response to this proposal the updated cohort data was used by Paxton *et al.* [22] and Crump [23] as the basis for new risk estimates. Paxton *et al.*, using a variant of logistic regression compared risk estimates based upon exposures used by NIOSH and Crump. On the basis of the more conservative exposure estimates of Rinsky, Paxton determined that

a threshold for ANLL occurred at exposures below 50-ppm-years. The Crump analysis compared risk estimates for exposures estimated by Rinsky *et al.* [8], Crump and Allen [15], and Paustenbach *et al.* [24]. The latter estimates were based on job-specific information and interviews with surviving workers, and they incorporated potential peak exposures for some job tasks. Crump [23] also compared results from nonlinear as well as linear dose-response models. Finally in 1996 the ACGIH set a new recommendation of a 0.5-ppm TLV with a 15-min short-term exposure limit (STEL) of 2.5 ppm [25].

Europe and Other Nations

The current occupational standard for benzene is 1 ppm in the EU [26], Korea [27], and Germany [28]. In the People's Republic of China, the occupational standard has been recently reduced from 14 to 5 ppm [29]. With the exception of the Chinese occupational standard, the other national standards are all based in some way or strongly consider the results of the Pliofilm cohort.

Unresolved Issues

At present significant uncertainties remain before the risks of benzene-induced hematopoietic disease can be fully and accurately understood. The primary uncertainties are the identification of those hematopoietic malignancies that are causally related to benzene exposure and the nature of the dose-response relationship between benzene exposure and those malignancies. The latter issue includes the question of whether or not a true or apparent threshold exists for benzene-induced leukemias or lymphomas.

Cell Type

At present there is consensus that benzene can induce ANLL. Other cell types of hematological malignancies including lymphocytic leukemias and lymphomas have been alleged as being causally linked to benzene; however, the support for these cell types is weaker and multiple studies have reported inconsistent results. Myelodysplastic syndrome is

closely associated with some forms of ANLL and is often considered as potentially due to benzene exposure. Current regulatory risk assessments are based upon the assumption that benzene may be related to all or most leukemia cell types.

Dose-Response and Threshold

The mechanism by which benzene exposure results in hematologic malignancies is not currently understood, but there is a consensus that it must involve a multistep process that involves both genetic alterations resulting in clonal abnormalities and non-genetic interference with the normal regulation of hematopoiesis [30, 31]. This epigenetic dysregulation may be at the level of differentiation and/or proliferation of one or more cell types during hematopoiesis. The specifics of the mechanism have important ramifications for risk assessment in that they can define the shape of the dose-response relationship (*see Dose-Response Analysis*) between benzene exposure and leukemia response. The existence of a real or effective threshold for induction of leukemia will depend upon the specifics of the leukemogenic mechanism.

Biomonitoring – Breath Benzene, Phenol, S-Phenylmercapturic Acid (SPMA)

A significant research effort is under way to identify biomarkers that could be used to extend the range of the dose-response relationship for benzene-induced leukemogenesis. The current problem in this field is separating those phenomena that are biomarkers of effect (and causally related to induction of leukemia) from those that are merely biomarkers of exposure. The latter may be useful for exposure monitoring but bear no causal relationship to the induction of disease.

Summary

Benzene-induced leukemogenesis lies at the intersection of medicine, industrial hygiene, toxicology, and epidemiology. All of these disciplines have critical contributions to make in a risk assessment of benzene-induced leukemogenesis. At this time all of these disciplines suffer from critical uncertainties that

need to be resolved for a complete understanding of this phenomenon.

References

- [1] Wallace, L. (1988). Major sources of exposure to benzene and other volatile organic chemicals, *Risk Analysis* **10**(1), 59–64.
- [2] ATSDR (2003). *Toxicological Profile for Benzene, Agency for Toxic Substances and Disease Registry*, U.S. Public Health Service, Atlanta.
- [3] Santesson, C.G. (1897). Chronic poisoning with coal-tar benzene: four deaths. Clinical and pathological-anatomical observations of several colleagues and illustrating animal experiments, *Archiv fur Hygiene (Munich)* **31**, 336–376.
- [4] Hamilton, A. (1931). Benzene (benzol) poisoning: general review, *Archives of Pathology* **11**, 601–637.
- [5] Aksoy, M., Erdem, S. & Dincol, G. (1974). Acute leukemia due to chronic exposure to benzene, *The American Journal of Medicine* **52**, 162–166.
- [6] Vigliani, E.C. & Forni, A. (1976). Benzene and leukemia, *Environmental Research* **11**, 122–127.
- [7] Infante, P.F., Rinsky, R.A., Wagoner, J.K. & Young, R.J. (1977). Leukemia in benzene workers, *Lancet* **2**, 76–78.
- [8] Rinsky, R.A., Young, R.J. & Smith, A.B. (1981). Leukemia in benzene workers, *American Journal of Industrial Medicine* **2**, 217–245.
- [9] Hayes, R.B., Yin, S.N., Dosemeci, M., Li, G.-L., Wacholder, S., Travis, L.B., Li, C.-Y., Rothman, N., Hoover, R.N. & Linet, M.B. (1997). Benzene and the dose-related incidence of hematologic neoplasms in China, *Journal of the National Cancer Institute* **89**, 1065–1071.
- [10] Glass, D.C., Gray, C.N., Jolley, D.J., Gibbons, C., Sim, M.R., Fritschi, L., Adams, G.G., Bisby, J.A. & Manuell, R. (2003). Leukemia risk associated with low level benzene exposure, *Epidemiology* **14**, 569–577.
- [11] Gun, R.C., Ryan, P., Roder, D., Morgenstern, M., Pratt, N., Griffith, E. & McDermott, B. (2005). *Health Watch Twelfth Report*, Department of Public Health, Adelaide University, Adelaide.
- [12] McCurley, T.L. & Macon, W.R. (2004). Diagnosis and classification of non-Hodgkin lymphomas, *Wintrobe's Clinical Hematology*, 11th Edition, Lippincott Williams & Wilkins, Vol. 2, pp. 2301–2323.
- [13] Snyder, R. & Kalf, G.F. (2004). A perspective on benzene leukemogenesis, *Critical Reviews in Toxicology* **24**(3), 177–209.
- [14] Rothman, N., Smith, M.T., Hayes, R.B., Traver, R.D., Hoener, B., Campleman, S., Li, G.-L., Dosemeci, M., Linet, M., Zhang, L., Xi, L., Wacholder, S., Lu, W., Meyer, K.B., Titenko-Holland, N., Stewart, J.T., Yin, S. & Ross, D. (1997). Benzene poisoning, a risk factor for hematological malignancy, is associated with the NQO1⁶⁰⁹C → T mutation and rapid fractional excretion of chlorzoxazone, *Cancer Research* **57**, 2839–2842.
- [15] Crump, K.S. & Allen, B. (1984). *Quantitative Estimates of Risk of Leukemia from Occupational Exposure to Benzene*, Report to U.S. Occupational Safety and Health Administration.
- [16] OSHA (1987). Occupational exposure to benzene; final rule, *Federal Register* **52**, 34460–34578.
- [17] USEPA (1985). *Interim Quantitative Cancer Unit Risk Estimate Due to Inhalation of Benzene*, EPA-600/X-85-022.
- [18] Wong, O. (1978). An industrywide mortality study of chemical workers occupationally exposed to benzene II: dose-response analyses, *British Journal of Industrial Medicine* **44**, 382–395.
- [19] Ott, M.G., Townsend, J.C., Fishbeck, W.A. & Langner, R.A. (1978). Mortality among individuals occupationally exposed to benzene, *Archives of Environmental Health* **33**, 3–10.
- [20] IRIS (2000). *Benzene in USEPA Integrated Risk Information System*, at <http://www.epa.gov/ncea/iris/subst/0276.htm>.
- [21] IARC (1982). Some industrial chemicals and dyestuffs, *IARC Monographs on the Evaluation of Carcinogenic Risks of Chemicals to Humans*, International Agency for Research on Cancer, Lyon, Vol. 29.
- [22] Paxton, M.B., Chinchilli, V.M., Brett, S.M. & Rodricks, J.V. (1994). Leukemia risk associated with benzene exposure in the Pliofilm cohort, *Risk Analysis* **14**, 155–161.
- [23] Crump, K.S. (1994). Benzene-induced leukemia: a sensitivity analysis of the Pliofilm cohort with additional follow-up and new exposure estimates, *Journal of Toxicology and Environment Health* **42**, 219–242.
- [24] Paustenbach, D.J., Price, P.S., Ollison, W., Blank, C., Jernigan, J.D., Bass, R.D. & Peterson, H.D. (1992). Reevaluation of benzene exposure for the Pliofilm (rubber worker) cohort (1936–1976), *Toxicology and Environmental Health* **36**, 177–231.
- [25] ACGIH (2006). *Documentation of Threshold Limit Values for Benzene*, Cincinnati.
- [26] EC (1997). *Council Directive 97/42/EC of 27 June 1997 Amending for the First Time Directive 90/394/EEC on the Protection of Workers from the Risks Related to Exposure to Carcinogens at Work*, 97/42/EC 4–6.
- [27] Kang, S.-K., Lee, M.-Y., Kim, T.-K., Lee, J.O. & Ahn, Y.S. (2005). Occupational exposure to benzene in South Korea, *Chemico – Biological Interactions*, **153–154**, 65–74.
- [28] DFG (2003). *List of MAK and BAT Values*, Wiley-VCH, Weinheim.
- [29] Liang, Y.-X., Wong, O., Ye, X.-B., Miao, L.-Z., Zhou, Y.-M., Wu, Q.-E., Qian, H.-J. & Fu, H. (2005). an overview of published benzene exposure data by industry in China, 1960–2003, *Chemico – Biological Interact*, **153**(154), 55–64.
- [30] Smith, M.T. & Fanning, E.W. (1997). Report on the workshop entitled: modeling chemically-induced

leukemia – implications for benzene risk assessment, *Leukemia Research* **21**, 361–374.

- [31] Irons, R.D. & Stillman, W.S. (1996). The process of leukemogenesis, *Environmental Health Perspectives*, **104**(Suppl. 6), 1239–1246.

Further Reading

Capleton, A.C. & Levy, L.S. (2005). An overview of occupational benzene exposures and occupational exposure limits in Europe and North America, *Chemico-Biological Interactions* **153–154**, 43–53.

Related Articles

Environmental Hazard

Environmental Risks

Statistics for Environmental Toxicity

What are Hazardous Materials?

PATRICK BEATTY

Bioequivalence

Two drug formulations are bioequivalent if their absorption characteristics are closely similar. The most important characteristics are the extent and rate of absorption, which together define the bioavailability of a drug formulation. Thus, *bioequivalence* is the comparable bioavailability of drug formulations.

Kinetic Measures of Bioavailability

Measures After a Single Drug Administration

When drugs are not administered directly into the systemic circulation, as in the case of oral intake, at least a fraction of the dose should first be absorbed. As a result, the concentration in blood and plasma initially rises (Figure 1). The drug is then distributed to various tissues and also starts being eliminated from the body. Consequently, after reaching a peak, the concentration declines (Figure 1).

The principal measures (metrics) characterizing bioavailability, and therefore bioequivalence, are the area under the curve (AUC), contrasting concentration with time, and the maximum concentration, $C_{\max}$, which is observed at the time of $T_{\max}$ (Figure 1). AUC is a measure of the extent of absorption. AUC rises as the absorbed fraction of drug dose, i.e. the amount reaching the circulation, increases.

$C_{\max}$ is the most widely used index for the evaluation of absorption rates, especially in comparative studies. $C_{\max}$ actually reflects several processes of drug disposition, including the extent of absorption. However, if, in two drug formulations, all processes except the absorption rate have the same magnitudes (often a good assumption, at least approximately), then differences of $C_{\max}$ values indeed indicate deviations between absorption rates. A higher $C_{\max}$ signals (everything else being equal) a faster absorption rate.

$T_{\max}$ is also determined by various processes of drug disposition. Consequently, contrasts in $T_{\max}$ reflect deviations between absorption rates only if magnitudes characterizing other processes of drug disposition remain the same. Under this condition, a smaller $T_{\max}$ indicates a higher absorption rate.

AUC, $C_{\max}$, and $T_{\max}$ are often referred to as model-free measures of bioavailability because they

are determined independently of assumed pharmacokinetic models.

Measures After Repeated Drug Administrations

Following repeated administrations of a given drug dose, D , the concentration in plasma and blood eventually reaches a quasi-steady state. In this condition, after each administration of the drug, the concentration first rises, passes through a maximum, $C_{\max}$, and then declines toward a minimum or so-called trough value, $C_{\min}$. The time profile of concentrations is illustrated in Figure 2.

The kinetic measures of bioavailability parallel, in the steady state, the metrics noted for single administration. AUC is again an index of the extent of absorption. However, it is evaluated after repeated administrations by measuring the area recorded during the time interval between two dosings (the dosing or maintenance interval, T). This AUC is numerically identical to the value obtained following a single drug administration, provided that the drug exhibits first-order kinetics (i.e. the rate of change in the concentration is proportional to the concentration) and that this is identical under the two conditions.

Absorption rates are usually less important when considering repeated drug administrations than after a single drug administration. Nevertheless, $C_{\max}$ is still important, especially as an index of drug safety; unusually high concentrations could indicate a danger of toxicity.

Additional measures of bioavailability are considered later.

Assessment of Bioequivalence

Several statisticians have made distinguished contributions since the early 1970s toward developing the methodology for the assessment of bioequivalence. They notably include Carl M. Metzler and Wilfred J. Westlake and also, among others, Sharon Anderson, Walter W. Hauck, Jochen Mau, & Bruce E. Rodda. Chow and Liu [1] reviewed the statistical aspects of the evaluation of bioequivalence.

Procedures that are applied most widely at present are briefly described. Sauter *et al.* [2] illustrated examples for detailed calculations evaluating bioequivalence following single drug administrations and in the steady state.

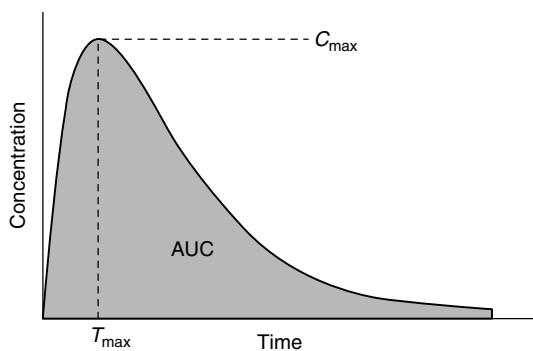


Figure 1 Time course of plasma concentration following a single oral administration of a drug

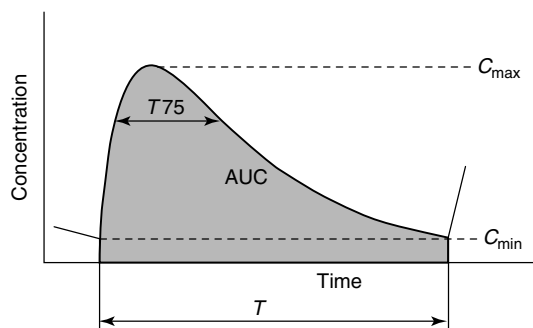


Figure 2 Time course of plasma concentration following repeated oral administrations of a drug

Two One-Sided Tests Procedure

The two one-sided tests procedure is nowadays widely applied for the assessment of bioequivalence. Yee [3] presented some features of the approach, and Schuirmann [4] elaborated them. The goal is to compare the mean kinetic responses of a reference (R) formulation and an investigated test (T) drug product. The null hypothesis to be tested is that of bioinequivalence. The test is subdivided into two one-sided problems. They assume bioinequivalence if the difference between the means is either less than or equal to a regulatory criterion θ_1 or (/and) larger than or equal to another regulatory value θ_2 :

$$H_{01} : \mu_T - \mu_R \leq \theta_1; \quad H_{02} : \mu_T - \mu_R \geq \theta_2 \quad (1)$$

The null hypothesis could be rejected in favor of an alternative hypothesis indicating bioequivalence:

$$H_a : \theta_1 < \mu_T - \mu_R < \theta_2 \quad (2)$$

If both null hypotheses are true, their evaluation at the $\alpha/2$ significance level indicates a consumer risk on both sides with this probability.

The two null hypotheses are, in practice, assessed by the use of confidence intervals. Various approaches have been proposed. The application of the shortest interval at the α level [4, 5] has been widely adopted. It yields, at a given significance level (and consumer risk), the highest power, i.e. the smallest risk for producers when the two formulations are in fact bioequivalent [6].

Implementation of the Two One-Sided Tests Procedure

Most kinetic quantities are considered to have multiplicative character (their multiplication and division – e.g. whether their magnitudes should be raised or lowered by a factor of 2 or 10 – appear to be relevant). Correspondingly, their errors are also thought to be multiplicative and not additive [6, 7]. Therefore, they are typically evaluated and compared in their logarithmic form. Consequently, μ_T and μ_R are estimated after the logarithmic transformation of the investigated quantity, e.g. from $\log AUC$ or $\log C_{max}$. The regulatory limits are usually considered to be symmetrical. Consequently, $\theta_2 = -\theta_1$ in the logarithmic scale. Times of the observations are not regarded to have multiplicative character. Moreover, they are recorded only at discrete sampling points. Therefore, T_{max} is generally not transformed and not assessed with regulatory limits. $\alpha = 0.10$ is widely applied.

The bioequivalence of two drug formulations is evaluated generally in two-period, two-sequence, crossover trials. The kinetic responses (e.g. AUC) estimated for both formulations are contrasted in each subject. From the difference of individual logarithmic responses, their average and its 90% confidence interval are calculated. Bioequivalence is declared if the limits are in the regulatory range, between θ_1 and θ_2 . The values of θ_1 and θ_2 are considered below.

Distribution-Free Procedure

Hauschke *et al.* [8] described a nonparametric procedure that, in its implementation of the two-sided hypotheses, took into account the structure of two-period, two-sequence, crossover studies. The kinetic parameters recorded with the test and reference

formulations (X_T and X_R) are contrasted within each individual:

$$X_i = X_{Ti} - X_{Ri}, \quad i = 1, \dots, n \quad (3)$$

Let the number of subjects in the two sequences of drug administration be n_1 and n_2 , with $n_1 + n_2 = n$. A total of $n_1 n_2$ differences in the two sequences,

$$X_{j1} - X_{j^*2}, \quad j = 1, \dots, n_1, j^* = 1, \dots, n_2 \quad (4)$$

are formed and sorted by magnitude. The median of the pairwise differences is a Hodges–Lehmann point estimator of $2(\mu_T - \mu_R)$. Ninety percent confidence limits are also obtained from the ranked differences.

Rapidly Evolving Issues

Various issues remain unresolved about the evaluation of bioequivalence. Two important topics are discussed that have particular relevance in biostatistics. They are developing rapidly and, therefore, their resolution in the near future is anticipated.

Individual Bioequivalence

The criteria discussed so far for the acceptance of bioequivalence involve the comparison of average kinetic parameters. Thus, the resulting similarity of two drug formulations is referred to as *average bioequivalence*. It ensures that the efficacy and safety of the new drug product is, on average, similar to that of the reference formulation. It is recognized that the efficacy and safety of the reference product was thoroughly evaluated during its development.

Average bioequivalence ensures that an individual who had not been exposed to either drug formulation would have generally similar responses to both products. Thereby, the *prescribability* of the test formulation is demonstrated. When, as often happens, patients are already receiving the reference product, then its substitution by the test formulation is contemplated. The issue is, therefore, the *switchability* from one formulation to another [9]. Thus, the similarity of responses and kinetic parameters *within* individuals, and therefore *individual bioequivalence*, becomes important.

To assess individual bioequivalence, the intrasubject variances of the reference and test products (σ_{WR}^2 and σ_{WT}^2) need to be estimated. This can be

accomplished if three- or four-period crossover trials are conducted. This design also enables the estimation of the subject $\times$ formulation interaction (σ_{SF}^2), which can be a major source of lack of individual bioequivalence.

The procedures proposed by several authors for the evaluation of individual bioequivalence can be separated into two principal categories [10, 11]. Probability-based approaches assume that deviations between individual kinetic parameters for the two formulations (X_R and X_T) would remain within a specified range and have a probability $\Pr(|X_T - X_R| \leq r)$, which should be sufficiently high. Moment-based procedures consider the second moment of $X_T - X_R$. A measure would extend that given for average bioequivalence, equation (2), by including terms for σ_{SF}^2 , σ_{WR}^2 , and σ_{WT}^2 . A one-sided criterion is generally

$$(\mu_T - \mu_R)^2 + \delta \sigma_{SF}^2 + \phi \sigma_{WR}^2 + \gamma \sigma_{WT}^2 < \theta_1 \quad (5)$$

The regulatory criterion θ_1 as well as the coefficients δ , ϕ , and η will have to be determined. For instance, it has been suggested that $\delta = \phi = 1$ and $\gamma = -1$ [12]. Sheiner [13] recommended $\delta = \gamma = 1$ and $\phi = 0$. Other values have also been proposed for the coefficients.

An unscaled measure such as that given above can be divided by a combination of σ_{WR}^2 and σ_{WT}^2 . The resulting scaled bioequivalence criterion differs intrinsically from its unscaled counterpart. It is anticipated that scaled comparisons will be particularly useful for assessing the bioequivalence of drugs that exhibit high intraindividual variability; the analysis of their equivalence is very difficult and, at times, even impossible by applying the methodology of average bioequivalence.

Similar conclusions can be drawn for probability-based measures of individual bioequivalence since they have close relationships to the corresponding moment-based measures [14, 15].

Kinetic Measures for the Evaluation of Bioequivalence

There is general agreement that AUCs measure the extent of absorption well, and that the application of the two one-sided tests procedure to relative AUCs appropriately evaluates the equivalence of extents of absorption of two drug formulations. Consequently, two drug products are considered to be bioequivalent

by this criterion if the 90% confidence interval around the geometric average of individual AUC ratios is between 0.80 and 1.25 [16].

A similar consensus has not been reached about metrics evaluating the equivalence of absorption rates. $C_{\max}$ is used most frequently. The various regulatory agencies apply differing criteria to indicate equivalence. For example, the Food and Drug Administration in the United States has requirements for $C_{\max}$ s that parallel those for AUCs: the 90% confidence interval around the geometric average of individual $C_{\max}$ ratios should be between 0.80 and 1.25. The European Union recognizes that $C_{\max}$ is determined generally with larger variation than AUC. Therefore, its condition for the equivalence of $C_{\max}$ s is less demanding: the stated 90% confidence limits should be between 0.70 and 1.43. Canadian regulatory requirements do not invoke confidence limits and are, therefore, even less demanding: it is sufficient for the declaration of bioequivalence if the geometric average of individual $C_{\max}$ ratios is between 0.80 and 1.25.

In addition to the diversity of regulatory criteria, questions have been raised about the usefulness of $C_{\max}$ for determining the equivalence of absorption rates. $C_{\max}$ notably reflects various kinetic quantities and processes, including the extent of absorption. Moreover, $C_{\max}$ responds to changes in absorption rates very insensitively, particularly in the steady state. The statistical properties of the metric are also unfavorable. Interestingly, its variation in the steady state can be higher or lower than that observed after a single drug administration, depending on the contributions of various sources of variation [17].

Therefore, reasonably, alternative measures have been suggested for assessing the equivalence of absorption rates after a single drug administration. They include, following a single drug administration, $C_{\max}/\text{AUC}$, AUC measured until the peak of the reference formulation (partial AUC), and the intercept obtained by linear extrapolation from the ratios of concentrations of the two formulations measured in the early stage of a study [18–20].

After repeated drug administrations, in the steady state, the most frequently applied metric is, in addition to $C_{\max}$, the peak – trough fluctuation, PTF; $\text{PTF} = (C_{\max} - C_{\min})/C_{\text{ave}}$, where C_{ave} and $C_{\min}$ are the average and minimum concentrations, respectively, during a dosing interval [21]. Other measures include the Swing $[(C_{\max} - C_{\min})/(C_{\min})]$, the AUC

above C_{ave} normalized by the total AUC within a dosing interval (AUCF), and the duration (T75) of the concentration peak at the level of 3/4 of its adjusted height, i.e. at $0.75C_{\max} + 0.25C_{\min}$ (see Figure 2).

Little is known, at present, about the properties of these metrics, which are being explored extensively by several investigators.

References

- [1] Chow, S.C. & Liu, J.P. (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies*, Marcel Dekker, New York.
- [2] Sauter, R., Steinijans, V.W., Diletti, E., Böhm, A. & Schulz, H.-U. (1992). Presentation of results from bioequivalence studies, *International Journal of Clinical Pharmacology, Therapeutics and Toxicology* **30**, 233–256.
- [3] Yee, K.F. (1986). The calculation of probabilities in rejecting bioequivalence, *Biometrics* **42**, 961–965.
- [4] Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability, *Journal of Pharmacokinetics and Biopharmaceutics* **15**, 657–680.
- [5] Westlake, W.J. (1981). Bioequivalence testing – a need to rethink, *Biometrics* **37**, 589–594.
- [6] Steinijans, V.W. & Hauschke, D. (1990). Update on the statistical analysis of bioequivalence studies, *International Journal of Clinical Pharmacology, Therapeutics and Toxicology* **28**, 105–110.
- [7] Westlake, W.J. (1988). Bioavailability and bioequivalence of pharmaceutical formulations, in *Biopharmaceutical Statistics for Drug Development*, K.E. Peace, ed, Marcel Dekker, New York, pp. 329–352.
- [8] Hauschke, D., Steinijans, V.W. & Diletti, E. (1990). A distribution-free procedure for the statistical analysis of bioequivalence studies, *International Journal of Clinical Pharmacology, Therapeutics and Toxicology* **28**, 72–78.
- [9] Anderson, S. & Hauck, W.W. (1990). Considerations of individual bioequivalence, *Journal of Pharmacokinetics and Biopharmaceutics* **18**, 259–273.
- [10] Anderson, S. (1993). Individual bioequivalence: a problem of switchability (with discussion), *Biopharmaceutics Report* **22**, 1–11.
- [11] Schall, R. (1995). Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar, *Biometrics* **51**, 615–626.
- [12] Schall, R. & Luus, H.E. (1993). On population and individual bioequivalence, *Statistics in Medicine* **12**, 1109–1124.
- [13] Sheiner, L.B. (1992). Bioequivalence revisited, *Statistics in Medicine* **11**, 1777–1788.
- [14] Holder, D.J. & Hsuan, F. (1993). Moment-based criteria for determining bioequivalence, *Biometrika* **80**, 835–846.

- [15] Schall, R. (1995). Unified view of individual, population and average bioequivalence, in *BioInternational 2: Bioavailability, Bioequivalence and Pharmacokinetic Studies*, H.H. Blume & K.K. Midha, eds, Medpharm, Stuttgart, pp. 91–106.
- [16] Cartwright, A.C., Gundert-Remy, U., Rauws, G., McGilveray, I., Salmonson, T. & Walters, S. (1991). International harmonization and consensus DIA meeting on bioavailability and bioequivalence testing requirements and standards, *Drug Information Journal* **25**, 471–482.
- [17] Zha, J. & Endrenyi, L. (1997). Variation of the peak concentration following single and repeated drug administrations in investigations of bioavailability and bioequivalence, *Journal of Biopharmaceutical Statistics* **7**, 191–204.
- [18] Chen, M.L. (1992). An alternative approach for assessment of rate of absorption in bioequivalence studies, *Pharmaceutical Research* **9**, 1380–1385.
- [19] Endrenyi, L. & Al-Shaikh, P. (1995). Sensitive and specific determination of the equivalence of absorption rates, *Pharmaceutical Research* **12**, 1856–1864.
- [20] Endrenyi, L., Fritsch, S. & Yan, W. (1991). $C_{\max}$ /AUC is a clearer measure than $C_{\max}$ for absorption rates in investigations of bioequivalence, *International Journal of Clinical Pharmacology, Therapeutics and Toxicology* **29**, 394–399.
- [21] Steinijans, V.W., Sauter, R., Jonkman, J.H.G., Schulz, H.U., Stricker, H. & Blume, H. (1989). Bioequivalence studies: single vs multiple doses, *International Journal of Clinical Pharmacology, Therapeutics and Toxicology* **27**, 261–266.

LASZLO ENDRENYI

Article originally published in Encyclopedia of Biostatistics, 2nd Edition (2005, John Wiley & Sons, Ltd).

Blinding

The validity of any research experiment or study is threatened by bias. Bias may arise, for example, when the patient or the person administering the treatment knows which treatment is being allocated [1]. This type of bias is known as *ascertainment bias* or *information bias* [2]. In order to reduce bias, it is necessary to ensure that the patients and the investigators are blind to (do not know) which treatment is allocated to whom. The purpose of blinding is to prevent conscious and unconscious bias associated with the expectations of patients and investigators.

Different types of bias can be prevented or reduced by blinding. For example, blinding of patients and health care providers helps prevent *performance bias* that may occur if biased supplemental care or treatment (sometimes called *cointervention*) is provided preferentially to one of the treatment groups (e.g., the placebo group) [3]. One can also blind other study personnel such as data collectors, outcome assessors, and/or data analysts. This helps to reduce the *detection bias* that arises when the knowledge about the assignment of patients influences the assessment of outcome.

The terms *single*, *double*, and *triple* blinding are often used to communicate the blinding status of the persons involved in a study. The terminology *double-blinded* usually means that neither the participants nor the investigators responsible for monitoring the participants know which intervention each participant is receiving. In a *single-blinded* study, either the participants or the trial investigators/evaluators remain unaware of intervention assignments throughout the trial. *Triple blind* usually means a double-blind trial that also maintains a blind data analysis [1].

Sometimes the nature of the study precludes blinding of participants and/or health care providers [1, 4]. Such studies include those involving a surgical procedure, comparison of medical devices, changes in lifestyle (e.g., eating habits, exercise, and cigarette smoking), and learning techniques. However, even when the study participants and those administering treatment cannot be blinded, blinding of outcome assessment could still be possible and helps prevent detection bias [3, 5].

Reporting the blinding status of the persons involved in a research study and the blinding methodology is necessary in order for the reader to be able to judge the validity of the study [4]. However, the blinding efforts are often poorly reported [5–7]. This is partially due to the ambiguity in the blinding terminology: the terms such as *single*-, *double*-, and *triple-blinded* are not well defined and vary from textbook to textbook and from physician to physician [2, 5, 6]. Therefore, stating that a study is double-blinded is not enough: the authors should explicitly state who were blinded and how they were blinded, and provide information about the blinding mechanism (e.g., caplets, tablets, etc.) and similarity of the treatment characteristics (appearance, taste, and administration) [2, 6, 8].

While all researchers agree that the blinding methodology should be clearly reported, there has been a considerable debate about the value of assessing the success of blinding at the end of the trials. Fergusson *et al.* [7] argue that researchers should assess and report the efficacy of blinding, while others argue that asking the patients or their clinicians at the end of a trial which drug they thought they were taking confounds failures in blinding with successes in pretrial hunches about treatment efficacy or adverse events [9]. For example, if the treatment is highly effective, the patients might correctly guess which group they were assigned to. In this case, the end of study test for blindness would lead to the incorrect conclusion that blinding was unsuccessful. Sackett [9] recommends rigorous testing for blindness before the trial begins, but not during the trial and never at the end of the trial.

In conclusion, the objective of blinding is to contribute to the elimination of bias that may be introduced intentionally or unintentionally by the various individuals involved in a research study. Careful reporting of the blinding status of all the individuals involved in a research study and the blinding mechanisms helps the reader to judge the validity of the study and ensures the credibility of its results.

References

- [1] Friedman, L.M., Furberg, C.D. & DeMets, D.L. (1998). *Fundamentals of Clinical Trials*, 3rd Edition, Springer-Verlag.
- [2] Schulz, K.F. (2002). The landscape and lexicon of blinding in randomized trials, *Annals of Internal Medicine* 136(3), 254–259.

- [3] Juni, P., Altman, D.G. & Egger, M. (2001). Assessing the quality of controlled clinical trials, *British Medical Journal* **323**, 42–46.
- [4] Redmond, C. & Colton, T. (2001). *Biostatistics in Clinical Trials*, John Wiley & Sons.
- [5] Schulz, K.F. & Grimes, D.A. (2002). Blinding in randomized trials: hiding who got what, *The Lancet* **359**, 696–700.
- [6] Devereaux, P.J., Manns, B.J., Ghali, W.A., Quan, H., Lacchetti, C., Montori, V.M., Bhandari, M. & Guyatt, G.H. (2001). Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials, *Journal of the American Medical Association* **285**, 2000–2003.
- [7] Fergusson, D., Glass, K.C., Waring, D. & Shapiro, S. (2004). Turning a blind eye: the success of blinding reported in a random sample of randomized, placebo controlled trials, *British Medical Journal* **328**, 432.
- [8] Montori, V.M., Bhandari, M., Devereaux, P.J., Manns, B.J., Ghali, W.A. & Guyatt, G.H. (2002). In the dark: the reporting of blinding status in randomized controlled trials, *Journal of Clinical Epidemiology* **55**, 787–790.
- [9] Sackett, D. (2004). Letter to the editor: turning a blind eye. Why we don't test for blindness at the end of our trials, *British Medical Journal* **328**, 1136.

Related Articles

Randomized Controlled Trials

ILANA BELITSKAYA-LEVY

Bonus–Malus Systems

When the actuary sets a price for an insurance policy, he has to analyze many factors, namely

- the expected loss of the policy;
- the expenses associated with writing the policy and settling the claims;
- the investment returns owing to the inversion of the business cycle;
- the capital costs.

Although all these elements are important, we concentrate here on the determination of the so-called pure premium, i.e., the expectation of the losses aggregated over a 1-year period (typically insurance policies are written for a 1-year period). The other three elements are briefly described.

Administrative expenses are incurred for writing new policies and renewing existing policies. They correspond to the salaries of the employees as well as the cost of marketing efforts to get new business. Administrative expenses are also incurred for settling the claims. They essentially correspond to the salaries of the employees as well as costs to defray experts and lawyers. These expenses are random, as it is not possible to forecast exactly what the costs will be in the future.

In motor insurance, when there are claims with bodily injuries, it often takes years before claims are completely settled. For example, claims with bodily injuries may take more than 15 years to be settled in Belgium. Because the premium is paid up front, but the claims are settled later (often years later), the insurer benefits from the investment returns on the accumulated premiums. This aspect may be taken into account to give rebates on the expected costs to the policyholders.

On the other hand, we should not forget the shareholders who provide capital to the insurance company. That capital is vital to attract policyholders. Indeed, policyholders want to make (almost) sure that the insurer will be able to compensate them (or the third parties) when there is a claim. Should the accumulated premiums be insufficient (owing to the stochastic process governing the claims generation, or because of the actuary used a wrong model), the shortfall is paid by using the capital provided by the shareholders. Obviously, shareholders want to be

remunerated for providing capital at risk and this translates into a margin that should be charged over and above the pure premium.

Denoting the annual number of losses by N , and the cost of the i th loss by X_i , we can see that the annual aggregate loss is given by $S = X_1 + \dots + X_N$. Under the assumption that the X_i s are independent and identically distributed, and independent of N , the pure premium is given by $\mathbb{E}S = \mathbb{E}N\mathbb{E}X_1$.

In automobile insurance, the number of claims is known to vary among policyholders more than the severity of the claims does. We will now focus this discussion on the number of claims filed by a policyholder.

The determination of $\mathbb{E}N$ (actually we have a model for the distribution of N) is done in two steps:

1. An *a priori* tariff is applied, based on the known characteristics of the policyholder. Declining the policy may be an option of the tariff. Imposing deductibles and/or limits on the financial compensations may also be an option in the *a priori* tariff.
2. An *a posteriori* tariff is applied after the policyholder has been observed for some insurance periods leading to bonuses or maluses. Cancellation of the policy is actually part of the *a posteriori* tariff of the insurance company.

In this chapter, we briefly comment on a particular type of *a posteriori* tarification, namely, bonus–malus systems.

An *a posteriori* tariff is made necessary when it is not possible to discover all the characteristics of the policyholders such that the *a priori* tariff is missing important elements.

We discuss bonus–malus systems in the framework of motor third party liability insurance because the claim frequency is relatively large for that kind of insurance. Fire policies have much lower claim frequencies (and also less hidden characteristics), making the use of bonus–malus systems less applicable.

Bonus–malus systems have been comprehensively studied by [1], with an update in [2]. A modern treatment of the subject is provided by [3], including a detailed list of references.

The rest of this chapter is organized as follows. The Section titled “Poisson Regression” presents the *a priori* segmentation in a Poisson regression framework. The section titled “Mixed Poisson Regression”

introduces a random effect in the Poisson model to account for heterogeneity. Bayesian updates are deduced in the section titled “Bayesian Analysis”. The section titled “Bonus–Malus Scales” discusses practical bonus–malus scales and introduces their mathematical treatment. Bonus hunger, moral hazard, and signal theory are briefly discussed in the final section.

Poisson Regression

In motor insurance, one usually uses the following covariates in order to categorize the policyholders:

- age, sex, profession, civil status, etc. of the driver;
- type of car, color, power, etc.;
- type (length, deductible, comprehensive cover, etc.) of the contract.

All these covariates allow the classification of the policyholders into relatively homogeneous classes. This is usually done through Poisson regressions. We therefore assume that the number of claims filed by the i th policyholder is Poisson distributed, $N_i \sim Po(\lambda_i)$ with expected frequency $\lambda_i = d_i \exp(score_i)$ where

- d_i denotes the risk exposure of policyholder i (i.e., the time during which the policy is in force);
- $score_i$ is the score of policyholder i modeled by a linear regression: $score_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$;
- $(x_{i1}, \dots, x_{ip})$ denotes the vector of covariates.

The score allows to rank policyholders and the exponential link ensures that the annual frequency will be positive.

The β parameters are easily estimated by standard maximum likelihood techniques.

We assume that, after a statistical analysis of the portfolio, the actuary has m *a priori* classes of drivers, having expected frequency λ_k , $k = 1, \dots, m$.

In the section titled “Bayesian Analysis”, we allow the d_i and $score_i$ to vary with time.

Mixed Poisson Regression

A Poisson model implies equidispersion: $\mathbb{E}[N_i] = \mathbb{V}[N_i]$. However empirical analyses show that there remains some heterogeneity within the classes. This is owing to the nonobservable effects, e.g., usage

of drugs or alcoholic drinks, knowledge of highway code, aggressiveness, etc. Some policyholders with characteristics x_i may have a frequency larger or smaller than λ_i . This is typically modeled by introducing a random effect in the Poisson regression: $N_i | \Theta = \theta \sim Po(\lambda_i \theta)$ such that $\mathbb{E}\Theta = 1$, i.e., we find on average the observed frequency $\mathbb{E}N_i = \lambda_i \mathbb{E}\Theta = \lambda_i$. This mixed Poisson regression model induces overdispersion, $\mathbb{V}[N_i] = \lambda_i + \lambda_i^2 \mathbb{V}[\Theta] \geq \mathbb{E}[N_i] = \lambda_i$, which is observed on empirical data sets.

In other words, the premium for each policyholder in each risk class is an average. The policies are affected by the hidden features modeled by the random effect Θ . The latter is revealed by the claims of the policyholders. Using that information, the conditional distribution of Θ is modified, implying an adaptation of the premium. This justifies the use of a *posteriori* rating schemes, as described in the section titled “Bayesian Analysis”.

Mostly for mathematical ease, the random effect is modeled by a gamma distribution, $\Theta \sim \text{Gam}(a, a)$:

$$f_{\Theta}(\theta) = \frac{1}{\Gamma(a)} a^a \theta^{a-1} \exp(-a\theta), \quad \theta > 0 \quad (1)$$

Then it is possible to show that the unconditional distribution of N_i is negative binomial:

$$\mathbb{P}[N_i = k] = \binom{a+k-1}{k} \left(\frac{\lambda_i}{a+\lambda_i} \right)^k \left(\frac{a}{a+\lambda_i} \right)^a, \quad k = 0, 1, \dots \quad (2)$$

Here again, the β parameters, as well as the a parameter are easily estimated by standard maximum likelihood techniques.

Note that other models are possible for the random effect. Inverse Gaussian or lognormal distributions are suitable alternatives. However, the mathematics becomes much simpler with the gamma distribution.

Bayesian Analysis

Assume now that we observe a policy for T periods with length $d_{i1}, \dots, d_{iT}$, i.e., the number of claims $N_{i1}, \dots, N_{iT}$ have been filed and the past *a priori* annual frequencies $\lambda_{i1}, \dots, \lambda_{iT}$ have been recorded. λ_{it} is computed as follows:

$$\lambda_{it} = d_{it} \exp(score_{it}), \quad t = 1, \dots, T \quad (3)$$

where

$$score_{it} = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}(t), \quad t = 1, \dots, T \quad (4)$$

where the covariate $x_{ij}(t)$ may now change with time. This is always the case for the variable age. It also happens when the policyholder changes his domicile or his car, for instance.

Conditionally to Θ , the random variables $N_{i1}, \dots, N_{iT}$ are assumed to be independent and Poisson distributed. Dependence between claim numbers is generated by the residual heterogeneity modeled by Θ . It is therefore relevant to analyze the conditional distribution of Θ . When Θ is gamma distributed, $\Theta \sim \text{Gam}(a, a)$, we have the following:

$$\Theta | N_{i1}, \dots, N_{iT} \sim \text{Gam}(a + N_{i\bullet}, a + \lambda_{i\bullet}) \quad (5)$$

where $N_{i\bullet} = \sum_{t=1}^T N_{it}$ and $\lambda_{i\bullet} = \sum_{t=1}^T \lambda_{it}$.

This immediately implies

$$\mathbb{E}[\Theta | N_{i1}, \dots, N_{iT}] = \frac{a + N_{i\bullet}}{a + \lambda_{i\bullet}} \quad (6)$$

Furthermore, we have that

$$\begin{aligned} \mathbb{E}[N_{i(T+1)} | N_{i1}, \dots, N_{iT}] &= \lambda_{i(T+1)} \\ &\times \mathbb{E}[\Theta | N_{i1}, \dots, N_{iT}] \end{aligned} \quad (7)$$

from which we conclude

$$\mathbb{E}[N_{i(T+1)} | N_{i1}, \dots, N_{iT}] = \lambda_{i(T+1)} \frac{a + N_{i\bullet}}{a + \lambda_{i\bullet}} \quad (8)$$

For instance, for a policyholder having constant expected frequency $\lambda = 0.10$ and a random effect modeled by a gamma distribution with $a = 1.25$, we can draw a table as shown in Table 1.

Table 1 *A posteriori* corrections for a policyholder with expected frequency $\lambda = 0.10$ and $a = 1.25$

$t/N_{i\bullet}$	0(%)	1(%)	2(%)
1	93	167	241
2	86	155	224
3	81	145	210
4	76	136	197
5	71	129	186
6	68	122	176
7	64	115	167
8	61	110	159

This bonus–malus table is not easy to implement in practice. This is why insurance companies resort to so-called bonus–malus scales.

Bonus–Malus Scales

Let us assume a bonus–malus scale with $s + 1$ levels ℓ numbered from 0 to s . To each level ℓ corresponds a premium level r_ℓ , also called *relativity*. The relativities are multiplied by a base premium (BP) to obtain the premium paid by the policyholder (see Table 2).

When the driver makes a claim, he suffers an increase of his position in the scale, resulting in an increase of his premium: this is the malus. When the driver is claim free, he gets a bonus by going down in the scale, resulting in a lower premium. Each bonus–malus system has well-defined transition rules showing the behavior of the policyholders within the scale. The driver never goes below level $\ell = 0$ or above level $\ell = s$.

There is no model to define the transition rules. These rules are set by the insurance company as a function of its commercial approach of the market. The relativities should then be computed in a fair way.

Because the cost of the claims is essentially a function of the third party, it cannot be explained by the behavior of the policyholder. Therefore, the severity of the claims usually is not taken into account in the transition rules. This, however, leads to the hunger for bonus described in the section titled “Bonus Hunger, Moral Hazard, and Signal Theory”.

As an example, let us mention the transition rules of the old Belgian bonus–malus scale (see Table 3). Before 2004, a unique bonus–malus scale was imposed to all insurers by law. Since 2004, the market has been freed. However, most insurers have

Table 2 Description of a bonus–malus scale

Level	Premium	Relativity
s	b_s	r_s
$\vdots$	$\vdots$	$\vdots$
ℓ	$b_\ell = r_\ell \times BP$	r_ℓ
$\vdots$	$\vdots$	$\vdots$
1	b_1	r_1

Table 3 Old Belgian bonus–malus scale

Class	Relativities (%)	Class	Relativities (%)	Class	Relativities (%)
22	200	15	105	7	69
21	160	14	100	6	66
20	140	13	95	5	63
19	130	12	90	4	60
18	123	11	85	3	57
17	117	10	81	2	54
16	111	9	77	1	54
		8	73	0	54

applied only cosmetic changes to the old scale we present now. The scale has 23 steps numbered 0 to 22. A new driver enters the scale at level 11 (or 14 if he has professional use of the car). Each year the policyholder goes one step down. Each reported accident is penalized by a five steps increase. In other words, when the policyholder does not claim, he gets a one step bonus; if he reports one accident, he has a malus of four steps; if he reports two claims, he has a malus of nine steps, etc. The policyholder never goes below step 0 or above step 22. Moreover, there is a superbond rule allowing a driver being above level 14, with four consecutive claim-free years to be sent to level 14.

In practice, such a bonus–malus scale can be represented by a finite Markov chain. Indeed, the knowledge of the current level and of the number of claims of the current year suffices to determine the next level in the scale. Moreover, the Markov chains are generally irreducible, meaning that all states are always accessible in a finite number of steps from all other states (having a bonus–malus scale whose associated Markov chain is not irreducible would be difficult to justify to the policyholders). Moreover, bonus–malus scales have a bonus level with maximal reward: policyholders being in that level will remain in that level after a claim-free year. Both these properties ensure that the Markov chain associated to the bonus–malus scale has a limiting distribution, or steady state distribution, which is also the stationary distribution.

Norberg [4] suggested the minimization of the expected squared difference between the true relative premium Θ (i.e., the correction to apply to the policyholder according to our model with a random effect), and the relative premium r_L of the scale (i.e., the correction to apply according to the simplified model based on a bonus–malus scale), $\mathbb{E}(\Theta - r_L)^2$,

to get the relativities r_ℓ . Pitrebois *et al.* [5] have extended this methodology to the case with *a priori* segmentation. It gives the following:

$$r_\ell = \frac{\sum_{k=1}^m w_k \int_0^\infty \theta \pi_\ell(\lambda_k \theta) dF_\Theta(\theta)}{\sum_{k=1}^m w_k \int_0^\infty \pi_\ell(\lambda_k \theta) dF_\Theta(\theta)} \quad (9)$$

where

- F_Θ is the cumulative density function (cdf) of Θ ;
- w_k is the weight of the k th class;
- λ_k is the expected frequency within the k th class;
- $\pi_\ell(\lambda_k \theta)$ is the probability that a policyholder with frequency $\lambda_k \theta$ is at the level ℓ at steady state.

This formula recognizes the interaction between *a priori* and *a posteriori* tarification. Indeed, an insurer using few, or even no *a priori* covariates needs stronger *a posteriori* corrections, whereas an insurer that would be able to discover all the relevant characteristics (even those being hidden) of the driver actually does not need any *a posteriori* correction. Note that the old bonus–malus systems in force in the EU did not recognize that fact. Insurers were, indeed, allowed to apply the *a priori* segmentation of their choice, whereas the bonus–malus scale was imposed by law, independently of the *a priori* segmentation.

Results regarding the Belgian bonus–malus scale have been obtained by Pitrebois *et al.* [6], based on real life data.

Bonus Hunger, Moral Hazard, and Signal Theory

An important effect of bonus–malus scales is the so-called bonus hunger. Policyholders causing small claims may be tempted to defray themselves the third party to avoid maluses. This strategy is very natural and implies censoring: small claims are not observed. Actually, the bonus hunger helps the insurer in reducing its administrative expenses. Indeed, the latter are the most important (in relative terms) for small claims. Not being reported to the insurer, administration costs are saved.

Lemaire [7] has described an algorithm allowing the policyholder to obtain his optimal retention as a function of his level in the scale, as well as his claiming frequency.

The bonus hunger effect is important to cope with when an insurance company wants to change a bonus–malus scale. The new scale may indeed have another effect on the bonus hunger of the policyholders. Lemaire’s algorithm allows to find the optimal strategy of the policyholders. However, it requires the true claim amount distribution and the true frequency distribution. These are not easy to obtain owing to the censoring effect. Attempts to cope with these parameters have been done in [8] and in [3].

Another reason why *a posteriori* rating schemes are more relevant is that they counteract moral hazard. When the premium is fixed, together with the absence of deductible (which is often the case in motor third party liability insurance), there is no incentive for the policyholder to avoid claims. Because (s)he knows that (s)he has a cover, the policyholder may adopt a more aggressive style of driving without having to support the potential financial consequences. This would be the realization of moral hazard risk: the cover is now influenced consciously by human behavior. It is clear that moral hazard should be avoided as far as possible. Imposing deductibles is a classical way to counteract moral hazard. *A posteriori* rating schemes clearly allow for counteracting moral hazard behaviors: if the policyholder claims, his (her) premium will increase at the next renewal. So there is a financial incentive for the policyholder not to claim.

Bonus–malus scales may also be used to discover the hidden characteristics of the policyholders.

Assume that an insurer proposes two scales to its policyholders: a soft scale (with mild penalties) and a strong scale (with heavy maluses). Bad risks (not necessarily revealed through their *a priori* characteristics) will be tempted to choose the soft scale. By doing so, they actually signal themselves to the insurance company. This is an application of the signal theory.

References

- [1] Lemaire, J. (1985). *Automobile Insurance: Actuarial Models*, Kluwer Academic Publishers, Netherlands.
- [2] Lemaire, J. (1995). *Bonus-Malus Systems in Automobile Insurance*, Kluwer Academic Publishers, Boston.
- [3] Denuit, M., Maréchal, X., Pitrebois, S. & Walhin, J.F. (2007). *Actuarial Modelling of Claim Counts. Risk Classification, Credibility and Bonus-Malus Scales*, John Wiley & Sons.
- [4] Norberg, R. (1976). A credibility theory for automobile bonus systems, *Scandinavian Actuarial Journal* 92–107.
- [5] Pitrebois, S., Denuit, M. & Walhin, J.F. (2003). Setting a bonus-malus scale in the presence of other rating factors: Taylor’s work revisited, *ASTIN Bulletin* 33, 419–436.
- [6] Pitrebois, S., Denuit, M. & Walhin, J.F. (2003). Fitting the Belgian bonus-malus system, *Belgian Actuarial Bulletin* 33, 419–436.
- [7] Lemaire, J. (1977). La soif du bonus, *ASTIN Bulletin*, 9, 181–190.
- [8] Walhin, J.F & Paris, J. (2001). The true claim amount and frequency distributions within a bonus-malus system, *ASTIN Bulletin* 30, 391–403.

JEAN-FRANÇOIS WALHIN

Burn-in Testing: Its Quantification and Applications

Definitions

1. In MIL-STD-883C [1, 2], method 1015.3, burn-in test is defined as follows:
Burn-in is a test performed for the purpose of screening or eliminating marginal devices, those with inherent defects or defects resulting from manufacturing aberrations which cause time and stress dependent failures (*see Reliability Integrated Engineering Using Physics of Failure; Hazard and Hazard Ratio*).
2. Kuo and Kuo in [3] define the burn-in more clearly as follows:
Burn-in combines the appropriate electrical conditions with the appropriate thermal conditions to accelerate the aging of a component or device. In other words, burn-in is a process during which electronic components or systems are operated under specific electrical and thermal conditions to determine the early life of components and products (*see Product Risk Management: Testing and Warranties; Design of Reliability Tests; Repair, Inspection, and Replacement Models*) and the substandard ones are forced to fail and thus removed from the production.

Why Burn-in?

In reliability and risk analysis (*see Multistate Systems; Imprecise Reliability; Lifetime Models and Risk Assessment; Mathematics of Risk and Reliability; A Select History*) the lifetime of a population of products are often described by using a graphical model called the *bathtub curve* (*see Hazard and Hazard Ratio; Accelerated Life Testing*) in Figure 1 [4]. The bathtub curve is composed of three periods: an infant mortality period with a decreasing failure rate, followed by a life period known as *useful life* with a low, relatively constant failure rate, and concluding with a wear-out period that exhibits an increasing failure rate.

Many types of equipment and systems exhibit an inherent decreasing failure rate characteristic during their early operating life. A relatively high early failure rate is usually attributed to the inherent variability in the production process. Burn-in has long been recognized as a useful method for detecting and eliminating components and systems with early failure risk before customer delivery. Without burn-in, defective components are delivered to customers, resulting in costly field repairs, damage to the reputation of the manufacturers, and thus loss of business.

Also from Figure 1, if the operating temperature K is higher, then the product's life will be shortened, as well as the burn-in period. This is the reason why many products will be burned-in at a higher temperature before they are released to customers, thus saving time and cost of the testing process.

Burn-in testing is not a product improvement technique. Submitting a population of a product to burn-in does not improve each unit's individual reliability. However, burn-in improves the reliability of the whole population by weeding out the bad portion of the population, resulting in a more reliable and homogeneous population of units. Also, burn-in usually provides insightful information about the various failure mechanisms that cause early failures and provides valuable feedback to the quality and reliability engineers who oversee the design and the manufacturing process to help them analyze the variations and causes that lead to the emergence of these early failure mechanisms.

Burn-in Test Types and Applications

The methods used to stress all the nodes in a device depend on the complexity and failure mechanisms of the integrated circuit (IC). Therefore, they fall into the following categories:

1. Static burn-in test

The simplest type of burn-in test is static, or steady-state, burn-in method. Static burn-in maintains a steady-state bias on each device under high temperature for a number of hours to accelerate the migration of impurities to the surface so that a failure will occur. A static system is cheaper and simpler, and is used to precipitate contamination-related failure mechanisms. It is, however, less effective than dynamic burn-in for large scale integration (LSI) and very large scale integration (VLSI) devices.

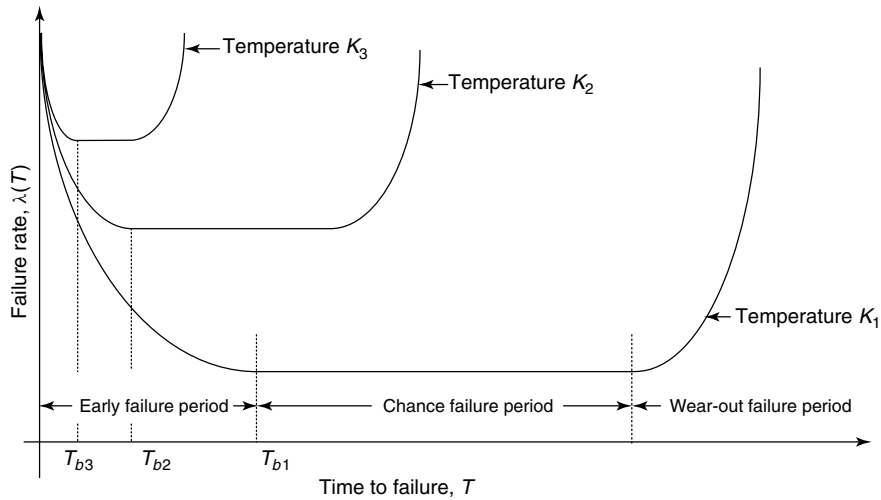


Figure 1 The bathtub curve

2. Dynamic burn-in test

The dynamic burn-in test method uses power sources such as voltage, clock signals, and address signals to ensure all internal nodes are reached during temperature stressing. It is more effective at detecting early failures in complex devices. It is also more expensive and requires more dedicated burn-in boards. It should be noted that a dynamic burn-in system can also be used for static burn-in, but not the other way around.

3. Test during burn-in

A subset of dynamic burn-in is test during burn-in (TDBI). TDBI adds functional testing and, possibly, monitoring of component outputs to show how they are responding to specific input stimuli. TDBI is the most comprehensive burn-in technique, especially when coupled with scan-based technology. It has been used primarily for dynamic random access memories (DRAMs), but it is applicable to all large memories owing to their long electrical test times. Normally, the electrical testing is performed after burn-in to detect any potential failures that may occur during their useful life period. TDBI is not appropriate for the microprocessors and other VLSI circuits of EPROM (erasable and programmable read only memory).

4. High-voltage stress tests

High-voltage stress tests are categorized as burn-in screens because of their device-aging acceleration features due to the application of higher voltage, time,

and temperature. For these tests, the distribution of the voltage stress throughout the IC is accomplished by carefully designed dynamic and functional operation. IC memory suppliers have used high-voltage stress tests in lieu of dynamic burn-in as a means to uncover oxide defects in metal oxide semiconductor (MOS) ICs. A typical high-voltage stress test involves cycling through all addresses (for a memory) using selected memory data patterns for 2 s, both high and low (logically speaking), with 7.5 V forcing function being applied, for a 5-V rated part, for example.

Other classifications have been proposed by Foster [5], according to which the generally used burn-in methods can be categorized as follows:

1. Elevated temperature plus power – the cheapest but the least effective method.
2. Elevated temperature plus power with all inputs reverse biased, or the so-called high-temperature reverse bias (HTRB) method – a method with moderate cost and reasonable effectiveness for most devices.
3. Elevated temperature, power, dynamic excitation of inputs, and full loading of all outputs – an effective and expensive method.
4. Optimum biasing combined with temperature in the range of 200–300 °C, or the so-called high-temperature operating test (HTOT) – an expensive and difficult-to-carry-out method, which is

not applicable to plastic devices owing to the high temperatures involved.

Burn-in Test Conditions

In MIL-STD-883C, the following six basic test conditions are specified for the burn-in test:

Test condition A – steady state, reverse bias

This test condition is suitable for use on all types of circuits, both linear and digital. In this test, as many junctions as possible will be reverse biased to the specified voltage.

Test condition B – steady state, forward bias

This test condition can be used on all digital type circuits and some linear types. In this test, as many junctions as possible will be forward biased to the specified voltage.

Test condition C – steady state, power, and reverse bias

This test condition can be used on all digital type circuits and some linear types where the inputs can be reverse biased and the output can be biased for maximum power dissipation or *vice versa*.

Test condition D – parallel or series excitation

This test condition is suitable for use on all circuit types. Parallel or series excitation, or any combination thereof, is permissible. However, all circuits must be driven with an appropriate signal to simulate, as closely as possible, circuit application, and all circuits should have the maximum load applied. The excitation frequency should not be less than 60 Hz.

Test condition E – ring oscillator

In this test condition, the output of the last circuit is normally connected to the input of the first circuit. The series will be free running at a frequency established by the propagation delay of each circuit and the associated wiring, and the frequency shall not be less than 60 Hz. In the case of circuits that cause phase inversion, an odd number of circuits shall be used. Each circuit in the ring shall be loaded to its rated maximum capacity.

Test condition F – temperature-accelerated test

Under this test condition, microcircuits are subjected to bias(es) at an ambient test temperature, typically from 151 to 300 °C, which considerably exceeds their maximum rated junction temperature. It is generally

found that microcircuits will not operate properly at these elevated temperatures. Therefore, special attention should be given to the choice of bias circuits and conditions to assure that important circuit areas are adequately biased without incurring any damaging overstresses to other areas of the circuit. To select the accelerated test conditions properly, it is recommended that an adequate sample of devices be exposed to the intended high temperature while measuring voltage(s) and current(s) at each device terminal to assure that the applied electrical stresses do not induce damaging overstresses.

The specified test temperature is the minimum actual ambient temperature to which all devices in the working area of the chamber shall be exposed. This should be assured by making whatever adjustments are necessary in the chamber profile, loading, location of control or monitoring instruments, and the flow of air or other suitable gas or liquid chamber medium. The ambient burn-in test temperature for conditions A–E shall be 125 °C minimum.

Burn-in Time Determination

The determination of the correct burn-in time requires the use of reliability methodologies, as well as the costs involved, *versus* achievement of the optimum failure rate [6].

Method 1: Determine the Burn-in Time Using a Histogram

A histogram is a very quick and useful method to determine the burn-in time when a large amount of time-to-failure data are available from testing or field returned products. The following example illustrates this method in details.

Example 1 The time-to-failure data are as given in columns 1 and 2 in Table 1 [1]. The data are based on 900 devices subjected to burn-in, and their performance is monitored in 10-h intervals to determine the number of devices that fail in each interval. Draw the failure rate histogram and decide the best burn-in time.

Solution to Example 1

1. Calculate the number of units surviving after each 10-h period and list them in column 3 of Table 1.

Table 1 Data of failures

1	2	3	4
Life (h)	Number of failures, N_F	Number of units surviving period, N_{BP}	Failure rate $\hat{\lambda} = \frac{N_F(\Delta T)}{N_{BP}(T) \times \Delta T}$
0–10	33	$900 - 33 = 867$	$\frac{33}{900 \times 10} = 0.003667$
10–20	31	$867 - 31 = 836$	$\frac{33}{900 \times 10} = 0.003667$
20–30	28	$836 - 28 = 808$	$\frac{33}{900 \times 10} = 0.003667$
30–40	25	$808 - 25 = 783$	$\frac{33}{900 \times 10} = 0.003667$
40–50	23	$783 - 23 = 760$	$\frac{33}{900 \times 10} = 0.003667$
50–60	21	$760 - 21 = 739$	$\frac{33}{900 \times 10} = 0.003667$
60–70	20	$739 - 20 = 719$	$\frac{33}{900 \times 10} = 0.003667$
70–80	19	$719 - 19 = 700$	$\frac{33}{900 \times 10} = 0.003667$
80–90	18	$700 - 18 = 682$	$\frac{33}{900 \times 10} = 0.003667$
90–100	18	$682 - 18 = 664$	$\frac{33}{900 \times 10} = 0.003667$
100+	664	$664 - 664 = 0$	–
$N = 900$			

- Use the following equation to estimate the failure rates in each 10-h period and list them in column 4.

$$\hat{\lambda}(T) = \frac{N_F(\Delta T)}{N_{BP}(T) \times \Delta T} \quad (1)$$

where $\hat{\lambda}(T)$: average estimate of the failure rate in each ΔT time interval, $N_F(\Delta T)$: number of devices failing in ΔT time interval, $N_{BP}(T)$: number of devices at the beginning of each ΔT interval, and ΔT : time interval for which the $\hat{\lambda}(T)$ is calculated.

- Draw the histogram in Figure 2.
- In Figure 2, the failure rate decreases with time and approaches a stable, or constant, value at about 84 h. So the optimum burn-in time is about 84 h.

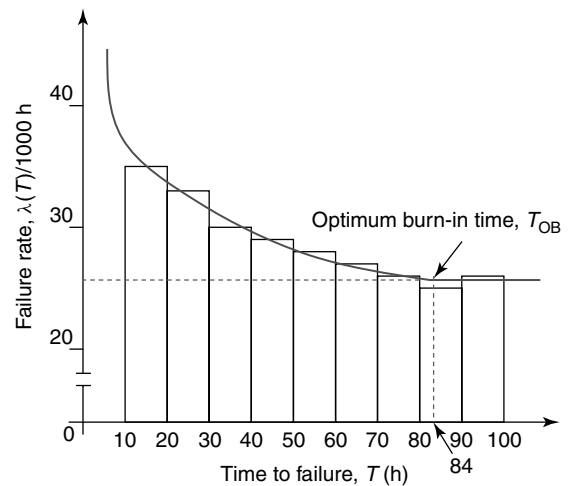


Figure 2 The histogram of the failure rate function

Method 2: Determine the Burn-in Time Using the Mixed Population Model

In contrast to the histogram method, the mixed population model provides an analytical way to decide the best burn-in time, as well as to quantify the percentages of different subpopulations in the whole product.

Let p denote the proportion, or percentage, of the i th subpopulation in the entire population, where

$$\sum_{i=1}^n p_i = 1, \quad 0 \leq p_i \leq 1 \quad (2)$$

The reliability and the probability density functions are then given, respectively, by

$$R(T) = \sum_{i=1}^n p_i R_i(T) \quad (3)$$

and

$$f(T) = \sum_{i=1}^n p_i f_i(T) \quad (4)$$

where $R(T)$ and $f(T)$ are the reliability function and the *pdf* of the entire population, respectively.

The failure rate function of the population (*see Reliability Data*) is given by

$$\lambda(T) = \frac{f(T)}{R(T)} = \frac{\sum_{i=1}^n p_i f_i(T)}{\sum_{i=1}^n p_i R_i(T)} \quad (5)$$

This model can be used to describe the failure behavior of the population, which contains n mutually exclusive and well-separated subpopulations with early, chance, and wear-out failure modes. This characteristic makes it very flexible to give a wide variety of failure rate curves, including the classic bathtub curve. Usually, there are three subpopulations, which are taken to be Weibull distributed.

Once the failure rate function, $\lambda(T)$, is determined, it is easy to plot this function *versus* time and find out the best burn-in time, at which the failure rate approaches a low and constant value.

Example 2 A total of 1000 units are tested to failure and after analysis we have the following results:

100 of them are early failures with the following Weibull distribution parameters: $\gamma = 0$, $\beta = 0.5$, $\eta = 20$ h; 700 of them are chance failures having a constant failure rate of $\lambda = 0.0003$ fr/h; and 200 of them are wear-out failures with the following Weibull distribution parameters:

$$\gamma = 0, \beta = 3.5, \eta = 4000 \text{ h}$$

Write down the equation for $\lambda_{e,c,w}(T)$, plot it, and decide the best burn-in time.

Solution to Example 2

The percentage of early failures in the entire population is

$$p_1 = \frac{N_1}{N} = \frac{100}{1000} \times 100\% = 10\% \quad (6)$$

the reliability function is

$$R_1(T) = e^{-\left(\frac{T}{\eta}\right)^\beta} = e^{-\left(\frac{T}{20}\right)^{0.5}} \quad (7)$$

and the *pdf* is

$$\begin{aligned} f_1(T) &= \frac{\beta}{\eta} \left(\frac{T}{\eta}\right)^{\beta-1} e^{-\left(\frac{T}{\eta}\right)^\beta} \\ &= \frac{0.5}{20} \left(\frac{T}{20}\right)^{-0.5} e^{-\left(\frac{T}{20}\right)^{0.5}} \end{aligned} \quad (8)$$

The percentage of chance failures in the entire population is

$$p_2 = \frac{N_2}{N} = \frac{700}{1000} \times 100\% = 70\% \quad (9)$$

the reliability function is

$$R_2(T) = e^{-\lambda T} = e^{-0.0003T} \quad (10)$$

and the *pdf* is

$$f_2(T) = \lambda e^{-\lambda T} = 0.0003 e^{-0.0003T} \quad (11)$$

The percentage of wear-out failures in the entire population is

$$p_3 = \frac{N_3}{N} = \frac{200}{1000} \times 100\% = 20\% \quad (12)$$

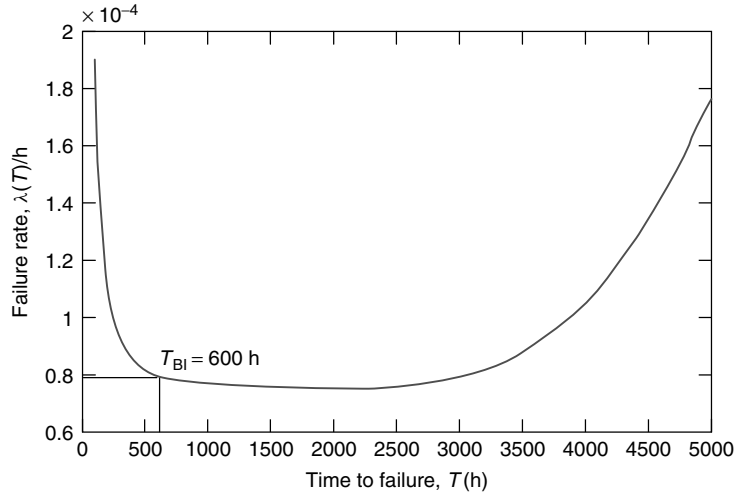


Figure 3 Failure rate curve for equation (5)

the reliability function is

$$R_1(T) = e^{-\left(\frac{T}{4000}\right)^{3.5}} \quad (13)$$

and the *pdf* is

$$\begin{aligned} f_1(T) &= \frac{\beta}{\eta} \left(\frac{T}{\eta}\right)^{\beta-1} e^{-\left(\frac{T}{\eta}\right)^{\beta}} \\ &= \frac{3.5}{4000} \left(\frac{T}{4000}\right)^{2.5} e^{-\left(\frac{T}{4000}\right)^{3.5}} \end{aligned} \quad (14)$$

Using equation (5)

$$\begin{aligned} \lambda_{e,c,w}(T) &= \frac{f(T)}{R(T)} = \frac{\sum_{i=1}^n p_i f_i(T)}{\sum_{i=1}^n p_i R_i(T)} \\ &= \frac{10\% \cdot \frac{0.5}{20} \left(\frac{T}{20}\right)^{-0.5} e^{-\left(\frac{T}{20}\right)^{0.5}} + 70\% \cdot 0.0003 e^{-0.0003T} + 20\% \cdot \frac{3.5}{4000} \left(\frac{T}{4000}\right)^{2.5} e^{-\left(\frac{T}{4000}\right)^{3.5}}}{10\% \cdot e^{-\left(\frac{T}{20}\right)^{0.5}} + 70\% \cdot e^{-0.0003T} + 20\% \cdot e^{-\left(\frac{T}{4000}\right)^{3.5}}} \end{aligned} \quad (15)$$

It may be seen from Figure 3 that the failure rate curve has an initial period during which the failure rate decreases sharply to a constant, useful life level, stays at this level for quite a long period,

and then increases as wear-out sets in. The burn-in time is the time after which the failure rate remains constant. For this numerical example the burn-in time is approximately $T = 600$ h.

References

- [1] Kececioglu, D.B. & Sun, F.-B. (2003). *Burn In Testing – It's Quantification and Optimization*, DEStech Publications, Lancaster, pp. 1–6, 15–27.
- [2] MIL-STD-883C (1983). *Test Methods and Procedures for Microelectronics*, August.

- [3] Kuo, W. & Kuo, Y. (1983). Facing the headaches of early failures: a state-of-the-art review of burn-in decisions, *Proceedings of the IEEE* 71(11), 1257–1266.

-
- [4] *Quantifying Optimum Burn-In Period*. <http://www.weibull.com/hotwire/issue58/hottopics58.htm>.
- [5] Foster, R.C. (1976). Why consider screening, burn-in, and 100-percent testing for commercial devices, *IEEE Transactions on Manufacturing Technology* **MFT-5**(3), 52–58.
- [6] Kececioglu, D.B. (2003). *Reliability Engineering Handbook*, DEStech Publications, Lancaster, Vol. 1, pp. 628–631.

DIMITRI B. KECECIOGLU AND XIAOFANG
CHEN

1,3-Butadiene

Physical and Chemical Characteristics

1,3-Butadiene (BD) (CAS #106-99-00) is a colorless gas with a structure consisting of four carbons linked by two double bonds at the first and third carbons (MW = 54.1). BD is flammable with a lower explosion limit of 2% (20 000 ppm). It is a major monomer used in the manufacture of rubber and plastics. During World War II, BD was used in large quantities to make synthetic rubber to replace the supplies of natural rubber that were no longer available to the allied forces. At that time, BD was considered to have low toxicity and the occupational threshold limit value was set at 1000 ppm (the level of exposure that the typical worker can experience without an unreasonable risk of disease or injury).

Exposure Potential

Exposure to BD outside the occupational arena is common [1], but this occurs at lower concentrations. Average annual ambient air concentrations in the United States are reported to range from 9 ppt to 3.2 ppb with an overall national mean concentration of 0.12 ppb. Motor vehicle emissions are the largest source of ambient BD, but BD is also present in cigarette smoke and in emissions from other combustion sources.

Acute Toxicity

Extremely high concentrations of BD (>100 000 ppm) cause irritation of the respiratory tract, eyes, and skin and induce suppression of the central nervous system in both animals and humans [2].

Noncancer Chronic Toxicity

Noncancer effects of BD exposure have been studied in animals. The reproductive and developmental toxicity of BD was studied in rats and mice [3] exposed to up to 1000 ppm BD on critical gestation days. No fetal toxicities were observed in rats but mice exposed to as low as 200 ppm BD had fetal anomalies. In bioassay studies, in which animals are

exposed to varying levels of BD over approximately 2 years, atrophy of the gonads was observed in male mice at exposure levels of 625 ppm BD or higher, and in female mice after exposure to 62.5 ppm [4, 5]. No such effects were observed in rats [6]. Rats and mice also exhibited different responses to BD on the hematopoietic and immune systems. Rats exposed for 13 weeks to 8000 ppm BD showed no signs of general toxicity, including hematotoxicity [7]. Mice exposed in a similar fashion had suppressed immune systems and toxicity to the bone marrow [8]. Later studies confirmed the hematotoxicity of BD in mice, which developed a macrocytic anemia following chronic, near lifespan exposure to 1250 ppm BD, indicating that the bone marrow was a target organ for BD toxicity in mice [9].

Potential for Carcinogenicity

Because BD was used in such large volumes, 2-year rodent bioassays were conducted to determine if rats or mice would suffer any adverse health effects, and in particular, cancer, if they were exposed to BD at selected air concentrations for most of their lifetimes. In the rat bioassay, animals were exposed to 1000 or 8000 ppm for about 2 years [6]. In line with previous expectations, the BD showed little noncancer toxicity and was a weak carcinogen in the rats. The tumors observed were mainly in hormonally dependent tissues (mammary glands and uterus in females; thyroid, pancreas, and testis in males) and only at the higher exposure concentration, with the exception that an increase in mammary gland tumors was observed in females at both exposure concentrations. In contrast to the rats, the mice exposed over most of their lifespan to much lower concentrations of BD (6.5–1250 ppm) indicated that BD was a potent multisite carcinogen in this species, resulting in neoplasms of the heart, lung, mammary glands, and ovaries [4, 5]. Increased lung tumors were observed in female mice even at the lowest exposure concentration of 6.25 ppm BD. The incidence of thymic lymphomas in the mice exposed to concentrations of 200 ppm BD or higher led to high mortality, but this result was considered to be the result of activation of an endogenous retrovirus in the specific strain of mice tested and not relevant to risk in humans [9].

Metabolism of BD

A major question raised by the animal studies was which species, if either, was predictive of human sensitivity to BD. In an attempt to help answer this question, many researchers studied differences in how the rats and mice metabolize BD [2]. Rats, mice, and humans all have common ways to get rid of foreign organic compounds through metabolism (alteration of the form of the chemical). Enzymes in the body first oxidize the compound to a reactive state, which can then be hydrolyzed to a form that can be excreted in the urine. For 1,3-butadiene, enzymatic oxidation can lead to formation of reactive epoxides at either or both of the double bond sites. These epoxides can then be hydrolyzed to diols, which form conjugates with reduced glutathione, and are then excreted in the urine as mercapturic acids. Unfortunately, sometimes the reactive form (the epoxides) reacts with the genetic material, the DNA, before the epoxide can be hydrolyzed for excretion. This reaction with DNA can lead to alterations (mutations) in the genetic material that result in loss of control of cell growth, and cancer. Thus the oxidation of BD is considered to be a toxication pathway, while hydrolysis and conjugation are detoxication pathways.

BD can be oxidized enzymatically to one of the three reactive epoxides, a monoepoxide, a diepoxide, or an epoxydiol. The epoxydiol is a compound in which one double bond is oxidized to an epoxide and the second double bond is occupied by two hydroxyl groups. This epoxide can be formed by hydrolysis of either the monoepoxide or the diepoxide. The diepoxide is a much more potent mutagen than monoepoxide or the epoxydiol [10] and the basic difference between the rats and mice was found to be that the mice form much more of the diepoxide than do the rats. This evidence is based on a comparison of metabolites in the blood of exposed rats and mice and suggestive evidence based on the urinary metabolites of BD in humans, rats, and mice [11].

Another difference between the species that influences the toxicity of BD is the balance between the activities of the oxidizing enzymes (forming toxic, reactive metabolites) and the hydrolyzing or conjugating enzymes (detoxification activities) in the livers of the different species. The oxidizing enzymes and the hydrolytic enzymes are both located in the microsomes located in the rough endoplasmic reticulum surrounding the nucleus where the DNA

is located [12]. If the hydrolyzing enzymes are active enough, the toxic oxidation products can be detoxicated quickly before they can react with the DNA. If the hydrolyzing activity is low, the monoepoxide can be further oxidized to the much more potent mutagen, the diepoxide. Among several species studied, the mouse has the lowest liver hydrolytic activity and also the highest ratio of oxidation to hydrolysis activity. Primates, including humans, have about a 30-fold higher liver hydrolytic activity than the mouse and a ratio of oxidation activity over hydrolytic activity about 1/10th that of the mouse. The rat has a similar ratio of these activities as that of the primates [12]. This imbalance between the toxication and the initial detoxication pathways in the mouse suggests one reason for the exquisite sensitivity of the mouse to the carcinogenic action of BD.

Metabolites as Biomarkers for BD

Biomarkers are endogenous indicators of the exposure to or the effect of exposures to foreign substances and can be useful in risk assessment. Biomarkers of exposure can help define exposures of individuals in epidemiology studies. A major study conducted in Czechoslovakian workers at a butadiene production plant compared exposures of individual worker based on in-depth occupational hygiene documentation, including personal monitors, with exposures determined by the blood or urine concentrations of selected metabolite biomarkers in the workers [13]. The biomarker that agreed best with the occupational hygiene measurements was a hemoglobin adduct formed from the epoxy diol metabolite [14].

The mercapturic acids excreted in the urine were also found to be useful as biomarkers of exposure. The urinary mercapturic acids allow one to distinguish between the mercapturic acid formed from conjugation of glutathione directly with the monoepoxide and the mercapturic acid formed by conjugation with the hydrolyzed monoepoxide (the epoxy diol). These two types of mercapturic acids can be used to provide a rough estimate of the hydrolytic detoxication activity occurring in different species. The ratios of the amount of these mercapturic acids in the urine indicate that mice have the least hydrolytic detoxication, rats are intermediate, and humans have the highest hydrolytic activity [12].

The urinary biomarkers of the formation of the different types of butadiene epoxides are of special interest. Some urinary biomarkers are compounds that are the breakdown products of the interaction of hemoglobin or DNA with the epoxides. Because the diepoxide appears to be the major genotoxic and carcinogenic metabolite of BD, a large research effort has been focused on developing a biomarker for the formation of the diepoxide. Recently a diepoxide-specific adduct (to DNA or to hemoglobin) has been identified and should be useful in determining the amount of the potent mutagen formed after exposure to BD [15]. A comparison of the biomarkers in the urine of rats, mice, and humans confirmed that much more of the diepoxide is formed in mice than in rats, and that human metabolism of BD is more similar to that of rats than that of mice [11, 12, 15]. A confirmation of the low amount of the diepoxide formed in humans is a biomarker of an effect of the BD metabolite. Exposure of human lymphocytes to the diepoxide *in vitro* results in induction of sister chromatid exchanges (SCE) in the cells, but no such effects were observed in workers exposed to 1–3 ppm BD [14]. An indication of a possible sensitive subpopulation has been found in similar *in vitro* studies. Lymphocytes from persons that do not have one of the conjugating detoxication enzymes are more sensitive to induction of SCEs following BD exposure than are lymphocytes from persons with the conjugating enzyme [14].

Toxicity Based on Epidemiology Studies

With the species dependency of the BD toxicity found in the animal studies, risk assessors turned to epidemiology studies, which determine associations between BD exposures and adverse health effects in humans. The epidemiology studies focused mainly on two groups of workers: the styrene–butadiene synthetic rubber (SBR) workers and the workers at BD production facilities. By far, the largest group was the workers in the SBR industry (approximately 18 000), and there was evidence of an association between BD exposure and an increase in mortality from leukemia (all types combined) [16, 17]. In later follow-up studies, it was confirmed that working in the SBR industry is associated with an increased risk of leukemia [18], but it has not been possible to tease out the responsible agent as being either butadiene or styrene. However, styrene has not been

associated with leukemia in workers exposed to high concentrations of styrene in other industries [16].

In the smaller group of workers producing BD (approximately 2800 workers), a group in which there was less confounding due to exposure to chemicals other than BD, deaths from lymphohematopoietic cancers (non-Hodgkin's lymphoma and to a lesser extent, leukemia) were increased [19] but no exposure–response relationship was found. In summary, current epidemiology provides reasonable evidence that work in the SBR industry is associated with increased risk of lymphohematopoietic cancer, but the causative agent(s) is not completely certain.

Regulatory Values

Various agencies have performed risk characterizations on the health effects of exposure to BD. The noncancer health risks are usually based on animal studies, while the cancer risks are based on epidemiology studies. The International Agency for Research on Cancer classifies BD as a known human carcinogen (Group 1) based on sufficient animal and human data. The US Environmental Protection Agency (EPA) classifies BD as carcinogenic to humans by inhalation with a lifetime risk of 3×10^{-5} for a $1 \mu\text{g m}^{-3}$ (0.45 ppb) exposure, based on the epidemiology studies of the SBR cohort. The same value calculated by the California EPA [20] is specified as a cancer potency factor of 1.7×10^{-4} . The United Kingdom, using an expert panel approach, concluded that there was no evidence of increased health risk in workers exposed to less than 1000 ppb and recommended 1 ppb as a protective level for the annual average for an ambient lifetime exposure [21]. The United States Occupational Safety and Health Administration has lowered its permissible exposure limit for workers from 1000 to 1 ppm (8-h time weighted average) and its short-term limit (15 min) to 5 ppm [22]. The American Conference of Industrial Hygienists lists a threshold limit value for BD as 2 ppm, which is the airborne exposure concentration for which it is believed that nearly all workers can be exposed on a daily, 8-h time weighted average basis, without adverse health effects [23].

References

- [1] USEPA (U.S. Environmental Protection Agency) (2004). *National-Scale Air Toxics Assessment (NATA)*.

- [2] Himmelstein, M.W., Acquavella, J.F., Recio, L., Medinsky, M.A. & Bond, J.A. (1997). Toxicology and epidemiology of 1,3-butadiene, *Critical Reviews in Toxicology* **27**, 1–108.
- [3] Morrissey, R.E., Schwetz, B.A., Hackett, P.L., Sikov, M.R., Hardin, B.D., McClanahan, B.J., Decker, J.R. & Mast, T.J. (1990). Overview of reproductive and developmental toxicity studies of 1,3-butadiene in rodents, *Environmental Health Perspectives* **86**, 79.
- [4] Huff, J.E., Melnick, R.L., Solleveld, H.A., Haseman, J.K., Powers, M. & Miller, R.A. (1985). Multiple organocarcinogenicity of 1,3-butadiene in B6C3F1 mice after 60 weeks of inhalation exposure, *Science* **227**, 548–549.
- [5] Melnick, R.L., Huff, J., Chou, B.J. & Miller, R.A. (1990). Carcinogenicity of 1,3-butadiene in B6C3F1 mice at low exposure concentrations, *Cancer Research* **50**, 6592–6599.
- [6] Owen, P.E., Glaister, J.R., Gaunt, I.F. & Pullinger, D.H. (1987). Inhalation toxicity studies with 1,3-butadiene III. Two year toxicity/carcinogenicity studies in rats, *American Industrial Hygiene Association Journal* **48**, 407–413.
- [7] Crouch, C.N., Pullinger, D.H. & Gaunt, I.F. (1979). Inhalation toxicity studies with 1,3-butadiene II. 3 month toxicity studies in rats, *American Industrial Hygiene Association Journal* **40**, 796.
- [8] Thurmond, L.M., Lauer, L.D., House, R.V., Stillman, W.S., Irons, R.D., Steinhage, W.H. & Dean, J.H. (1986). Effect of short-term inhalation exposure to 1,3-butadiene on murine immune function, *Toxicology and Applied Pharmacology* **86**, 170.
- [9] Irons, R.D., Smith, C.N., Stillman, W.S., Shah, R.S., Steinhagen, W.H. & Leiderman, L.J. (1986). Macrocytic-megaloblastic anemia in male B6C3F1 mice following chronic exposure to 1,3-butadiene, *Toxicology and Applied Pharmacology* **83**, 95–100.
- [10] Cochrane, J.E. & Skipek, T.R. (1994). Mutagenicity of butadiene and its epoxide metabolites: I. Mutagenic potential of 1,2-epoxybutene, 1,2,3,4-diepoxybutane, and 3,4,1,2-butanediol in cultured human lymphoblasts, *Carcinogenesis* **15**, 713–717.
- [11] Henderson, R.F., Thornton-Manning, J.R., Bechtold, W.E. & Dahl, A.R. (1996). Metabolism of 1,3-butadiene: species differences, *Toxicology* **113**, 17–22.
- [12] Henderson, R.F. (2001). Species differences in the metabolism of olefins: implications for risk assessment, *Chemico-Biological Interactions* **135–136**, 53–64.
- [13] Albertini, R.J., Sram, R.J., Vacek, P.M., Lynch, J., Nicklas, J.A., van Sittert, N.J., Boogaard, P.J., Henderson, R.F., Swenberg, J.A., Tate, A.D., Ward Jr, J.B., Wright, M., Ammenheuser, M.M., Binkova, B., Blackwell, W., de Zwart, F.A., Krako, D., Krone, J., Megens, H., Musilova, P., Rajska, G., Ranasinghe, A., Rosenblatt, J.I., Rossner, P., Rubes, J., Sullivan, L., Upton, P. & Zwinderman, A.H. (2003). *Biomarkers in Czech Workers Exposed to 1,3-Butadiene: A Transitional Epidemiologic Study*, Research Report Health Effects Institute, pp. 1–141, No. 116.
- [14] Swenberg, J.E., Koc, H., Upton, P.B., Georgieva, N., Ranasinghe, A.M., Walker, V.E. & Henderson, R.F. (2001). Using DNA and hemoglobin adducts to improve the risk assessment of butadiene, *Chemico-Biological Interactions* **135–136**, 387–403.
- [15] Boysen, G., Georgieva, N.I., Upton, P.B., Walker, V.E. & Swenberg, J.A. (2007). N-terminal globin adducts as biomarkers for formation of butadiene derived epoxides, *Chemico-Biological Interactions* **166**, 84–92.
- [16] Delzell, E., Sathikumar, N., Hovinga, M., Macaluso, M., Julian, J., Larson, R., Cole, P. & Muir, D.C. (1996). A follow-up study of synthetic rubber workers, *Toxicology* **113**(1–3), 182–189.
- [17] Macalus, M., Larson, R., Delzell, E., Sathikumar, N., Hovinga, M., Julian, J., Muir, D. & Cole, P. (1996). Leukemia and cumulative exposure to butadiene, styrene and benzene among workers in the synthetic rubber industry, *Toxicology* **113**(1–3), 190–202.
- [18] Sathikumar, N., Graff, J., Macaluso, M., Maldonado, G., Matthews, R. & Delzell, E. (2005). An updated study of mortality among North American synthetic rubber industry workers, *Occupational and Environmental Medicine* **62**(12), 822–829.
- [19] Divine, B.J. & Hartman, C.M. (1996). Mortality update of butadiene production workers, *Toxicology* **113**(1–3), 169–181.
- [20] California EPA (1992). *Hot Spots Unit Risk and Cancer Potency Values*, <http://www.oehha.ca.gov> (accessed Nov 2005).
- [21] Expert Panel on Air Quality Standards (2002). *Second Report on 1,3-butadiene*. DEFRA, London.
- [22] Department of Labor, Occupational Safety and Health Administration (1996). *29 CFR Part 1910. Occupational Exposure to 1,3-Butadiene; Final Rule*, Federal Register, November 4, 1996.
- [23] American Conference of Governmental Industrial Hygienists (2005). *Threshold Limit Values for Chemical Substances and Physical Agents*, Cincinnati.

Related Articles

Cancer Risk Evaluation from Animal Studies

ROGENE F. HENDERSON

Cancer Risk Evaluation from Animal Studies

Just as animal models are used in preclinical trials of pharmaceutical agents before testing is begun in humans, studies of chemicals found in the workplace or in the general environment are conducted in experimental animals to assess potential human health risks. In addition to identifying adverse responses prior to human trials, animal studies are performed in drug development programs to estimate and compare therapeutic doses with those doses that induce adverse reactions. The ratio between the toxic dose and the therapeutic dose, called the *therapeutic index*, is a measure of the relative safety of a drug for a specific treatment (*see Toxicity (Adverse Events)*). Adverse health concerns are minimized for drugs that produce therapeutic effects at doses much lower than those that cause toxic effects. Drugs with a low therapeutic index require much more monitoring to achieve the therapeutic effect with minimal toxicity. In contrast to pharmaceuticals, a therapeutic value is not a recognized benefit from environmental or occupational exposure to toxic chemicals and carcinogens (*see Environmental Carcinogenesis Risk*). Exposure standards for such agents are typically based on the nature of the induced health effect and its potency. The likelihood of an adverse effect occurring in humans is often reduced when actual environmental or occupational exposures are substantially lower than those that induce toxic effects in animal models.

The utility of experimental animal studies for predicting human disease is based on similarities between species in biological processes such as absorption and metabolism of xenobiotic agents and in disease induction mechanisms. Because it is unethical to test for adverse health effects, such as cancer, in humans through intentional exposures, animal models are widely used to screen and evaluate agents for potential human health risks. Also, animal studies have other advantages compared to human studies. For example, animal studies eliminate the need to wait for a high incidence of human cancers, which may take as long as 30 years beginning from time of first exposure to clinical manifestation of disease [1], before implementing public health protective strategies. In addition, because exposure

conditions can be finely controlled in animal studies, they are easier to interpret, assign causality, and estimate response potency. In contrast, epidemiological studies of occupational or environmental human cancer risks typically have limited sensitivity and limited exposure information, especially at times early in tumor development, and when confounding factors are not always known or accounted for [2]. The major disadvantages of animal studies are that they require extrapolations across species and dose. In addition, animal studies do not capture the full range of interindividual variability in disease susceptibility due to differences in genetic factors, health status, diet, life style, and exposures to other agents.

Most national and international public health agencies, including the International Agency for Research on Cancer (IARC) [1], the National Toxicology Program (NTP) [3], and the United States Environmental Protection Agency (US EPA), [4], utilize data from animal studies to make decisions on cancer risk in humans. The perspective adopted by these agencies that in the absence of adequate data in humans, “it is biologically plausible that agents for which there is sufficient evidence of carcinogenicity in experimental animals also present a carcinogenic hazard to humans” is based on the fact that all known human carcinogens that have been studied adequately in experimental animals produce positive carcinogenic results [1]. Hence, even in the absence of any human data, public health agencies classify agents as possibly/probably or likely to be carcinogenic in humans [1] or reasonably anticipated to be a human carcinogen [3] if there is sufficient evidence in animals demonstrating (a) increased incidence of malignant or combined malignant and benign tumors in two or more species or at multiple sites; or (b) increased incidence in two or more independent studies in one species; or (c) increased incidence in a single study in one species if malignant tumors occur to an unusual degree in incidence, site, type, or age of onset. Several agents that were considered to be possible human carcinogens based on animal data were later confirmed as human carcinogens when reliable epidemiology data (usually occupational exposures) became available; among others, these include 1,3-butadiene, cadmium, diethylstilbestrol, formaldehyde, ethylene oxide, and vinyl chloride [5].

This chapter focuses on the following: (a) design of carcinogenicity studies on environmental agents

in laboratory animals, (b) evaluation of experimental carcinogenicity studies, and (c) assessment of human cancer risk based on animal tumor data.

Design of Carcinogenicity Studies on Environmental Agents in Laboratory Animals

The design of experimental carcinogenicity studies is critical for the identification of adverse health effects caused by specific environmental agents (*see Environmental Hazard*) and the characterization of dose–response relationships (*see Dose–Response Analysis*). For example, due to design deficiencies, including small group size, lack of controls, short study duration, and low levels of exposure, early studies on *benzene* (*see Benzene*) in animals, failed to detect carcinogenic effects, even though epidemiology studies had demonstrated a causal association between benzene exposure and leukemia in humans. Subsequent studies that were better designed established benzene as a potent, multisite carcinogen in rats and mice [6, 7].

The Test Chemical

In order to ensure that the agent under study is responsible for any observed effects and that contaminants are not the cause or modifier of that response, the chemical should be tested at high purity. However, testing of technical grades of certain chemicals may be done to mimic human exposures. Prior to exposing animals to the agent, it is necessary to demonstrate its purity and its stability under the conditions of exposure and storage. If the agent degrades or evaporates during exposure, the accuracy of the targeted administered dose is compromised and degradation products may contribute to any observed response.

Animal Models

Rats and mice are the two species most frequently used in cancer bioassays because (a) they are responsive to tumor induction resulting from exposure to carcinogens, (b) they have life spans of about 2.5 years, and (c) studies of up to 1000 animals can be performed in reasonably sized animal rooms. The selected strains of animal models

should have good longevity, genetic stability, and few spontaneous diseases that might shorten their life span, mask any chemical-induced effects, or impair metabolism/elimination of the test agent [8]. Importantly, it is usually difficult to detect a chemically induced response in an organ with a high natural or background tumor rate. Both sexes of two species are typically used to identify any sex-specific responses and to confirm multiple species effects. Although most carcinogenicity studies are conducted in few strains of rats and mice, suggestions have been made to use multiple strains of inbred mice to increase genetic diversity in these bioassays [9]. However, the ability to detect true effects in multiple strains *versus* single strains is a function of the heterogeneity of response across the strains, the strength of response in the sensitive strains, and the inherent rate of the specific tumor types in the various strains [10].

Exposure Duration

Typical carcinogenicity studies in rats and mice involve exposures beginning at 6 weeks of age and continuing for 2 years; this exposure period corresponds roughly with early adulthood through most of an occupational life span [11]. However, because cancer susceptibility may be greater with exposure to mutagens and endocrine disrupting chemicals during growth and early developmental stages, earlier periods of exposure should be included in animal carcinogenicity studies when fetal or childhood exposure to such agents might occur. The 2-year duration limit was selected to minimize late-developing background tumor responses that might preclude the detection of chemical-induced effects [12] and to reflect an occupational life span. Care must be taken, however, to select strains that have reasonable 2-year survival rates. Studies with exposure durations shorter than 2 years are also problematic because of their reduced sensitivity to detect increases in late-appearing tumors that are related to treatment with the test agent [11].

Dose Selection

A major shortcoming of the rodent cancer bioassay is its limited statistical power to detect small to moderate responses [13]. Power is the probability of detecting an effect (rejecting the null hypothesis)

when an effect exists; it is influenced by the sample size, the background rate, and the magnitude of the true response [14]. For example, in order to achieve statistical significance at $p \leq 0.05$ level (the chance of rejecting the null hypothesis when it is true is 5% or less) in a bioassay consisting of 50 animals in each of the three dose groups, a tumor incidence of 38% (19/50) or greater would be necessary in the high dose group if the incidence in the control group was 20% (10/50). If animal group size were only 25, then an incidence of 40% (10/25) would not be statistically significant if the control incidence was 20% (5/25). A negative finding in an underpowered study provides no assurance of the absence of health risks in exposed populations.

Because of the limited statistical power of the cancer bioassay when group size is only about 50 animals per sex dose, higher doses are typically used to identify potential carcinogenic hazards and multiple dose groups are used to characterize dose–response relationships [15, 16]. If carcinogenicity studies were conducted at actual environmental exposure levels, extremely large group sizes (i.e., several hundred to thousands of animals per group) would be needed in order to avoid false negative results.

In order to estimate the maximally tolerated dose or the minimally toxic dose (MTD) for use in the cancer bioassay, data are generated in preliminary prechronic studies typically of 4–13 weeks duration and evaluated for effects on body weight, survival, clinical changes, and incidence and severity of pathologic lesions. The MTD should not cause increases in mortality other than from chemically induced tumors. Therefore, doses that induce life-threatening toxic lesions in a prechronic study would be excluded from use in a cancer bioassay. The identification and use of an MTD is a critical aspect in the design of experimental carcinogenicity studies of low statistical power. Lower doses (1/2 MTD and 1/4 to 1/10 MTD) are used in case the highest dose selected for the chronic study is found to be too high (excessive mortality) and to provide data for dose–response analyses. Pharmacokinetic information is also used to ensure that no more than one of the selected doses is above a level that approaches saturation of absorption, metabolic activation, or detoxification activities. A much better characterization of the true dose–response relationship can be achieved with several dose groups rather than only two widely spaced dose groups.

Evaluation of Experimental Carcinogenicity Studies

The conduct and evaluation of a cancer bioassay requires a multidisciplinary effort, including expertise from toxicologists, laboratory animal veterinarians, chemists, histologists, pathologists, cellular/molecular biologists, and statisticians. The interpretation of tumor data may be affected by the thoroughness of the histopathological evaluations, methods of statistical analysis, and mechanistic considerations.

Histopathology

The detection of toxic effects and neoplastic lesions in animals depends on the thoroughness of the necropsies and the histopathologic examinations performed. The US NTP requires microscopic examination of all organs and tissues, approximately 40 per animal [17]. In some cases, multiple sectioning of an organ in exposed and control groups may be necessary to obtain a truer estimate of the incidence of neoplastic lesions, especially for small lesions that may not be detected at necropsy [18]. Although diagnostic criteria have been established for most observable lesions, it is not unusual for pathologists to disagree in their judgment of lesions, especially those that are part of a continuum of progressive change. After diagnoses by the study pathologist, all NTP studies undergo an independent quality assessment (QA) pathology review. This is followed by a pathology working group review (typically 8–10 pathologists) that seeks to resolve discrepancies in diagnoses between the original and the QA pathologists [19]. Multiple independent evaluations provide greater confidence on the reliability of the histopathology data.

In addition to statistical analyses described below, several additional factors may contribute to the interpretation of tumor data, including: (a) the occurrence of uncommon *versus* common tumors, (b) evidence of progression of lesions, such as benign to malignant where it is appropriate to combine [20], (c) tumor occurrence with reduced latency, (d) multiplicity in a site-specific tumor response, (e) evidence of metastases, and (f) supporting evidence of proliferative preneoplastic lesions at the same site or detection of the same lesion in the other sex or species. Because cancer development is a multistep process with a long time period between exposure and

manifestation of metastatic neoplasia, evidence of enhanced disease progression (e.g., reduced latency, tumor multiplicity, metastasis) in an animal study is another indication that the agent promotes cancer development.

Statistics

Statistical analyses of tumor data include pair-wise comparisons of tumor incidence in control groups *versus* each treatment group and analysis of exposure related trends. Tumor incidence is the number of animals with a particular neoplasm at a specific anatomic site divided by the number of animals in which that site was examined. If mortality in a dose group is different from that of controls, it is important that these analyses be based on survival-adjusted tumor rates [14]. The reason for this adjustment is that if animals died early from causes other than tumors at the organ of interest, then those animals would not have been on study long enough to provide a contribution of risk equivalent to other groups having better survival. For example, the poly-k quantal response method adjusts site-specific tumor rates within each treatment group for the number of animal years at risk due to differences in intercurrent mortality [21]. Failure to adjust for differences in survival can result in misinterpretations of a true site-specific effect and unreliable estimates of cancer risk.

Although comparisons between the concurrent control group and the exposure groups are the most appropriate for identifying chemical-induced effects, comparisons with historical control data may also be helpful in interpreting whether or not observed differences are treatment-related, e.g., the observation of a small number of rarely occurring tumors in treated animals may not reach statistical significance, yet may exceed historical control rates [22]. For meaningful comparisons, the conditions of the current study must be similar to those in the historical database and diagnostic criteria must be identical. Thus, comparisons must be specific for the species, sex, and strain of animals, the route of exposure, and the diet. A survival-adjusted statistical test for evaluating dose-response trends based on historical control data has been developed to facilitate the interpretation of studies, including those in which exposures are associated with small numbers of rare tumors [23].

Mechanistic Considerations

Results from animal or *in vitro* studies that address molecular and cellular processes involved in tissue responses to environmental compounds (i.e., mechanistic studies) have been used to upgrade or downgrade cancer risk classifications of agents that have inadequate or limited evidence of carcinogenicity in humans. For example, based on mechanistic data, IARC and NTP upgraded ethylene oxide [3, 24] and 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) [3, 25] to “known human carcinogens”. In both cases, the evidence of carcinogenicity was limited in humans but sufficient in experimental animals. The upgrading of ethylene oxide was based largely on induction of chromosomal aberrations in peripheral lymphocytes, micronuclei in bone marrow cells, and hemoglobin adducts in exposed workers. Chromosomal alterations are associated with various stages in the multistep process of carcinogenesis. For TCDD, the upgrading was based on data demonstrating that the multisite carcinogenicity of this chemical in experimental animals was due to a mechanism involving activation of the aryl hydrocarbon receptor and studies showing that this receptor is highly conserved across species, and functions similarly in humans and experimental animals. Thus, even though human epidemiological data were at that time not sufficient alone to reach the “known human carcinogen” classification category, these authoritative bodies considered the combination of animal data and mechanistic data sufficient to support the upgraded “weight-of-evidence” conclusions. The term *weight of evidence* is used by the US EPA [4] to reflect evaluations that are “based on the combined strength and coherence of inferences appropriately drawn from all of the available information”.

Similarly, vinyl bromide and vinyl fluoride should be treated as “known human carcinogens” since they act in the same way as vinyl chloride, an agent that has been shown to be causally linked with increased cancer risk in humans. These three chemicals are metabolized by enzymes found in humans to reactive intermediates that form identical promutagenic DNA adducts [26] and experimental carcinogenicity studies have shown that all three chemicals induce multiple tumor types including liver angiosarcomas in rats and mice [27].

Estimating Human Cancer Risk from Animal Tumor Data

Risk assessment provides a systematic approach for characterizing the nature and probability of adverse effects (i.e., health risks) occurring in individuals or populations exposed to hazardous agents and often serves as the basis for risk-management decisions on whether and to what extent human exposure to such agents should be controlled (*see* **Environmental Health Risk**). The National Academy of Sciences (NAS)/National Research Council [28] developed guidelines for the conduct of risk assessments in the US Federal Government. The risk assessment paradigm developed by the NAS consists of four parts: hazard identification, dose–response assessment, exposure assessment, and risk characterization.

Dose–response analyses (*see* **Dose–Response Analysis**) provide a basis for determining *causality* (*see* **Causality/Causation**) from experimental cancer data, evaluating the consistency of mechanistic hypotheses, and estimating low-dose risk. When high-quality human data are available, such as from occupational cohorts, they are preferentially used to estimate human cancer risk at environmental exposure levels and set occupational and environmental exposure standards [4].

When adequate human data are not available, dose–response tumor data from studies in laboratory animals serve as the basis for estimating cancer risks in exposed humans [4]. Site-specific survival-adjusted tumor rates from animal studies are assumed to be relevant for determining human cancer risk unless a mode-of-action has been unequivocally established that contradicts the acceptance of that assumption. A quantitative risk assessment requires conversion of animal doses to equivalent human doses. If a verified physiology based pharmacokinetic model is available for the specific agent, this might be used to describe the internalized dose of the agent or the time-dependent tissue specific concentration of the parent compound or its metabolites. Such a model may be used to characterize internal dose metrics in humans by replacing physiological and biochemical parameters in the animal model with those specific for humans (e.g., breathing volumes, organ sizes, cardiac output, metabolic rate constants, etc.). Computer-based dosimetry models consist of a series of mathematical equations that represent, in quantitative terms, the complex

biological processes that affect the absorption, distribution, metabolism, and elimination of the agent in the intact animal. These models can accommodate parameter values that represent the range and distribution of activities that exist in human populations. Reliable models can provide estimates of tissue dose as a function of duration and frequency of exposure to various environmental exposure levels [29]. However, one must be cautious of assumptions in models that have not been validated because they could lead to inaccurate estimates of tissue dose in specific populations. If a verified dosimetry model is not available, human equivalent exposures are typically obtained by scaling animal doses to humans as a function of body surface area (i.e., body weight^{0.75}) [4].

The next step in the risk assessment process is to use the data on animal cancer to estimate cancer risk at human exposure levels. This is done by fitting empirical dose–response models to the tumor incidence data and an appropriate human dose metric (e.g., a physiology based model estimate of the daily internalized concentrations of reactive metabolites or the rodent scaled equivalent human dose) to extend the dose–response relationship to low doses or to estimate response rates at the low end of the observed dose–response range (i.e., in the 1–10% range of response). Alternatively, a mechanistic-based model may be used if there is sufficient data to establish the mechanism of tumor induction and validated model parameters are available for the key kinetic events associated with that mechanism. For studies with large differences in mortality rates among dose groups, a time-to-tumor model should be used if incidence values were not corrected for survival differences.

By extending the dose–response curve, estimates can be made of extra cancer risk at any dose, including human exposure levels, or estimates can be made of exposure levels that are associated with specific risks (e.g., 1/100, 1/1000, and 1/1 000 000). In their most recent cancer risk assessment guidelines, the US EPA [4] recommends modeling the dose–response data to determine the central estimate of the effective dose associated with 1% (ED₀₁) to 10% (ED₁₀) risk and the associated lower 95% confidence limits on those dose estimates (LED₀₁ or LED₁₀). For example, a modified Weibull model (equation 1) was fit to survival-adjusted tumor data and concentrations of chloroprene or 1,3-butadiene to which mice

were exposed in chronic studies to characterize the shape of the dose–response curve and estimate ED_{10} values and the 95% upper and lower confidence limits [30].

$$P(\text{dose}) = 1 - e^{-(\text{intercept} + \text{scale} \times \text{dose}^{\text{shape}})} \quad (1)$$

$P(\text{dose})$ is the probability of a neoplasm prior to study termination for animals administered a defined dose. The parameters intercept, scale, and shape are estimated *via* maximum-likelihood estimation using a binomial likelihood based on the survival-adjusted tumor data. ED_{10} values obtained from the dose–response curves represent exposure concentrations associated with an excess cancer risk of 10% at each site. A likelihood-ratio test is used to test the hypothesis that the shape parameter equals one. If the shape parameter is equal to one, the dose–response is consistent with a linear model; if the shape parameter is greater than one, the dose–response exhibits sublinear behavior.

A graphical example of this modified Weibull curve with the ED_{10} and the LED_{10} for oral cavity tumors in male rats exposed to chloroprene is given in Figure 1. In this example, the shape parameter is not significantly different from one, so the dose–response curve is fairly linear. The estimated ED_{10} is 21.0 ppm and the lower endpoint of its 95% confidence interval is 6.0 ppm.

From the LED_{10} or the LED_{01} value, corrected for background, a straight line is drawn to the origin (zero risk) for mutagenic carcinogens or chemicals for which the mode-of-action has not been established [4]. Cancer risks at lower exposure levels or doses associated with a specific risk level (e.g., 1/1 000 000 increased lifetime risk) are estimated from the slope of this line (e.g., $0.01/LED_{01}$). These risks can then be compared to age, sex, and race-dependent incidences of specific cancers in the US population reported in the Surveillance, Epidemiology, and End Results data set [31] to determine the relative risk of developing or dying from cancer due to a particular exposure. Relative risk is the incidence of disease in an exposed population divided by the incidence of that disease in the unexposed population. For example, consider that a low-dose extrapolation of an animal carcinogenicity study indicates that a particular exposure level of an agent is associated with an additional leukemia risk of 9/1000 at age 70. The probability of women in the general US population dying from leukemia at that age is 6.6/1000, then the relative risk of dying from leukemia due to that exposure is 2.4. In other words, 70-year old women who received that exposure have 2.4 times the probability of dying from leukemia than do 70-year old women who were unexposed. [The estimated incidence in the unexposed population (P_0) is 6.6/1000, while the estimated

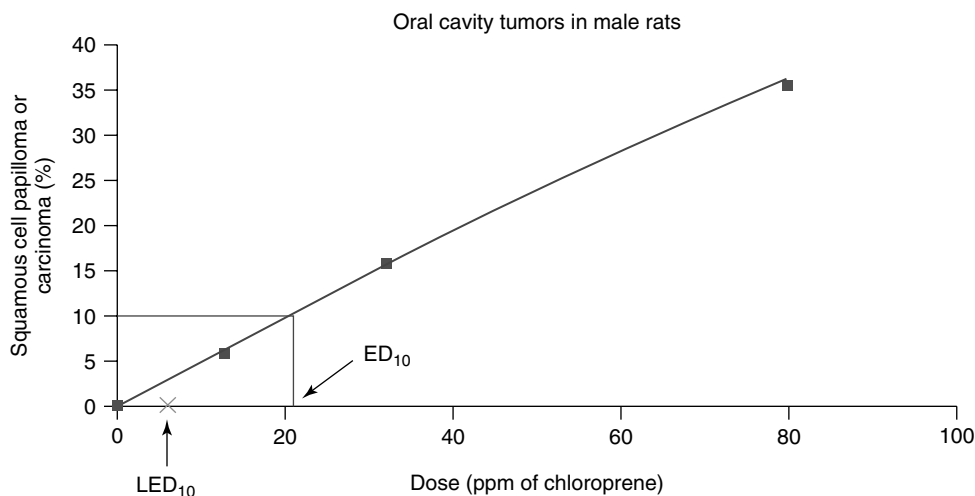


Figure 1 Weibull curve for oral cavity tumors in male rats exposed to chloroprene showing locations of the ED_{10} and LED_{10} [Reproduced from [30]. © Oxford University Press, 1999.]

incidence in the exposed population (P_d) is the background risk (6.6/1000) plus the additional risk associated with exposure (9/1000). Relative risk = $P_d/P_0 = (0.0066 + 0.009)/0.0066 = 2.4$.]

Adjustments to the LED_{10} (or LED_{01}) derived slope factors may be necessary if the data used for the risk assessment do not adequately capture exposures associated with differential susceptibilities. For example, risk levels derived from a study of male workers would not be adequate for female workers if the exposure involves an agent such as 1,3-butadiene that is associated with increased breast cancer risk. To deal with potentially higher lifetime cancer risks due to early life exposures, the US EPA issued a supplementary guidance [32] for adjusting cancer potency estimates for mutagenic carcinogens when data are not available to assess susceptibility from early life exposures to such agents. Based on combined analyses of multiple animal and epidemiological data sets, the US EPA recommends default adjustment factors of 10-fold for exposures occurring before 2 years of age and 3-fold for exposures between 2 and 16 years of age; no adjustments were recommended for exposures occurring after 16 years of age. However, when chemical specific data are available they should be used in preference to default adjustments. Data for carcinogens that act by a nonmutagenic mode-of-action were considered to be too limited and the mechanisms too diverse to support the establishment of default cancer potency adjustment factors for early life exposures to such agents.

Site Concordance

There are several examples in which human cancer sites correspond with one animal species, but not the second species (e.g., hematopoietic cancers are induced by benzene in mice and humans, but not in rats) and there are also examples of site correspondence among rats, mice, and humans [33]. However, because the etiologies of most cancers are not known and the basis for differences in species susceptibility are not fully understood, it would not be prudent to limit considerations of human cancer risk only to organs or tissues in which tumors arise in animal studies.

There are several reasons why site correspondence should not be a requirement for assessing human relevance and causality. First, exposure factors might contribute to differences in sites of tumor induction;

some of these include the route, frequency, duration, intensity, and age at onset of exposure. Exposure conditions are known to affect sites and dose-response for tumor induction in experimental studies. For example, gestational or early childhood exposure is frequently important in tumor induction; consequently, animal or human studies that do not capture these early stages of life may not identify all susceptible organ sites or may underestimate the true carcinogenic risk of the agent. As noted above, for mutagenic agents the US EPA [32] recommends adjustment in risk estimates for studies that did not include childhood exposures. Second, human susceptibility to environmental carcinogens varies for a large number of reasons including genetic factors, health status, diet, lifestyle (e.g., smoking, alcohol consumption), age, and other exposures (e.g., medications, occupational experiences). Because cancer development is the likely result of interactions among environmental factors and individual susceptibility factors, differences in sites of tumor induction among individuals are not unusual for known human carcinogens. For example, although the lung is the most common cancer site among cigarette smokers, many people do not develop lung tumors but instead develop cancers of the bladder, kidney, nasal cavity, lip, esophagus, or pancreas, and some smokers do not develop any diagnosed cancers during their lifetime. Numerous host susceptibility factors could account for lack of site correspondence among individuals in exposed populations or between experimental animals and humans. Because of the much greater genetic diversity in humans compared to strains of laboratory animals, the range of expected human response may include subgroups that are less, equal, or more sensitive than the animal models used in the experimental cancer studies.

Conclusions

Data from properly designed and evaluated studies in experimental animals have been and continue to be reliable sources of information for the identification of potential human health hazards and the estimation of risks in exposed populations. Properly designed animal studies must include: (a) animal models that are sensitive to the endpoints under investigation, (b) detailed characterization of the agent and the administered doses, (c) adequately challenging doses

(MTD) and durations of exposure (at least 2 years for rats and mice) that would allow a reliable determination of whether or not the agent poses a health hazard, (d) sufficient numbers of animals per dose group to have adequate statistical power to detect a true effect, (e) multiple dose groups to allow characterization of dose–response relationships, (f) complete and peer-reviewed histopathological diagnoses, and (g) pairwise comparisons with the control group and analyses of trends based on survival-adjusted tumor rates. Mechanistic information and pharmacokinetic models that have been adequately tested may impact on the characterization of dose–response relationships. Setting occupational or environmental exposure standards based on animal carcinogenicity data provides a prudent public health approach to reduce the burden of cancer in exposed populations. More work is necessary to understand the basis for interindividual differences in susceptibility and to properly establish health-based exposure standards that protect the most vulnerable segments of the general population. Until the etiology of environmentally induced cancers and the basis for human susceptibility are better understood, site correspondence should not be a requirement for judging human relevance and causality.

Acknowledgment

This work was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Institute of Environmental Health Sciences (NIEHS).

References

- [1] International Agency for Research on Cancer (IARC) (2006). *Preamble to the IARC Monographs* <http://monographs.iarc.fr/ENG/Preamble/CurrentPreamble.pdf>.
- [2] Stayner, L.T. & Smith, R.J. (1992). Methodologic issues in using epidemiologic studies of occupational cohorts for cancer risk assessment, *Epidemiologia e Prevenzione* **14**, 32–39.
- [3] National Toxicology Program. Report on Carcinogens (2004). *Carcinogen Profiles 2004*, 11th Edition, US Department of Health and Human Services, Research Triangle Park.
- [4] US EPA (2005). *Guidelines for Carcinogen Risk Assessment*, EPA/630/P-03/001F, US Environmental Protection Agency, Washington, DC.
- [5] Huff, J. (1993). Chemicals and cancer in humans: first evidence in experimental animals, *Environmental Health Perspectives* **100**, 201–210.
- [6] Maltoni, C., Ciliberti, A., Cotti, G., Conti, B. & Bel-poggi, F. (1989). Benzene, an experimental multipotential carcinogen: results of the long-term bioassays performed at the Bologna Institute of Oncology, *Environmental Health Perspectives* **82**, 109–124.
- [7] Huff, J.E., Haseman, J.K., DeMarini, D.M., Eustis, S., Maronpot, R.R., Peters, A.C., Persing, R.L., Chrisp, C.E. & Jacobs, A.C. (1989). Multiple-site carcinogenicity of benzene in Fischer 344 rats and B6C3F1 mice, *Environmental Health Perspectives* **82**, 125–163.
- [8] Rao, G.N., Birnbaum, L.S., Collins, J.J., Tennant, R.W. & Skow, L.C. (1988). Mouse strains for chemical carcinogenicity studies: overview of a workshop, *Fundamental and Applied Toxicology* **10**, 385–394.
- [9] Festing, M.F. (1995). Use of a multistrain assay could improve the NTP carcinogenesis bioassay, *Environmental Health Perspectives* **103**, 44–52.
- [10] Kissling, G.E. (2005). Power calculations for multiple-strain designs, Presented at the *NTP Workshop: Animal Models for the NTP Rodent Cancer Bioassay: Strains & Stocks – Should We Switch?*, Research Triangle Park, June 16–17, 2005.
- [11] Haseman, J., Melnick, R., Tomatis, L. & Huff, J. (2001). Carcinogenesis bioassays: study duration and biological relevance, *Food and Chemical Toxicology* **39**, 739–744.
- [12] Solleveld, H.A., Haseman, J.K. & McConnell, E.E. (1984). Natural history of body weight gain, survival, and neoplasia in the F344 rat, *Journal of the National Cancer Institute* **72**, 929–940.
- [13] Haseman, J.K. (1983). Statistical support of the proposed National Toxicology Program protocol, *Toxicologic Pathology* **11**, 77–82.
- [14] Haseman, J.K. (1984). Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies, *Environmental Health Perspectives* **58**, 385–392.
- [15] Griesemer, R.A. (1992). Dose selection for animal carcinogenicity studies: a practitioner's perspective, *Chemical Research in Toxicology* **5**, 737–741.
- [16] Bucher, J.R. (2002). The National Toxicology Program rodent bioassay: designs, interpretations and scientific contributions, *Annals of the New York Academy of Science* **982**, 198–207.
- [17] National Toxicology Program (NTP) (2005). *11th Report on Carcinogens 2004*, at http://ntp.niehs.nih.gov/files/SPECIFICATIONS_2006Oct.pdf.
- [18] Eustis, S.L., Hailey, J.R., Boorman, G.A. & Haseman, J.K. (1994). The utility of multiple-section sampling in the histopathological evaluation of the kidney for carcinogenicity studies, *Toxicologic Pathology* **22**, 457–472.
- [19] Boorman, G.A., Montgomery Jr C.A., Eustis, S.L., Wolfe, M.J., McConnell, E.E. & Hardisty J.F. (1985). Quality assurance in pathology for rodent carcinogenicity studies, in *Handbook of Carcinogen Testing*, H. Milman, E. Weisburger, eds, Noyes Publications, Park Ridge, pp. 345–357.

- [20] McConnell, E.E., Solleveld, H.A., Swenberg, J.A. & Boorman, G.A. (1986). Guidelines for combining neoplasms for evaluation of rodent carcinogenesis studies, *Journal of the National Cancer Institute* **76**, 283–289.
- [21] Portier, C.J. & Bailer, A.J. (1989). Testing for increased carcinogenicity using a survival-adjusted quantal response test, *Fundamental and Applied Toxicology* **12**, 731–737.
- [22] Haseman, J.K., Huff, J. & Boorman, G.A. (1984). Use of historical control data in carcinogenicity studies in rodents, *Toxicologic Pathology* **12**, 126–135.
- [23] Peddada, S., Dinse, E. & Kissling, G.E. (2007). Incorporating historical data when comparing tumor incidence rates, *Journal of the American Statistical Association* (in press).
- [24] International Agency for Research on Cancer (1994). Ethylene oxide, *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans* **60**, 73–159.
- [25] International Agency for Research on Cancer (1997). 2,3,7,8-Tetrachlorodibenzo-para-dioxin, *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans* **69**, 33–343.
- [26] Guengerich, F.P., Kim, D.H. & Iwasaki, M. (1991). Role of human cytochrome P-450 IIE1 in the oxidation of many low molecular weight cancer suspects, *Chemical Research in Toxicology* **4**, 168–179.
- [27] Melnick, R.L. (2002). Carcinogenicity and mechanistic insights on the behavior of epoxides and epoxide-forming chemicals, *Annals of the New York Academy of Science* **982**, 177–189.
- [28] National Research Council (NRC) (1983). *Risk Assessment in the Federal Government: Managing the Process*, Committee on the Institutional Means for Assessment of Risk to Public Health, Commission on Life Sciences, NRC, National Academy Press, Washington, DC.
- [29] Kohn, M.C. (1995). Achieving credibility in risk assessment models, *Toxicology Letters* **79**, 107–114.
- [30] Melnick, R.L., Sills, R.C., Portier, C.J., Roycroft, J.H., Chou, B.J., Grumbein, S.L. & Miller, R.A. (1999). Multiple organ carcinogenicity of inhaled chloroprene (2-chloro-1,3-butadiene) in F344/N rats and B6C3F1 mice and comparison of dose-response with 1,3-butadiene in mice, *Carcinogenesis* **20**, 867–878.
- [31] Surveillance, Epidemiology and End Results (SEER) (2007). *Surveillance, Epidemiology and End Results (SEER)*, at <http://seer.cancer.gov/>.
- [32] US, E.P.A. (2005). *Supplemental Guidance for Assessing Susceptibility from Early-Life Exposure to Carcinogens*, EPA/630/R-03/003F, US Environmental Protection Agency, Washington, DC.
- [33] Melnick, R.L. & Huff, J.E. (1993). 1,3-Butadiene induces cancer in experimental animals at all concentrations from 6.25 to 8000 parts per million, *IARC Scientific Publications* **127**, 309–322.

RONALD L. MELNICK AND GRACE E.
KISSLING

Canonical Modeling

Assessing Risk in Complex Systems

Under fortuitous conditions quantitative risk assessments may be approached with methods of probability theory. For example, to predict or diagnose the failure risk of a complicated entity such as a nuclear power plant, it is possible to construct fault trees (*see* **Fault Detection and Diagnosis**) or event trees that dissect a complex process into manageable modules (e.g., loss of valve function, a pipe break, a human operator falling asleep) whose failure probabilities can be computed with some reliability from a combination of the physics of the modules, factual knowledge of engineering principles, and expert heuristics [1]. The probabilities of these unit failures become the basis for the computation of all possible conditional probabilities, which are appropriately summed for an assessment of the likelihood of occurrence of the undesired event. The key assumption of this strategy is that all possible component failures and human errors can be quantified with some confidence (*see* **Human Reliability Assessment**). As soon as complicated environmental or health risk scenarios are to be characterized quantitatively, this strategy of dissection into modular risks becomes problematic. The limited applicability of fault tree analysis in environmental and human health has many reasons, but in the end the root cause is usually the enormous complexity of biological systems and the challenge of truly understanding their functioning. For instance, biological systems seldom have a pure tree structure, but instead cycle material back or control themselves with regulatory signals such as feedback inhibition. The challenge of understanding this complexity cannot be mastered with intuition alone but must be addressed with systematic, computational, and mathematical modeling methods, for three reasons.

First, even the simplest biological systems consist of more components than our minds can handle simultaneously. A yeast cell has about 6000 genes, humans have proteins numbering in the hundred thousands. Just naming such vast quantities and tracking them through time and space is difficult. Second, biological components may interact with each other in multitudinous ways, many of which we do not yet understand or even know. The same protein may

have different functional roles; the same gene may be spliced in different ways, thereby coding for different end products; lipids are important building blocks of membranes, but they may also serve as signaling molecules. Third, the interactions within biological systems are seldom linear, with the consequence that increasing an input does not necessarily evoke a similarly increasing output. Following Caesar's motto to divide and conquer (i.e., exploiting the principle of superposition), we are trained to apportion big problems into manageable tasks, and this strategy has been very successful for linear systems. However, for nonlinear systems, subdivision into subsystems becomes problematic when important structural or functional connections are disrupted in the dissection process. Thus, reliable insights into the function of biological systems require integrated representations and nonlinear methods of analysis.

Nature has not provided us with guidelines for how to represent nonlinear phenomena, and the range of possibilities is infinite. It is not even easy to compare objectively two functional representations of simple sigmoidal processes, as they are used as descriptions for growth processes and in high-dose to low-dose extrapolations (*see* **Low-Dose Extrapolation**) in toxicological risk assessment. Indeed, logistic, Hill, arc tangent, and many other functions could essentially give the same data fits [2]. Instead of trying to find "the true" or "the optimal" model, it is often useful to resort to *nonlinear canonical models*, which are based on some suitable type of approximation and have a prescribed mathematical format [3]. Best known among these are Lotka-Volterra models [4] and models of biochemical systems theory (BST [5]). Interestingly, these nonlinear canonical models typically consist of a combination of differentiation, logarithmic transformation, and summation. For instance, each differential equation in a Lotka-Volterra system may be written as the derivative of a logarithmic variable, which equals the weighted sum of the systems variables, while BST models derive from linearization in logarithmic space. Generic advantages of canonical models include the following (*see* [6]):

1. they are mathematically guaranteed to be correct in the vicinity of some nominal state of choice, the so-called *operating point*;
2. many biological observations have been modeled very well;

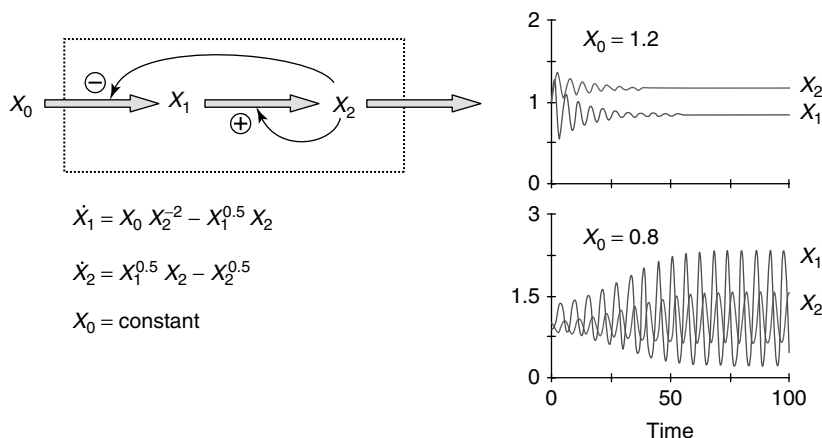


Figure 1 Generic pathway with inhibition and activation, represented in the manner of BST. Although seemingly simple, the pathway can exhibit qualitatively different responses

3. distinct guidelines are available for constructing and analyzing models;
4. efficient, customized computational tools for analysis, optimization and objective comparisons between competing models have been developed; and
5. they are rich enough in structure to represent essentially all types of smooth processes, no matter how complicated.

Thus, while canonical modeling structures appear at first restrictive, they often have genuine advantages over the infinite possibilities of *ad hoc* models.

As a generic example, consider the deceptively simple pathway in Figure 1 with a typical BST representation shown beneath. Although unbranched, its responses to changes in input X_0 are difficult to foresee. For instance, if $X_0 > 1$, X_1 and X_2 quickly reach a stable steady state. However, as soon as $X_0 < 1$, the steady state becomes unstable, and X_1 and X_2 oscillate in a stable limit cycle around the unstable point. BST models of practical importance and considerable size have been formulated, fully parameterized, and thoroughly analyzed to yield novel insights. Examples include a red blood cell model with close to 100 variables and models of

citric acid, TCA, purine, and sphingolipid metabolism with between 10 and 50 variables. Applications in environmental studies are reviewed in reference [3].

References

- [1] Fullwood, R.R. & Hall, R.E. (1988). *Probabilistic Risk Assessment in the Nuclear Power Industry*, Pergamon Press, Oxford.
- [2] Torres, N.V. & Voit, E.O. (2002). *Pathway Analysis and Optimization in Metabolic Engineering*, Cambridge University Press, Cambridge.
- [3] Voit, E.O. (2000). Canonical modeling: a review of concepts with emphasis on environmental health (invited). *Environmental Health Perspectives*, **108**(Suppl. 5), 895–909.
- [4] Takeuchi, Y. (1996). *Global Dynamical Properties of Lotka-Volterra Systems*, World Scientific, Singapore.
- [5] Savageau, M.A. (1976). *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*, Addison-Wesley, Reading.
- [6] Voit, E.O. (2000). *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*, Cambridge University Press, Cambridge, pp. xii and 530.

EBERHARD O. VOIT

Case–Control Studies

There are two primary types of nonexperimental studies in epidemiology. The first, the *cohort study* (see **Cohort Studies**) (also called the *follow-up study* or *incidence study*), is a direct analog of the experiment. Different exposure groups are compared, but it is the investigator who selects subjects to observe, and classifies these subjects by exposure status, rather than assigning them to exposure groups. The second, the *incident case–control study*, or simply the *case–control study*, employs an extra step of sampling from the source population for cases. A cohort study includes all persons in the population at risk of becoming a study case. In contrast, a case-control study selects only a sample of those persons, and does so partly on the basis of their final disease status. Thus, by design, a person’s outcome influences their chance of becoming a subject in the case-control study. This extra sampling step can make a case–control study much more efficient than a cohort study of the same population, but it introduces a number of subtleties and avenues for bias that are absent in typical cohort studies.

Conventional wisdom about case–control studies is that they do not yield estimates of effect that are as valid as measures obtained from cohort studies. This thinking may reflect common misunderstandings in conceptualizing case–control studies, but it also reflects concern about quality of exposure information and biases in case or control selection. For example, if exposure information comes from interviews, then cases will have usually reported the exposure information after learning of their diagnosis, which can lead to errors in the responses that are related to the disease (recall bias). While it is true that recall bias does not occur in prospective cohort studies, neither does it occur in all case–control studies. Exposure information that is taken from records whose creation predated disease occurrence will not be subject to recall bias. Similarly, while a cohort study may log information on exposure for an entire source population at the outset of the study, it still requires tracing of subjects to ascertain exposure variation and outcomes, and the success of this tracing may be related to exposure. These concerns are analogous to case–control problems of loss of subjects with unknown exposure and to biased selection

of controls and cases. Each study, whether cohort or case–control, must be considered on its own merits.

Conventional wisdom also holds that cohort studies are useful for evaluating the range of effects related to a single exposure, while case–control studies provide information only about the one disease that afflicts the cases. This thinking conflicts with the idea that case–control studies can be viewed simply as more efficient cohort studies. Just as one can choose to measure more than one disease outcome in a cohort study, it is possible to conduct a set of case–control studies nested within the same population using several disease outcomes as the case series. The case–cohort study (see the section titled “**Case–Cohort Studies**”) is particularly well suited to this task, allowing one control group to be compared with several series of cases. Whether or not the case–cohort design is the form of case–control study that is used, case–control studies do not have to be characterized as being limited with respect to the number of disease outcomes that can be studied.

For diseases that are sufficiently rare, cohort studies become impractical, and case–control studies become the only useful alternative. On the other hand, if exposure is rare, ordinary case–control studies are inefficient, and one must use methods that selectively recruit additional exposed subjects, such as special cohort studies or two-stage designs. If both the exposure and the outcome are rare, two-stage designs may be the only informative option, as they employ oversampling of both exposed and diseased subjects.

Ideally, a case–control study can be conceptualized as a more efficient version of a corresponding cohort study. Under this conceptualization, the cases in the case–control study are the same cases as would ordinarily be included in the cohort study. Rather than including all of the experience of the source population that gave rise to the cases (the study base), as would be the usual practice in a cohort design, controls are selected from the source population. The sampling of controls from the population that gave rise to the cases affords the efficiency gain of a case–control design over a cohort design. The controls provide an estimate of the prevalence of the exposure and covariates in the source population. When controls are selected from members of the population who were at risk for disease at the beginning of the study’s follow-up period, the case–control odds ratio (see **Odds and Odds Ratio**) estimates the risk ratio that would be obtained from a cohort

design. When controls are selected from members of the population who were noncases at the times that each case occurs, or otherwise in proportion to the person-time accumulated by the cohort, the case-control odds ratio estimates the rate ratio that would be obtained from a cohort design. Finally, when controls are selected from members of the population who were noncases at the end of the study's follow-up period, the case-control odds ratio estimates the incidence odds ratio that would be obtained from a cohort design. With each control selection strategy, the odds ratio calculation is the same, but the measure of effect estimated by the odds ratio differs. Study designs that implement each of these control selection paradigms will be discussed after topics that are common to all designs.

Common Elements of Case-Control Studies

In a cohort study, the numerator and denominator of each disease frequency (incidence proportion, incidence rate, or incidence odds) are measured, which requires enumerating the entire population and keeping it under surveillance. A case-control study attempts to observe the population more efficiently by using a control series in place of complete assessment of the denominators of the disease frequencies. The cases in a case-control study should be the same people who would be considered cases in a cohort study of the same population.

Pseudofrequencies and the Odds Ratio

The primary goal for control selection is that the exposure distribution among controls be the same as it is in the source population of cases. The rationale for this goal is that, if it is met, we can use the control series in place of the denominator information in measures of disease frequency to determine the ratio of the disease frequency in exposed people relative to that among unexposed people. This goal will be met if we can sample controls from the source population such that the ratio of the number of exposed controls (B_1) to the total exposed experience of the source population is the same as the ratio of the number of unexposed controls (B_0) to the unexposed experience of the source population, apart from sampling error. For most purposes, this goal need only be followed

within strata defined by the factors that are used for stratification in the analysis, such as factors used for restriction or matching.

Using person-time to illustrate, the goal requires that B_1 has the same ratio to the amount of exposed person-time (T_1) as B_0 has to the amount of unexposed person-time (T_0):

$$\frac{B_1}{T_1} = \frac{B_0}{T_0} \quad (1)$$

Here B_1/T_1 and B_0/T_0 are the control sampling rates – that is, the number of controls selected per unit of person-time. Suppose A_1 exposed cases and A_0 unexposed cases occur over the study period. The exposed and unexposed rates are then

$$I_1 = \frac{A_1}{T_1} \quad (2)$$

and

$$I_0 = \frac{A_0}{T_0} \quad (3)$$

We can use the frequencies of exposed and unexposed controls as substitutes for the actual denominators of the rates to obtain exposure-specific case-control ratios, or *pseudorates*:

$$\text{Pseudorate}_1 = \frac{A_1}{B_1} \quad (4)$$

and

$$\text{Pseudorate}_0 = \frac{A_0}{B_0} \quad (5)$$

These pseudorates have no epidemiologic interpretation by themselves. Suppose, however, that the control sampling rates B_1/T_1 and B_0/T_0 are equal to the same value r , as would be expected if controls are selected independently of exposure. If this common sampling rate r is known, the actual incidence rates can be calculated by simple algebra, since apart from sampling error, B_1/r should equal the amount of exposed person-time in the source population, and B_0/r should equal the amount of unexposed person-time in the source population: $B_1/r = B_1/(B_1/T_1) = T_1$ and $B_0/r = B_0/(B_0/T_0) = T_0$. To get the incidence rates, we need only multiply each pseudorate by the common sampling rate, r .

If the common sampling rate is not known, which is often the case, we can still compare the sizes of

the pseudorates by division. Specifically, if we divide the pseudorate for exposed by the pseudorate for unexposed, we obtain

$$\begin{aligned}\frac{\text{Pseudorate}_1}{\text{Pseudorate}_0} &= \frac{A_1/B_1}{A_0/B_0} = \frac{A_1/[(B_1/T_1)T_1]}{A_0/[(B_0/T_0)T_0]} \\ &= \frac{A_1/(r \cdot T_1)}{A_0/(r \cdot T_0)} = \frac{A_1/T_1}{A_0/T_0}\end{aligned}\quad (6)$$

In other words, the ratio of the pseudorates for the exposed and unexposed is an estimate of the ratio of the incidence rates in the source population, provided that the control sampling rate is independent of exposure. Thus, using the case-control study design, one can estimate the incidence rate ratio in a population without obtaining information on every subject in the population. Similar derivations in the section titled “Variants of the Case-Control Design” show that one can estimate the risk ratio by sampling controls from those at risk for disease at the beginning of the follow-up period (case-cohort design) and that one can estimate the incidence odds ratio by sampling controls from the noncases at the end of the follow-up period (cumulative case-control design). With these designs, the pseudofrequencies correspond to the incidence proportions and incidence odds, respectively, multiplied by common sampling rates.

There is a statistical penalty for using a sample of the denominators, rather than measuring the person-time experience for the entire source population: the precision of the estimates of the incidence rate ratio from a case-control study is less than the precision from a cohort study of the entire population that gave rise to the cases (the source population). Nevertheless, the loss of precision that stems from sampling controls will be small if the number of controls selected per case is large. Furthermore, the loss is balanced by the cost savings of not having to obtain information on everyone in the source population. The cost savings might allow the epidemiologist to enlarge the source population and so obtain more cases, resulting in a better overall estimate of the incidence rate ratio, statistically and otherwise, than would be possible using the same expenditures to conduct a cohort study.

The ratio of the two pseudorates in a case-control study is usually written as A_1B_0/A_0B_1 , and is sometimes called the *cross-product ratio*. The cross-product ratio in a case-control study can be viewed

as the ratio of cases to controls among the exposed subjects (A_1/B_1), divided by the ratio of cases to controls among the unexposed subjects (A_0/B_0). This ratio can also be viewed as the odds of being exposed among cases (A_1/A_0) divided by the odds of being exposed among controls (B_1/B_0), in which case it is termed the *exposure odds ratio*. While either interpretation will give the same result, viewing this odds ratio as the ratio of case-control ratios shows more directly how the control group substitutes for the denominator information in a cohort study and how the ratio of pseudofrequencies gives the same result as the ratio of the incidence rates, incidence proportion, or incidence odds in the source population, if sampling is independent of exposure.

Defining the Source Population

If the cases are a representative sample of all cases in a precisely defined and identified population, and the controls are sampled directly from this source population, the study is said to be *population-based* or a *primary-base* study. For a population-based case-control study, random sampling of controls may be feasible if a population registry exists or can be compiled. When random sampling from the source population of cases is feasible, it is usually the most desirable option.

Random sampling of controls does not necessarily mean that every person should have an equal probability of being selected to be a control. As explained above, if the aim is to estimate the incidence rate ratio, then we would employ longitudinal (density) sampling, in which a person's control selection probability is proportional to the person's time at risk. For example, in a case-control study nested within an occupational cohort, workers on an employee roster will have been followed for varying lengths of time, and a random sampling scheme should reflect this varying time to estimate the incidence rate ratio.

When it is not possible to identify the source population explicitly, simple random sampling is not feasible, and other methods of control selection must be used. Such studies are sometimes called *studies of secondary* bases, because the source population is identified secondarily to the definition of a case-finding mechanism. A secondary source population or *secondary* base is therefore a source population that is defined from (secondary to) a given case series.

Consider a case-control study in which the cases are patients treated for severe psoriasis at the Mayo Clinic. These patients come to the Mayo Clinic from all corners of the world. What is the specific source population that gives rise to these cases? To answer this question, we would have to know exactly who would go to the Mayo Clinic, if he or she had severe psoriasis. We cannot enumerate this source population because many people in it do not know themselves that they would go to the Mayo Clinic for severe psoriasis, unless they actually developed severe psoriasis. This secondary source might be defined as a population spread around the world that constitutes those people who would go to the Mayo Clinic if they developed severe psoriasis. It is this secondary source from which the control series for the study would ideally be drawn. The challenge to the investigator is to apply eligibility criteria to the cases and controls so that there is good correspondence between the controls and this source population. For example, cases of severe psoriasis and controls might be restricted to those in counties within a certain distance of the Mayo Clinic, so that at least a geographic correspondence between the controls and the secondary source population can be assured. This restriction might, however, leave very few cases for study.

Unfortunately, the concept of a secondary base is often tenuously connected to underlying realities, and can be highly ambiguous. For the psoriasis example, whether a person would go to the Mayo Clinic depends on many factors that vary over time, such as whether the person is encouraged to go by their regular physicians and whether the person can afford to go. It is not clear, then, how or even whether one could precisely define the secondary base, let alone draw a sample from it; thus it is not clear one could ensure that controls were members of the base at the time of sampling. We therefore prefer to conceptualize and conduct case-control studies as starting with a well-defined source population, and then identify and recruit cases and controls to represent the disease and exposure experience of that population. When one takes a case series as a starting point instead, it is incumbent upon the investigator to demonstrate that a source population can be operationally defined to allow the study to be recast and evaluated relative to this source. Similar considerations apply when one takes a control series as a starting point, as is sometimes done [1].

Case Selection

Ideally, case selection will amount to a direct sampling of cases within a source population. Therefore, apart from random sampling, all people in the source population who develop the disease of interest are presumed to be included as cases in the case-control study. It is not always necessary, however, to include all cases from the source population. Cases, like controls, can be randomly sampled for inclusion in the case-control study, so long as this sampling is independent of the exposure under study within the strata defined by the stratification factors that are used in the analysis. Of course, if fewer than all cases are sampled, the study precision will be lower in proportion to the sampling fraction.

The cases identified in a single clinic or treated by a single medical practitioner are possible case series for case-control studies. The corresponding source population for the cases treated in a clinic comprises all people who would attend that clinic and would be recorded with the diagnosis of interest, if they had the disease in question. It is important to specify “if they had the disease in question” because clinics serve different populations for different diseases, depending on referral patterns and the reputation of the clinic in specific speciality areas. As noted above, without a precisely identified source population, it may be difficult or impossible to select controls in an unbiased fashion.

Control Selection

The definition of the source population determines the population from which controls are sampled. Ideally, control selection will amount to a direct sampling of people within the source population. On the basis of the principles explained above regarding the role of the control series, many general rules for control selection can be formulated. Two basic rules are as follows: (a) Controls should be selected from the same population – the source population – that gives rise to the study cases. If this rule cannot be followed, there needs to be solid evidence that the population supplying controls has an exposure distribution identical to that of the population that is the source of cases, which is a very stringent demand that is rarely demonstrable. (b) Within strata defined by factors that are used for stratification in the analysis, controls should be selected independently

of their exposure status, in that the sampling rate for controls (r in the above discussion) should not vary with exposure.

If these rules and the corresponding case rule are met, then the ratio of pseudofrequencies will, apart from sampling error, equal the ratio of the corresponding measure of disease frequency in the source population. If the sampling rate is known, then the actual measures of disease frequency can also be calculated [2]. Wacholder *et al.* have elaborated on the principles of control selection in case-control studies [3–5].

When one wishes controls to represent person-time, sampling of the person-time should be constant across exposure levels. This requirement implies that the sampling *probability* of any person as a control should be proportional to the amount of person-time that person spends at risk of disease in the source population. For example, if in the source population one person contributes twice as much person-time during the study period as another person, the first person should have twice the probability of the second of being selected as a control.

This difference in probability of selection is automatically induced by sampling controls at a steady rate per unit time over the period in which cases occur (longitudinal, or density sampling), rather than by sampling all controls at a point in time (such as the start or end of the study). With longitudinal sampling of controls, a population member present for twice as long as another will have twice the chance of being selected.

If the objective of the study is to estimate a risk or rate ratio, it should be possible for a person to be selected as a control and yet remain eligible to become a case, so that person might appear in the study as both a control and a case. This possibility may sound paradoxical or wrong, but is, nevertheless, correct. It corresponds to the fact that in a cohort study, a case contributes to both the numerator and the denominator of the estimated incidence. If the controls are intended to represent person-time and are selected longitudinally, similar arguments show that a person selected as a control should remain eligible to be selected as a control again, and thus might be included in the analysis repeatedly as a control [6, 7].

Common Fallacies in Control Selection

In cohort studies, the study population is restricted to people at risk for the disease. Because they

viewed case-control studies as if they were cohort studies done backwards, some authors argued that case-control studies ought to be restricted to those at risk for exposure (i.e., those with exposure opportunity). Excluding sterile women from a case-control study of an adverse effect of oral contraceptives and matching for duration of employment in an occupational study are examples of attempts to control for exposure opportunity. Such restrictions do not directly address validity issues, and can ultimately harm study precision by reducing the number of unexposed subjects available for study [8]. If the factor used for restriction (e.g., sterility) is unrelated to the disease, it will not be a confounder, and hence the restriction will yield no benefit to the validity of the estimate of effect. Furthermore, if the restriction reduces the study size, the precision of the estimate of effect will be reduced.

Another principle sometimes used in cohort studies is that the study cohort should be “clean” at start of follow-up, including only people who have never had the disease. Misapplying this principle to case-control design suggests that the control group ought to be “clean”, including only people who are healthy, for example. Illness arising after the start of the follow-up period is not reason to exclude subjects from a cohort analysis, and such exclusion can lead to bias. Similarly, controls with illness that arose after exposure should not be removed from the control series. Nonetheless, in studies of the relation between cigarette smoking and colorectal cancer, certain authors recommended that the control group should exclude people with colon polyps, because colon polyps are associated with smoking and are precursors of colorectal cancer [9]. But such exclusion reduces the prevalence of the exposure in the controls below that in the actual source population of cases, and hence biases the effect estimates upward [10].

Sources for Control Series

The methods suggested below for control sampling apply when the source population cannot be explicitly enumerated, so random sampling is not possible. These methods should only be implemented subject to the reservations about secondary bases described above.

Neighborhood Controls. If the source population cannot be enumerated, it may be possible to select

controls through sampling of residences. This method is not straightforward. Usually, a geographic roster of residences is not available, so a scheme must be devised to sample residences without enumerating them all. For convenience, investigators may sample controls who are individually matched to cases from the same neighborhood. That is, after a case is identified, one or more controls residing in the same neighborhood as that case are identified and recruited into the study. If neighborhood is related to exposure, the matching should be taken into account in the analysis.

Neighborhood controls are often used when the cases are recruited from a convenient source, such as a clinic or hospital. Such usage can introduce bias, however, for the neighbors selected as controls may not be in the source population of the cases. For example, if the cases are from a particular hospital, neighborhood controls may include people who would not have been treated at the same hospital had they developed the disease. If being treated at the hospital from which cases are identified is related to the exposure under study, then using neighborhood controls would introduce a bias. For any given study, the suitability of using neighborhood controls needs to be evaluated with regard to the study variables on which the research focuses.

Hospital- or Clinic-Based Controls. As noted above, the source population for hospital- or clinic-based case-control studies is not often identifiable, since it represents a group of people who would be treated in a given clinic or hospital if they developed the disease in question. In such situations, a random sample of the general population will not necessarily correspond to a random sample of the source population. If the hospitals or clinics that provide the cases for the study only treat a small proportion of cases in the geographic area, then referral patterns to the hospital or clinic are important to take into account in the sampling of controls. For these studies, a control series comprising patients from the same hospitals or clinics as the cases may provide a less biased estimate of effect than general-population controls (such as those obtained from case neighborhoods or by random-digit dialing). The source population does not correspond to the population of the geographic area, but only to the people who would seek treatment at the hospital or clinic were they to develop the disease under study. While the latter population may

be difficult or impossible to enumerate or even define very clearly, it seems reasonable to expect that other hospital or clinic patients will represent this source population better than general-population controls. The major problem with any nonrandom sampling of controls is the possibility that they are not selected independently of exposure in the source population. Patients hospitalized with other diseases, for example, may be unrepresentative of the exposure distribution in the source population either because exposure is associated with hospitalization, or because the exposure is associated with the other diseases, or both. For example, suppose the study aims to evaluate the relation between tobacco smoking and leukemia using hospitalized cases. If controls are people hospitalized with other conditions, many of them will have been hospitalized for conditions associated with smoking. A variety of other cancers, as well as cardiovascular diseases and respiratory diseases, are related to smoking. Thus, a control series of people hospitalized for diseases other than leukemia would include a higher proportion of smokers than would the source population of the leukemia cases.

Limiting the diagnoses for controls to conditions for which there is no prior indication of an association with the exposure improves the control series. For example, in a study of smoking and hospitalized leukemia cases, one could exclude from the control series anyone who was hospitalized with a disease known to be related to smoking. Such an exclusion policy may exclude most of the potential controls, since cardiovascular disease by itself would represent a large proportion of hospitalized patients. Nevertheless, even a few common diagnostic categories should suffice to find enough control subjects, so that the exclusions will not harm the study by limiting the size of the control series. Indeed, in limiting the scope of eligibility criteria, it is reasonable to exclude categories of potential controls even on the suspicion that a given category might be related to the exposure. If wrong, the cost of the exclusion is that the control series becomes more homogeneous with respect to diagnosis and perhaps a little smaller. But if right, then the exclusion is important to the ultimate validity of the study.

On the other hand, an investigator can rarely be sure that an exposure is not related to a disease or to hospitalization for a specific diagnosis. Consequently, it would be imprudent to use only a single diagnostic category as a source of controls. Using a variety of

diagnoses has the advantage of potentially diluting the biasing effects of including a specific diagnostic group that is related to the exposure.

Excluding a diagnostic category from the list of eligibility criteria for identifying controls is intended simply to improve the representativeness of the control series with respect to the source population. Such an exclusion criterion does not imply that there should be exclusions based on disease history [11]. For example, in a case-control study of smoking and hospitalized leukemia patients, one might use hospitalized controls but exclude any who are hospitalized because of cardiovascular disease. This exclusion criterion for controls does not imply that leukemia cases who have had cardiovascular disease should be excluded; only if the cardiovascular disease was a cause of the hospitalization, should the case be excluded. For controls, the exclusion criterion should only apply to the cause of the hospitalization used to identify the study subject. A person who was hospitalized because of a traumatic injury and who is thus eligible to be a control would not be excluded if he or she had previously been hospitalized for cardiovascular disease. The source population includes people who have had cardiovascular disease, and they should be included in the control series. Excluding such people would lead to an underrepresentation of smoking relative to the source population and produce an upward bias in the effect estimates.

If exposure directly affects hospitalization (for example, if the decision to hospitalize is in part based on exposure history), the resulting bias cannot be remedied without knowing the hospitalization rates, even if the exposure is unrelated to the study disease or the control diseases. This problem was in fact one of the first problems of hospital-based studies to receive detailed analysis [12], and is often called *Berksonian bias*.

Other Diseases. In many settings, especially in populations with established disease registries or insurance-claims databases, it may be most convenient to choose controls from people who are diagnosed with other diseases. The considerations needed for valid control selection from other diagnoses parallel those just discussed for hospital controls. It is essential to exclude any diagnoses known or suspected to be related to exposure, and better still to include only diagnoses for which there is some evidence to indicate they are unrelated to exposure.

These exclusion and inclusion criteria apply only to the diagnosis that brought the person into the registry or database from which controls are selected. The history of an exposure-related disease should not be a basis for exclusion. If, however, the exposure directly affects the chance of entering the registry or database, the study will be subject to the Berksonian bias mentioned earlier for hospital studies.

Other Considerations for Subject Selection

Representativeness. Some textbooks have stressed the need for representativeness in the selection of cases and controls. The advice has been that cases should be representative of all people with the disease and that controls should be representative of the entire nondiseased population. Such advice can be misleading. A case-control study may be restricted to any type of case that may be of interest: female cases, old cases, severely ill cases, cases that died soon after disease onset, mild cases, cases from Philadelphia, cases among factory workers, and so on. In none of these examples would the cases be representative of all people with the disease, yet, in each one, perfectly valid case-control studies are possible [13]. The definition of a case can be virtually anything that the investigator wishes.

Ordinarily, controls should represent the source population for cases, rather than the entire nondiseased population. The latter may differ vastly from the source population for the cases by age, race, sex (e.g., if the cases come from a Veterans administration hospital), socioeconomic status, occupation, and so on – including the exposure of interest. One of the reasons for emphasizing the similarities rather than the differences between cohort and case-control studies is that numerous principles apply to both types of study, but are more evident in the context of cohort studies. In particular, many principles relating to subject selection apply identically to both types of study. For example, it is widely appreciated that cohort studies can be based on special cohorts, rather than on the general population. It follows that case-control studies can be conducted by sampling cases and controls from within those special cohorts. The resulting controls should represent the distribution of exposure across those cohorts, rather than the general population, reflecting the more general rule that controls should represent the source population of the cases in the study, not the general population.

Comparability of Information. Some authors have recommended that information obtained about cases and controls should be of comparable or equal accuracy, to ensure nondifferentiality (equal distribution) of measurement errors [3]. The rationale for this principle is the notion that nondifferential measurement error biases the observed association toward the null, and so will not generate a spurious association, and that bias in studies with nondifferential error is more predictable than in studies with differential error.

The comparability-of-information principle is often used to guide selection of controls and collection of data. For example, it is the basis for using proxy respondents instead of direct interviews for living controls, whenever case information is obtained from proxy respondents. Unfortunately, in most settings, the arguments for the principle are logically unsound. For example, in a study that used proxy respondents for cases, use of proxy respondents for the controls might lead to greater bias than use of direct interviews with controls, even if measurement error is differential. The comparability-of-information principle is therefore applicable only under very limited conditions. In particular, it would seem to be useful only when confounders and effect modifiers are measured with negligible error, and when measurement error is reduced by using comparable sources of information. Otherwise, the effect of forcing comparability of information may be as unpredictable as the effect of using noncomparable information.

Timing of Classification and Diagnosis. The principles for classifying persons, cases, and person-time units in cohort studies according to exposure status also apply to cases and controls in case-control studies. If the controls are intended to represent person-time (rather than persons) in the source population, one should apply principles for classifying person-time to the classification of controls. In particular, principles of person-time classification lead to the rule that controls should be classified by their exposure status as of their selection time. Exposures accrued after that time should be ignored. The rule necessitates that information (such as exposure history) be obtained in a manner that allows one to ignore exposures accrued after the selection time. In a similar manner, cases should be classified as of time of diagnosis or disease onset, accounting for any built-in lag periods or induction-period hypotheses.

Variants of the Case-Control Design

Nested Case-Control Studies

Epidemiologists sometimes refer to specific case-control studies as *nested* case-control studies when the population within which the study is conducted is a fully enumerated cohort, which allows formal random sampling of cases and controls to be carried out. The term is usually used in reference to a case-control study conducted within a cohort study, in which further information (perhaps from expensive tests) is obtained on most or all cases, but for economy is obtained from only a fraction of the remaining cohort members (the controls). Nonetheless, many population-based case-control studies can be thought of as nested within an enumerated source population.

Case-Cohort Studies

The *case-cohort study* is a case-control study in which the source population is a cohort, and every person in this cohort has an equal chance of being included in the study as a control, regardless of how much time that person has contributed to the person-time experience of the cohort or whether the person developed the study disease. This is a logical way to conduct a case-control study when the effect measure of interest is the ratio of incidence proportions rather than a rate ratio, as is common in perinatal studies. The average risk (or incidence proportion) of falling ill during a specified period may be written as

$$R_1 = \frac{A_1}{N_1} \quad (7)$$

for the exposed subcohort and

$$R_0 = \frac{A_0}{N_0} \quad (8)$$

for the unexposed subcohort, where R_1 and R_0 are the incidence proportions among the exposed and unexposed, respectively, and N_1 and N_0 are the initial sizes of the exposed and unexposed subcohorts. (This discussion applies equally well to exposure variables with several levels, but, for simplicity, we consider only a dichotomous exposure.) Controls should be selected such that the exposure distribution among them will estimate without bias the exposure distribution in the source population. In a case-cohort study,

the distribution we wish to estimate is among the $N_1 + N_0$ cohort members, not among their person-time experience [14–16].

The objective is to select controls from the source cohort such that the ratio of the number of exposed controls (B_1) to the number of exposed cohort members (N_1) is the same as the ratio of the number of unexposed controls (B_0) to the number of unexposed cohort members (N_0), apart from sampling error:

$$\frac{B_1}{N_1} = \frac{B_0}{N_0} \quad (9)$$

Here, B_1/N_1 and B_0/N_0 are the control sampling fractions (the number of controls selected per cohort member). Apart from random error, these sampling fractions will be equal if controls have been selected independently of exposure.

We can use the frequencies of exposed and unexposed controls as substitutes for the actual denominators of the incidence proportions to obtain “pseudorisks”:

$$\text{Pseudorisk}_1 = \frac{A_1}{B_1} \quad (10)$$

and

$$\text{Pseudorisk}_0 = \frac{A_0}{B_0} \quad (11)$$

These pseudorisks have no epidemiologic interpretation by themselves. Suppose, however, that the control sampling fractions are equal to the same fraction, f . Then, apart from sampling error, B_1/f should equal N_1 , the size of the exposed subcohort; and B_0/f should equal N_0 , the size of the unexposed subcohort: $B_1/f = B_1/(B_1/N_1) = N_1$ and $B_0/f = B_0/(B_0/N_0) = N_0$. Thus, to get the incidence proportions, we need only multiply each pseudorisk by the common sampling fraction, f . If this fraction is not known, we can still compare the sizes of the pseudorisks by division:

$$\begin{aligned} \frac{\text{Pseudorisk}_1}{\text{Pseudorisk}_0} &= \frac{A_1/B_1}{A_0/B_0} = \frac{A_1/[(B_1/N_1)N_1]}{A_0/[(B_0/N_0)N_0]} \\ &= \frac{A_1/fN_1}{A_0/fN_0} = \frac{A_1/N_1}{A_0/N_0} \end{aligned} \quad (12)$$

In other words, the ratio of pseudorisks is an estimate of the ratio of incidence proportions (risk ratio) in the source cohort if control sampling

is independent of exposure. Thus, using a case-cohort design, one can estimate the risk ratio in a cohort without obtaining information on every cohort member.

Thus far, we have implicitly assumed that there is no loss to follow-up or competing risks in the underlying cohort. If there are such problems, it is still possible to estimate risk or rate ratios from a case-cohort study, provided that we have data on the time spent at risk by the sampled subjects or we use certain sampling modifications [17]. These procedures require the usual assumptions for rate-ratio estimation in cohort studies, namely, that loss to follow-up and competing risks are either not associated with exposure or not associated with disease risk.

An advantage of the case-cohort design is that it facilitates conduct of a set of case-control studies from a single cohort, all of which use the same control group. Just as one can measure the incidence rate of a variety of diseases within a single cohort, one can conduct a set of simultaneous case-control studies using a single control group. A sample from the cohort is the control group needed to compare with any number of case groups. If matched controls are selected from people at risk at the time a case occurs (as in risk-set sampling, which is described in the section titled “Density Case-Control Studies”), the control series must be tailored to a specific group of cases. To have a single control series serve many case groups, another sampling scheme must be used. The case-cohort approach is a good choice in such a situation.

Wacholder has discussed the advantages and disadvantages of the case-cohort design relative to the usual type of case-control study [18]. One point to note is that, because of the overlap of membership in the case and control groups (controls who are selected may also develop disease and enter the study as cases), one will need to select more controls in a case-cohort study than in an ordinary case-control study with the same number of cases, if one is to achieve the same amount of statistical precision. Extra controls are needed because the statistical precision of a study is strongly determined by the numbers of distinct cases and noncases. Thus, if 20% of the source cohort members will become cases, and all cases will be included in the study, one will have to select 1.25 times as many controls as cases in a case-cohort study to insure that there will be as many

controls who never become cases in the study. On average, only 80% of the controls in such a situation will remain noncases; the other 20% will become cases. Of course, if the disease is uncommon, the number of extra controls needed for a case-cohort study will be small.

Density Case-Control Studies

Earlier, we described how case-control odds ratios will estimate rate ratios if the control series is selected so that the ratio of the person-time denominators T_1/T_0 is validly estimated by the ratio of exposed to unexposed controls B_1/B_0 . That is, to estimate rate ratios, controls should be selected so that the exposure distribution among them is, apart from random error, the same as it is among the person-time in the source population. Such control selection is called *density sampling* because it provides for estimation of relations among incidence rates, which have been called *incidence densities*.

If a subject's exposure may vary over time, then a case's exposure history is evaluated up to the time the disease occurred. A control's exposure history is evaluated up to an analogous index time, usually taken as the time of sampling; exposure after the time of selection must be ignored. This rule helps ensure that the number of exposed and unexposed controls will be in proportion to the amount of exposed and unexposed person-time in the source population.

The time during which a subject is eligible to be a control should be the time in which that person is also eligible to become a case, should the disease occur. Thus, a person in whom the disease has already developed or who has died is no longer eligible to be selected as a control. This rule corresponds to the treatment of subjects in cohort studies. Every case that is tallied in the numerator of a cohort study contributes to the denominator of the rate until the time that the person becomes a case, when the contribution to the denominator ceases. One way to implement this rule is to choose controls from the set of people in the source population who are at risk of becoming a case at the time that the case is diagnosed. This set is sometimes referred to as the *risk set* for the case, and this type of control sampling is sometimes called *risk-set sampling*. Controls sampled in this manner are matched to the case with respect to sampling time; thus, if time is related to exposure, the resulting data should be analyzed as matched data

[19]. It is also possible to conduct unmatched density sampling using probability sampling methods if one knows the time interval at risk for each population member. One then selects a control by sampling members with probability proportional to time at risk, and then randomly samples a time to measure exposure within the interval at risk.

As mentioned earlier, a person selected as a control, and who remains in the study population at risk after selection should remain eligible to be selected once again as a control. Thus, although unlikely in typical studies, the same person may appear in the control group two or more times. Note, however, that including the same person at different times does not necessarily lead to exposure (or confounder) information being repeated, because this information may change with time. For example, in a case-control study of an acute epidemic of intestinal illness, one might ask about food ingested within the previous day or days. If a contaminated food item was a cause of the illness for some cases, then the exposure status of a case or control chosen 5 days into the study might well differ from what it would have been 2 days into the study when the subject might also have been included as a control.

Cumulative ("Epidemic") Case-Control Studies

In some research settings, case-control studies may address a risk that ends before subject selection begins. For example, a case-control study of an epidemic of diarrheal illness after a social gathering may begin after all the potential cases have occurred (because the maximum induction time has elapsed). In such a situation, an investigator might select controls from that portion of the population that remains after eliminating the accumulated cases; that is, one selects controls from among noncases (those who remain noncases at the end of the epidemic follow-up).

Suppose that the source population is a cohort and that a fraction f of both exposed and unexposed noncases are selected to be controls. Then the ratio of pseudofrequencies will be

$$\frac{A_1/B_1}{A_0/B_0} = \frac{A_1/f(N_1 - A_1)}{A_0/f(N_0 - A_0)} = \frac{A_1/(N_1 - A_1)}{A_0/(N_0 - A_0)} \quad (13)$$

which is the incidence odds ratio for the cohort. The latter ratio will provide a reasonable approximation to the rate ratio, provided that the proportions falling ill

in each exposure group during the risk period are low, that is, less than about 20%, and that the prevalence of exposure remains reasonably steady during the study period. If the investigator prefers to estimate the risk ratio rather than the incidence rate ratio, the study odds ratio can still be used [20], but the accuracy of this approximation is only about half as good as that of the odds ratio approximation to the rate ratio [21]. The use of this approximation in the cumulative design is the basis for the common and mistaken notion that a rare-disease assumption is needed to estimate risk ratios in all case-control studies.

Prior to the 1970s, the standard conceptualization of case-control studies involved the cumulative design, in which controls are selected from noncases at the end of a follow-up period. As discussed by numerous authors [19, 22, 23], density designs and case-cohort designs have several advantages outside of the acute epidemic setting, including potentially much less sensitivity to bias from exposure-related loss to follow-up.

Case-Specular and Case-Crossover Studies

When the exposure under study is defined by proximity to an environmental source (e.g., a power line), it may be possible to construct a *specular* (hypothetical) control for each case, by conducting a “thought experiment”. Either the case or the exposure source is imaginarily moved to another location that would be equally likely were there no exposure effect; the case exposure level under this hypothetical configuration is then treated as the (matched) “control” exposure for the case [24]. When the specular control arises by examining the exposure experience of the case outside of the time in which exposure could be related to disease occurrence, the result is called a *case-crossover study*.

The classic *crossover* study is a type of experiment in which two (or more) treatments are compared, as in any experimental study. In a crossover study, however, each subject receives both treatments, with one following the other. Preferably, the order in which the two treatments are applied is randomly chosen for each subject. Enough time should be allocated between the two administrations so that the effect of each treatment can be measured and can subside before the other treatment is given. A persistent effect of the first intervention is called a *carryover effect*. A crossover study is only valid

to study treatments for which effects occur within a short induction period and do not persist, i.e., carryover effects must be absent, so that the effect of the second intervention is not intermingled with the effect of the first.

The *case-crossover* study is a case-control analogue of the crossover study [25]. For each case, one or more predisease or postdisease time periods are selected as matched “control” periods for the case. The exposure status of the case at the time of the disease onset is compared with the distribution of exposure status for that same person in the control periods. Such a comparison depends on the assumption that neither exposure nor confounders change with time in a systematic way.

Only a limited set of research topics are amenable to the case-crossover design. The exposure must vary over time within individuals, rather than stay constant. If the exposure does not vary within a person, then there is no basis for comparing exposed and unexposed time periods of risk within the person. Like the crossover study, the exposure must also have a short induction time and a transient effect; otherwise, exposures in the distant past could be the cause of a recent disease onset (a carryover effect).

Maclure [25] used the case-crossover design to study the effect of sexual activity on incident myocardial infarction. This topic is well suited to a case-crossover design because the exposure is intermittent and is presumed to have a short induction period for the hypothesized effect. Any increase in risk for a myocardial infarction from sexual activity is presumed to be confined to a short time following the activity. A myocardial infarction is an outcome well suited to this type of study because it is thought to be triggered by events close in time.

Each case in a case-crossover study is automatically matched with its control on all characteristics (e.g., sex and birth date) that do not change within individuals. A matched analysis of case-crossover data automatically adjusts for all such fixed confounders, whether or not they are measured. Control for measured time-varying confounders is possible using modeling methods for matched data. It is also possible to adjust case-crossover estimates for bias owing to time trends in exposure through use of longitudinal data from a nondiseased control group (case-time controls) [26]. Nonetheless, these trend adjustments themselves depend on additional

no-confounding assumptions and may introduce bias if those assumptions are not met [27].

Two-Stage Sampling

Another variant of the case-control study uses two-stage or two-phase sampling [28, 29]. In this type of study, the control series comprises a relatively large number of people (possibly everyone in the source population), from whom exposure information or perhaps some limited amount of information on other relevant variables is obtained. Then, for only a subsample of the controls, more detailed information is obtained on exposure or on other study variables that may need to be controlled in the analysis. More detailed information may also be limited to a subsample of cases. This two-stage approach is useful when it is relatively inexpensive to obtain the exposure information (e.g., by telephone interview), but the covariate information is more expensive to obtain (say, by laboratory analysis). It is also useful when exposure information already has been collected on the entire population (e.g., job histories for an occupational cohort), but covariate information is needed (e.g., genotype). This situation arises in cohort studies when more information is required than was gathered at baseline. This type of study requires special analytic methods to take full advantage of the information collected at both stages.

Case-Control Studies with Prevalent Cases

Case-control studies are sometimes based on prevalent cases rather than incident cases. When it is impractical to include only incident cases, it may still be possible to select existing cases of illness at a point in time. If the prevalence odds ratio in the population is equal to the incidence rate ratio, then the odds ratio from a case-control study based on prevalent cases can unbiasedly estimate the rate ratio. The conditions required for the prevalence odds ratio to equal the rate ratio are very strong, however, and a simple relation does not exist for age-specific ratios. If exposure is associated with duration of illness or migration out of the prevalence pool, then a case-control study based on prevalent cases cannot by itself distinguish exposure effects on disease incidence from the exposure association with disease duration or migration, unless the strengths of the latter associations are known. If the size of the exposed

or the unexposed population changes with time or there is migration into the prevalence pool, the prevalence odds ratio may be further removed from the rate ratio. Consequently, it is always preferable to select incident rather than prevalent cases when studying disease etiology.

Prevalent cases are usually drawn in studies of congenital malformations. In such studies, cases ascertained at birth are prevalent because they have survived with the malformation from the time of its occurrence until birth. It would be etiologically more useful to ascertain all incident cases, including affected abortuses that do not survive until birth. Many of these, however, do not survive until ascertainment is feasible, and thus it is virtually inevitable that case-control studies of congenital malformations are based on prevalent cases. In this example, the source population comprises all conceptuses, and miscarriage and induced abortion represent emigration before the ascertainment date. Although an exposure will not affect duration of a malformation, it may very well affect risks of miscarriage and abortion.

Other situations in which prevalent cases are commonly used are studies of chronic conditions with ill-defined onset times and limited effects on mortality, such as obesity and multiple sclerosis, and studies of health services utilization.

Conclusion

Epidemiologic research employs a range of study designs, including both experimental and nonexperimental studies. Among nonexperimental studies, cohort designs are sometimes thought to be inherently less susceptible to bias than case-control designs. Nonetheless, most of the biases that are associated with case-control studies are not inherent to the design, nor are cohort studies immune from bias. For example, recall bias will not occur in case-control study when exposure comes from records taken before disease onset, and selection bias can occur in cohort studies that suffer from loss to follow-up. No epidemiologic study is perfect, and this caution applies to cohort studies as well as case-control studies. A clear understanding of the principles of study design is essential for valid study design, conduct, and analysis, and for proper interpretation of results, regardless of the design.

Acknowledgment

This article is adapted from Modern Epidemiology, Third Edition, Rothman KJ, Greenland S, and Lash TL, eds., Lippincott Williams & Wilkins, 2008.

References

- [1] Greenland, S. (1985). Control-initiated case-control studies, *International Journal of Epidemiology* **14**, 130–134.
- [2] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Edition, Lippincott, Philadelphia, Chapter 21.
- [3] Wacholder, S., McLaughlin, J.K., Silverman, D.T. & Mandel, J.S. (1992). Selection of controls in case-control studies. I. Principles, *American Journal of Epidemiology* **135**, 1019–1028.
- [4] Wacholder, S., McLaughlin, J.K., Silverman, D.T. & Mandel, J.S. (1992). Selection of controls in case-control studies. I. Principles, *American Journal of Epidemiology* **135**, 1029–1041.
- [5] Wacholder, S., McLaughlin, J.K., Silverman, D.T. & Mandel, J.S. (1992). Selection of controls in case-control studies. I. Principles, *American Journal of Epidemiology* **135**, 1042–1050.
- [6] Lubin, J.H. & Gail, M.H. (1984). Biased selection of controls for case-control analyses of cohort studies, *Biometrics* **40**, 63–75.
- [7] Robins, J.M., Gail, M.H. & Lubin, J.H. (1986). More on biased selection of controls for case-control analyses of cohort studies, *Biometrics* **42**, 293–299.
- [8] Poole, C. (1986). Exposure opportunity in case-control studies, *American Journal of Epidemiology* **123**, 352–358.
- [9] Terry, M.B. & Neugut, A.L. (1998). Cigarette smoking and the colorectal adenoma-carcinoma sequence: a hypothesis to explain the paradox, *American Journal of Epidemiology* **147**, 903–910.
- [10] Poole, C. (1999). Controls who experienced hypothetical causal intermediates should not be excluded from case-control studies, *American Journal of Epidemiology* **150**, 547–551.
- [11] Lubin, J.H. & Hartge, P. (1984). Excluding controls: misapplications in case-control studies, *American Journal of Epidemiology* **120**, 791–793.
- [12] Berkson, J. (1946). Limitations of the application of 4-fold tables to hospital data, *Biometrics Bulletin* **2**, 47–53.
- [13] Cole, P. (1979). The evolving case-control study, *Journal of Chronic Diseases* **32**, 15–27.
- [14] Thomas, D.B. (1972). Relationship of oral contraceptives to cervical carcinogenesis, *Obstetrics and Gynecology* **40**, 508–518.
- [15] Kupper, L.L., McMichael, A.J. & Spirtas, R. (1975). A hybrid epidemiologic design useful in estimating relative risk, *Journal of the American Statistical Association* **70**, 524–528.
- [16] Miettinen, O.S. (1982). Design options in epidemiologic research: an update, *Scandinavian Journal of Work Environment Health* **8**(Suppl. 1), 7–14.
- [17] Flanders, W.D., DerSimonian, R. & Rhodes, P. (1990). Estimation of risk ratios in case-base studies with competing risks, *Statistics in Medicine* **9**, 423–435.
- [18] Wacholder, S. (1991). Practical considerations in choosing between the case-cohort and nested case-control design, *Epidemiology* **2**, 155–158.
- [19] Greenland, S. & Thomas, D.C. (1982). On the need for the rare disease assumption in case-control studies, *American Journal of Epidemiology* **116**, 547–553.
- [20] Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Application to cancer of the lung, breast and cervix, *Journal of the National Cancer Institute* **11**, 1269–1275.
- [21] Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analysis, *American Journal of Epidemiology* **125**, 761–768.
- [22] Sheehee, P.R. (1962). Dynamic risk analysis in retrospective matched-pair studies of disease, *Biometrics* **18**, 323–341.
- [23] Miettinen, O.S. (1976). Estimability and estimation in case-referent studies, *American Journal of Epidemiology* **103**, 226–235.
- [24] Zaffanella, L.E., Savitz, D.A., Greenland, S. & Ebi, K.L. (1998). The residential case-specular method to study wire codes, magnetic fields, and disease, *Epidemiology* **9**, 16–20.
- [25] Maclure, M. (1991). The case-crossover design: a method for studying transient effects on the risk of acute events, *American Journal of Epidemiology* **133**, 144–153.
- [26] Suissa, S. (1995). The case-time-control design, *Epidemiology* **6**, 248–253.
- [27] Greenland, S. (1996). Confounding and exposure trends in case-crossover and case-time-control designs, *Epidemiology* **7**, 231–239.
- [28] Walker, A.M. (1982). Anamorphic analysis: sampling and estimation for confounder effects when both exposure and disease are known, *Biometrics* **38**, 1025–1032.
- [29] White, J.E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease, *American Journal of Epidemiology* **115**, 119–128.

Related Articles

Absolute Risk Reduction

Epidemiology as Legal Evidence

History of Epidemiologic Studies

KENNETH J. ROTHMAN, SANDER GREENLAND
AND TIMOTHY L. LASH

Causal Diagrams

From their inception in the early twentieth century, causal systems models (more commonly known as *structural-equations models*) have been accompanied by graphical representations or path diagrams that provide compact summaries of qualitative assumptions made by the models. Figure 1 provides a graph that would correspond to any system of five equations encoding these assumptions:

1. independence of A and B
2. direct dependence of C on A and B
3. direct dependence of E on A and C
4. direct dependence of F on C and
5. direct dependence of D on B , C , and E .

The interpretation of “direct dependence” was kept rather informal and usually conveyed by causal intuition, for example, that the entire influence of A on F is “mediated” by C .

By the 1980s it was recognized that these diagrams could be reinterpreted formally as probability models, enabling use of graph theory in probabilistic inference, and allowing easy deduction of independence conditions implied by the assumptions [1]. By the 1990s it was further recognized that these diagrams could also be used as a formal tool for causal inference [2–5] and for illustrating sources of bias and their remedy in epidemiological research [6–15]. Given that the graph is correct, one can analyze whether the causal effects of interest (target effects or causal estimands) can be estimated from available data, or what additional observations are needed to validly estimate those effects. One can also see from the graph which variables should and should not be conditioned on (“controlled”) in order to estimate a given effect.

This article gives an overview of the following: (a) components of causal graph theory; (b) probability interpretations of graphical models; and (c) methodological implications of the causal and probability structures encoded in the graph, such as sources of bias and the data needed for their control. See **Causality/Causation** for a discussion of definitions of causation and statistical models for causal inference.

Basics of Graph Theory

As a befitting and well-developed mathematical topic, graph theory has an extensive terminology that once mastered provides access to a number of elegant results that may be used to model any system of relations. The term *dependence* in a graph, usually represented by connectivity, may refer to mathematical, causal, or statistical dependencies. The connectives joining variables in the graph are called *arcs*, *edge*, or *links*, and the variables are also called *nodes* or *vertices*. Two variables connected by an arc are *adjacent* or *neighbors* and arcs that meet at a variable are also adjacent. If the arc is an arrow, the tail (starting) variable is the *parent* and the head (ending) variable is the *child*. In causal diagrams, an arrow represents a “direct effect” of the parent on the child, although this effect is direct only relative to a certain level of abstraction, in that the graph omits any variables that might mediate the effect.

A variable that has no parent (such as A and B in Figure 1) is *exogenous* or *external*, or a *root* or *source* node, and is determined only by forces outside the graph; otherwise it is *endogenous* or *internal*. A variable with no children (such as D in Figure 1) is a *sink* or *terminal node*. The set of all parents of a variable X (all variables at the tail of an arrow pointing into X) is denoted $\text{pa}[X]$; in Figure 1, $\text{pa}[D] = \{B, C, E\}$.

A *path* or *chain* is a sequence of adjacent arcs. A *directed path* is a path traced out entirely along arrows tail-to-head. If there is a directed path from X to Y , X is an *ancestor* of Y and Y is a *descendant* of X . In causal diagrams, directed paths represent causal pathways from the starting variable to the ending variable; a variable is thus often called a *cause of its descendants* and an *effect of its ancestors*. In a *directed* graph the only arcs are arrows, and in an *acyclic* graph there is no feedback loop (directed path from a variable back to itself). Therefore, a directed acyclic graph (DAG) is a graph with only arrows for edges and no feedback loops (i.e., no variable is its own ancestor or its own descendant). A causal DAG represents a complete causal structure, in that all sources of dependence are explained by causal links; in particular, all common (shared) causes of variables in the graph are also in the graph.

A variable *intercepts* a path if it is in the path (but not at the ends); similarly, a set of variables S intercepts a path if it contains any variable intercepting

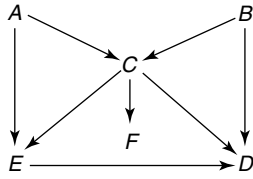


Figure 1 Example of causal diagram

the path. Variables that intercept directed paths are *intermediates* or *mediators* on the pathway. A variable is a *collider* on the path if the path enters and leaves the variable *via* arrowheads (a term suggested by the collision of causal forces at the variable). Note that being a collider is relative to a path; for example in Figure 1, C is a collider on the path $A \rightarrow C \leftarrow B \rightarrow D$ and a noncollider on the path $A \rightarrow C \rightarrow D$. Nonetheless, it is common to refer to a variable as a collider if it is a collider along any path (i.e., if it has more than one parent). A path is *open* or *unblocked* at noncolliders and *closed* or *blocked* at colliders; hence a path with no collider (like $E \leftarrow C \leftarrow B \rightarrow D$) is *open* or *active*, while a path with a collider (like $E \leftarrow A \rightarrow C \leftarrow B \rightarrow D$) is *closed* or *inactive*.

Some of the most important constraints imposed by a graphical model correspond to independencies arising from absence of open paths; e.g., absence of an open path from A to B in Figure 1 constrains A and B to be marginally independent (i.e., independent if no stratification is done). Nonetheless, the converse does not hold; i.e., presence of an open path allows but does not imply dependency. Independence may arise through cancellation of dependencies; as a consequence, even adjacent variables may be marginally independent; e.g., in Figure 1, A and E could be marginally independent if the dependencies through paths $A \rightarrow E$ and $A \rightarrow C \rightarrow E$ cancelled each other. The assumption of faithfulness, discussed below, is designed to exclude such possibilities.

Some authors use a bidirectional arc (two-headed arrow, $\leftrightarrow$) to represent the assumption that two variables share ancestors that are not shown in the graph; $A \leftrightarrow B$ then means that there is an unspecified variable U with directed paths to both A and B (e.g., $A \leftarrow U \rightarrow B$). Certain authors use $A \leftrightarrow B$ less specifically, however, in the way $A - B$ is used here to designate the presence of nondirected open paths

between A and B that are not intercepted by any variable in the graph.

Control: Manipulation versus Conditioning

The word *control* is used throughout science, but with a variety of meanings that are important to distinguish. In experimental research, to control a variable C usually means to manipulate or set its value. In observational studies, however, to control C (or more precisely, to control for C) more often means to condition on C , usually by stratifying on C or by entering C in a regression model. The two processes are very different physically and have very different representations and implications.

If a variable X is influenced by a researcher, the DAG would need an ancestor R of X to represent this influence. In the classical experimental case in which the researcher alone determines X , R , and X would be identical. In human trials, however, R more often represents just an *intention* to treat (with the assigned level of X), leaving X to be influenced by other factors that affect compliance with the assigned treatment R . In either case, R might be affected by other variables in the graph. For example, if the researcher uses age to determine assignments (an age-biased allocation), age would be a parent of R . Ordinarily, however, R would be exogenous as when R represents a randomized allocation.

In contrast, by definition in an observational study there is no such variable R representing the researcher's influence on X , and conditioning is substituted for experimental control. Conditioning on a variable C in a DAG can be represented by creating a new graph from the original graph to represent constraints on relations within levels (strata) of C implied by the constraints imposed by the original graph. This conditional graph can be found by the following sequence of operations [6], sometimes called *graphical moralization*.

1. If C is a collider, join ("marry") all pairs of parents of C by undirected arcs (here a dashed line will be used).
2. Similarly, if A is an ancestor of C and a collider, join all pairs of parents of A by undirected arcs.
3. Erase C and all arcs connecting C to other variables.

Figure 2 shows the graph derived from conditioning on C in Figure 1: the parents A and B of C are

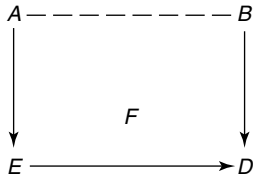


Figure 2 Graph derived from Figure 1 after conditioning on C

joined by an undirected arc, while C and all its arcs are gone. Figure 3 shows the result of conditioning on $F:C$ is an ancestral collider of F and so, again its parents A and B are joined, but only F and its single arc are erased. Note that, because of the undirected arcs, neither figure is a DAG.

Operations 1 and 2 reflect that if C depends on A and B through distinct pathways, the marginal dependence of A on B will not equal the dependence of A on B stratified on C (apart from special cases). To illustrate, suppose A and B are binary indicators (i.e., equal to 1 or 0), marginally independent, and $C = A + B$. Then among persons with $C = 1$, some will have $A = 1, B = 0$ and some will have $A = 0, B = 1$ (because other combinations produce $C \neq 1$). Thus when $C = 1$, A and B will exhibit perfect negative dependence: $A = 1 - B$ for all persons with $C = 1$.

Conditioning on a variable C closes open paths that pass through C . Conversely, conditioning on C opens paths that were blocked only at C or an ancestral collider A . In particular, conditioning on a variable may open a path even if it is not on the path, as with F in Figures 1 and 3. Thus the consequences of conditioning on C are often illustrated in the original (unconditional) graph by drawing a circle around C to denote the conditioning, then defining a path blocked by the circled C if C is a noncollider on the path, and by “marrying the parents” of C and

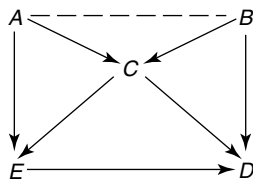


Figure 3 Graph derived from Figure 1 after conditioning on F

its collider ancestors (as in 1 and 2 above) to show the paths opened by conditioning. Thus if we circle C in Figure 1, it will block the $E-D$ paths $E \leftarrow C \leftarrow B \rightarrow D$ and $E \leftarrow A \rightarrow C \rightarrow D$ but unblock the path $E \leftarrow A \rightarrow C \leftarrow B \rightarrow D$ via the new arc between A and B . If we were to circle F but not C , no open path would be blocked, but A and B would again be joined by a dashed arc as in Figure 3.

A path is closed after conditioning on a set of variables S if S contains a noncollider along the path, or if the conditioning leaves the path closed at a collider; in either case S is said to block the path. Thus conditioning on S closes an open path if and only if S intercepts path, and opens a closed path if S contains no noncolliders on the path and every collider on the path is either in S or has a descendant in S . In Figure 1 the closed path $E \leftarrow A \rightarrow C \leftarrow B \rightarrow D$ will remain closed after conditioning on S if S contains A or B or if S does not contain C , but will be opened if S contains only C , F , or both.

Two variables (or sets of variables) in the graph are *d-separated* (or just separated) by a set S if, after conditioning on S , there is no open path between them. Thus in Figure 1, $\{A, C\}$ separates E from B , but $\{C\}$ does not (because conditioning on C alone results in Figure 2, in which E and B are connected via the open path A). In a DAG, $\text{pa}[X]$ *d-separates* X from every variable that is not affected by X (i.e., not a descendant of X). This feature of DAGs is sometimes called the *Markov condition*, expressed by saying the parents of a variable “screen off” the variable from everything but its effects. Thus in Figure 1 $\text{pa}[E] = \{A, C\}$, which separates E from B but not from D .

Bias and Confounding

There is considerable variation in the literature in the usage of terms like “bias,” “confounding,” and related concepts, which refer to dependencies that reflect more than just the effect under study. To capture these notions in a causal graph, we say that an open path between X and Y is a *biasing path* if it is not a directed path. The association of X with Y is then *unbiased* for the effect of X on Y if the only open paths from X to Y are the directed paths. Similarly, the dependence of Y on X is *unbiased given S* if, after conditioning on S , the open paths between X and Y are exactly (only and all) the directed paths in the starting graph. In such a case we say S is *sufficient* to

block bias in the X – Y dependence, and is *minimally sufficient* if no proper subset of S is sufficient.

Informally, confounding is a source of bias arising from causes of Y that are associated with but not affected by X (see **Confounding**). Thus we say an open nondirected path from X to Y is a *confounding path* if it ends with an arrow into Y . Variables that intercept confounding paths between X and Y are *confounders*. If a confounding path is present, we say *confounding* is present and that the dependence of Y on X is *confounded*. If no confounding path is present, we say the dependence is *unconfounded*, in which case the only open paths from X to Y through a parent of Y are directed paths. Similarly, the dependence of Y on X is *unconfounded* given S if, after conditioning on S , the only open paths between X and Y through a parent of Y are directed paths.

An unconfounded dependency may still be biased owing to nondirected open paths that do not end in an arrow into Y . These paths can be created when one conditions on a descendant of both X and Y . The resulting bias is called *Berksonian bias*, after its discoverer Joseph Berkson [16]. Most epidemiologists call this type of bias as “selection bias” [16]. Nonetheless, some writers (especially in econometrics) use “selection bias” to refer to confounding, while others call any bias created by conditioning as “selection bias” [14].

Consider a set of variables S that contains no effect (descendant) of X or Y . S is *sufficient* to block confounding if the dependence of Y on X is unconfounded given S . “No confounding” thus corresponds to sufficiency of the empty set. A sufficient S is called *minimally sufficient* to block confounding if no proper subset of S is sufficient. The initial exclusion from S of descendants of X or Y in these definitions arises first, because conditioning on X descendants can easily block directed (causal) paths that are part of the effect of interest, and second, because conditioning on X or Y descendants can unblock paths that are not part of the X – Y effect, and thus create new bias.

A *back-door path* from X to Y is a path that begins with a parent of X (i.e., leaves X from a “back door”) and ends at Y . A set S then satisfies the *back-door criterion* with respect to X and Y if S contains no descendant of X and there are no open back-door paths from X to Y after conditioning on S . In a DAG, the following simplifications occur.

1. All biasing paths are back-door paths.
2. The dependence of Y on X is unbiased whenever there are no open back-door paths from X to Y .
3. If X is exogenous, the dependence of any Y on X is unbiased.
4. All confounders are ancestors of either X or of Y .
5. A back-door path is open if and only if it contains a common ancestor of X and Y .
6. If S satisfies the back-door criterion, then S is sufficient to block X – Y confounding.

These conditions do not extend to non-DAGs like Figure 2. Also, although $\text{pa}[X]$ always satisfies the back-door criterion and hence is sufficient in a DAG, it may be far from minimal sufficient. For example, there is no confounding and hence no need for conditioning whenever X separates $\text{pa}[X]$ from Y (i.e., whenever the only open paths from $\text{pa}[X]$ to Y are through X).

As a final caution, we note that the biases dealt with by the above concepts are only confounding and selection biases. Biases due to measurement error and model-form misspecification require further structure to describe them.

Statistical Interpretations and Applications

A joint probability distribution for the variables in a graph is *compatible* with the graph if two sets of variables are independent given S whenever S separates them. For such distributions, two sets of variables will be statistically unassociated if there is no open path between them. Many special results follow for distributions compatible with a DAG. For example, if in a DAG, X is not an ancestor of any variable in a set T , then T and X will be independent given $\text{pa}[X]$. A distribution compatible with a DAG can thus be reduced to a product of factors $\text{Pr}(x|\text{pa}[X])$, with one factor for each variable X in the DAG; this is sometimes called the *Markov factorization* for the DAG. When X is a treatment, this condition implies that the probability of treatment is fully determined by the parents of X , $\text{pa}[X]$.

Suppose now we are interested in the effect of X on Y in a DAG, and we assume a probability model compatible with the DAG. Then, given a sufficient set S , the only source of association between X and Y within strata of S will be the directed paths from X to Y . Hence, the *net effect* of $X = x_1$ versus $X = x_0$ on Y when $S = s$ is defined as $\text{Pr}(y|x_1, s) - \text{Pr}(y|x_0, s)$,

the difference in risks of $Y = y$ at $X = x_1$ and $X = x_0$. Alternatively, one may use another effect measure such as the risk ratio $\Pr(y|x_1, s)/\Pr(y|x_0, s)$. A *standardized effect* is a difference or ratio of weighted averages of these stratum-specific $\Pr(y|x, s)$ over S , using a common weighting distribution. The latter definition can be generalized to include intermediate variables in S by allowing the weighting distribution to causally depend on X . Furthermore, given a set Z of intermediates along all directed paths from X to Y with $X-Z$ and $Z-Y$ unbiased, one can produce formulas for the $X-Y$ effect as a function of the $X-Z$ and $Z-Y$ effects (“front-door adjustment” [2, 4]).

The above form of standardized effect is identical to the forms derived under other types of causal models, such as potential outcome models. When S is sufficient, some authors go so far as to identify the $\Pr(y|x, s)$ with the distribution of potential outcomes given S . There have been objections to this identification on the grounds that not all variables in the graph can be manipulated, and that potential-outcome models do not apply to nonmanipulable variables. However, the objection loses force when X is an intervention variable. In that case, sufficiency of a set S implies that the potential-outcome distribution equals $\sum_s \Pr(y|x, s)\Pr(s)$, the risk of $Y = y$ given $X = x$ standardized to the S distribution.

Identification of Effects and Biases

In order to check the sufficiency and identify minimally sufficient sets of variables given a graph of the causal structure, one needs to see only whether the open paths from X to Y after conditioning are exactly the directed paths from X to Y in the starting graph. Mental effort may then be shifted to evaluating the reasonableness of the causal independencies encoded by the graph, some of which are reflected in conditional independence relations. This property of graphical analysis facilitates the articulation of necessary background knowledge for estimating effects, and eases teaching of algebraically difficult identification conditions.

As an example, spurious sample associations may arise if each variable affects selection into the study, even if those selection effects are independent. This phenomenon is a special case of the collider-stratification effect illustrated earlier. Its presence is easily seen by starting with a DAG that includes a

selection indicator $F = 1$ for those selected, or z otherwise, as well as the study variables, and then noting that we are always forced to examine associations within the $F = 1$ stratum (i.e., by definition, our observations stratify on selection). Thus, if selection (F) is affected by multiple causal pathways, we should expect selection to create or alter associations among the variables.

Figure 4 displays a situation common in randomized trials, in which the net effect of E on D is unconfounded, despite the presence of an unmeasured cause U of D . Unfortunately, a common practice in health and social sciences is to stratify on (or otherwise adjust for) an intermediate variable F between a cause E and effect D , and then claim that the estimated (F -residual) association represents that portion of the effect of E on D not mediated through F . In Figure 4 this would be a claim that, upon stratifying on F , the $E-D$ association represents the direct effect of E on D . Figure 5, however, shows the graph conditional on F , in which we see that there is now an open path from E to D through U , and hence the residual $E-D$ association is confounded for the direct effect of E on D .

The $E-D$ confounding by U in Figure 5 can be seen as arising from the confounding of the $F-D$ association by U in Figure 4. In a similar fashion, conditioning on C in Figure 1 opens the confounding path through A and B in Figure 2; this path can be seen as arising from the confounding of the $C-E$

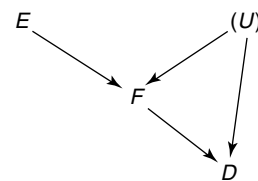


Figure 4 Graph with randomized E and confounded intermediate F

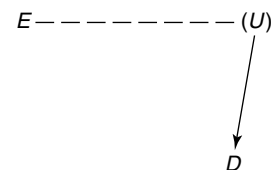


Figure 5 Graph derived from Figure 4 after conditioning on F

association by A and the $C-D$ association by B in Figure 1. In both examples, further stratification on either A or B blocks the created path and thus removes the new confounding.

Bias from conditioning on a collider or its descendant has been called *collider bias* [12, 15]. Starting from a DAG, there are two distinct forms of this bias: confounding induced in the conditional graph (Figures 2, 3, and 5), and Berksonian bias from conditioning on an effect of X and Y . Both biases can in principle be removed by further conditioning on variables along the biasing paths from X to Y in the conditional graph. Nonetheless, the starting DAG will always display ancestors of X or Y that, if known, could be used to remove confounding; in contrast, no variable need appear or even exist that could be used to remove Berksonian bias.

Figure 4 also provides a schematic for estimating the $F-D$ effect, as in randomized trials in which E represents assignment to or encouragement toward treatment F . One can put bounds on confounding of the $F-D$ association, and with additional assumptions remove it entirely through formulas that treat E as an *instrumental variable* (a variable associated with F such that every open path to D includes an arrow pointing into F) [4, 7, 15].

Questions of Discovery

While deriving statistical implications of graphical models is uncontroversial, algorithms that claim to discover causal (graphical) structures from observational data have been subject to strong criticism [17, 18]. A key assumption in certain “discovery” algorithms is a converse of compatibility called *faithfulness* [5]. A compatible distribution is *faithful* to the graph (or *stable*) if for all X, Y , and S , X and Y are independent given S only when S separates X and Y (i.e., the distribution entails no independencies other than those implied by graphical separation). Faithfulness implies that minimal sufficient sets in the graph will also be minimal for consistent estimation of effects. Nonetheless, there are real examples of near cancellation (e.g., when confounding obscures a real effect), which make faithfulness questionable as a routine assumption. Fortunately, faithfulness is not needed for the uses of the graphical models discussed here.

Whether or not one assumes faithfulness, the generality of graphical models is purchased with limitations on their informativeness. Causal diagrams

show whether the effects can be estimated from the given information, and can be extended to indicate effect direction when that is monotone [19]. Nonetheless, the nonparametric nature of the graphs implies that parametric concepts like effect-measure modification (heterogeneity of arbitrary effect measures) cannot be displayed by the basic graphical theory. Similarly, the graphs may imply that several distinct conditionings are minimal sufficient (e.g., both $\{A, C\}$ and $\{B, C\}$ are sufficient for the ED effect in Figure 1), but offer no further guidance on which to use. Open paths may suggest the presence of an association, but that association may be negligible even if nonzero. For example, bounds on the size of direct effects imply more severe bounds on the size of effects mediated in multiple steps (indirect effects), with the bounds becoming more severe with each step. As a consequence, there is often good reason to expect certain phenomena (such as the conditional $E-D$ confounding shown in Figures 2, 3, and 5) to be small in epidemiological examples. Thus, when quantitative information is used, graphical modeling becomes more a schematic adjunct than an alternative to causal modeling.

Other Reviews

Full technical details of causal diagrams and their relation to causal inference can be found in the books by Pearl [4] and Spirtes *et al.* [5]. Less technical reviews geared toward health scientists include Greenland *et al.* [6], Greenland and Brumback [9], and Glymour and Greenland [15].

References

- [1] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann Publishers, San Mateo.
- [2] Pearl, J. (1995). Causal diagrams for empirical research (with discussion), *Biometrika* **82**, 669–710.
- [3] Pearl, J. & Robins, J.M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables, *Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, San Francisco, pp. 444–453.
- [4] Pearl, J. (2000). *Causality*, Cambridge University Press, New York.
- [5] Spirtes, P., Glymour, C. & Scheines, R. (2001). *Causation, Prediction, and Search*, 2nd Edition, MIT Press, Cambridge.
- [6] Greenland, S., Pearl, J. & Robins, J.M. (1999). Causal diagrams for epidemiologic research, *Epidemiology* **10**, 37–48.

- [7] Greenland, S. (2000). An introduction to instrumental variables for epidemiologists, *International Journal of Epidemiology* **29**, 722–729 (Erratum: 2000, **29**, 1102).
- [8] Robins, J.M. (2001). Data, design, and background knowledge in etiologic inference, *Epidemiology* **12**, 313–320.
- [9] Greenland, S. & Brumback, B.A. (2002). An overview of relations among causal modelling methods, *International Journal of Epidemiology* **31**, 1030–1037.
- [10] Cole, S. & Hernán, M.A. (2002). Fallibility in estimating direct effects, *International Journal of Epidemiology* **31**, 163–165.
- [11] Hernán, M.A., Hernandez-Diaz, S., Werler, M.M. & Mitchell, A.A. (2002). Causal knowledge as a prerequisite for confounding evaluation, *American Journal of Epidemiology* **155**, 176–184.
- [12] Greenland, S. (2003). Quantifying biases in causal models: classical confounding versus collider-stratification bias, *Epidemiology* **14**, 300–306.
- [13] Jewell, N.P. (2004). *Statistics for Epidemiology*, Chapman & Hall/CRC, Boca Raton, section 8.3.
- [14] Hernán, M.A., Hernandez-Diaz, S. & Robins, J.M. (2004). A structural approach to selection bias, *Epidemiology* **15**, 615–625.
- [15] Glymour, M.M. & Greenland, S. (2008). Causal diagrams, in *Modern Epidemiology*, 3rd Edition, K.J. Rothman, S. Greenland & T.L. Lash, eds, Lippincott Williams & Wilkins, Philadelphia, Chapter 12.
- [16] Rothman, K.J., Greenland, S. & Lash, T.L. (2008). *Modern Epidemiology*, 3rd Edition, Lippincott Williams & Wilkins, Philadelphia.
- [17] Freedman, D.A. & Humphreys, P. (1999). Are there algorithms that discover causal structure? *Synthese* **121**, 29–54.
- [18] Robins, J.M. & Wasserman, L. (1999). On the impossibility of inferring causation from association without background knowledge, in *Computation, Causation, and Discovery*, C. Glymour & G. Cooper, eds, AAAI Press/The MIT Press, Menlo Park, pp. 305–321.
- [19] VanderWeele, T.J. & Robins, J.M. (2008). Signed directed acyclic graphs for causal inference, *Journal of the Royal Statistical Society series B*, in press.

Related Articles

Compliance with Treatment Allocation

SANDER GREENLAND AND JUDEA PEARL

Causality/Causation

While few people lack an intuitive understanding of cause and effect, a precise philosophical definition is challenging, and a practical epistemology for detecting causal relations even more so. In the practical realms of worldly sciences, methods have been established for conceptualizing, detecting, and modeling causation, some quite useful but others largely distracting.

Definition of Cause

Many attempts to define *cause* are circular, relying on synonyms (e.g., “a condition that produces an outcome”), and thus vacuous. The most widely accepted definition of cause is based on the concept of counterfactuals, and written analysis of it dates back to Hume’s [1] “An Enquiry Concerning Human Understanding”, which considers the fundamental and elusive nature of understanding causation:

“The only immediate utility of all sciences, is to teach us, how to control and regulate future events by their causes. Our thoughts and enquiries are, therefore, every moment, employed about this relation: Yet so imperfect are the ideas which we form concerning it, that it is impossible to give any just definition of cause, except what is drawn from something extraneous and foreign to it. Similar objects are always conjoined with similar. Of this we have experience. Suitably to this experience, therefore, we may define a cause to be an object, followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or in other words, where, if the first object had not been, the second never had existed.”

Expanding on the final quoted sentence (though setting aside the previous sentence, which others have simply ignored – see below), a cause of an effect is an event or state of the world such that the effect occurs given that the cause occurs or exists, while the effect does not occur if the cause does not occur or exist, *ceteris paribus*. This definition is generally referred to as the *counterfactual definition* because either the cause exists or it does not in the actual state of the world, and the definition depends on drawing conclusions about the status of the effect in a world where the status of the cause is contrary to actual fact.

Notice several implications of the definition: The *ceteris paribus* condition (i.e., that all other things are equal) allows an event or state of the world (hereafter, this conjunction is simplified to *agent*) to cause an effect under some circumstances and fail to do so in others; the latter include other circumstances where the effect does not occur in the presence of the potential causal agent, or where it occurs regardless of that agent. If the effect never occurs when the agent is absent, the agent is called *necessary* and if the effect always occurs whenever the agent occurs, the agent is called *sufficient*. In practical contexts, the terms *necessary* and *sufficient* are generally used conditionally rather than universally, with the implication that some intuitive list of worldly conditions is held constant. At a minimum, this takes the form of assuming that nothing varies beyond the range that we believe it can vary (e.g., when assessing causes of disease, we assume that a mysterious force will not radically alter human physiology).

In most practical science (i.e., once we depart from mathematical descriptions of symbolic logic), the definition of both the agent and the outcome become imprecise. For example, is an event the same event if it occurs 5 minutes sooner or is it slightly larger or slightly more intense? Is a year sooner, much larger or much more intense? If smoking accelerates someone’s fatal heart failure by 10 years, we would agree that it was a cause of death, but if it does so by 5 min, it seems inappropriate to attribute the death to smoking. The answers to these questions determine whether an agent is a cause, because if an outcome changes enough in the presence of an agent that it is considered a different event, then the agent caused the new version of the outcome to occur (and the old one to not occur). Clearly a large enough change constitutes a different outcome. However, if we define even the smallest changes as different outcomes, then almost everything would cause almost every event that is subsequent and proximate to it, and so the concept loses most of its practical value. Every context offers practical, generally fuzzy, definitions of “same” and “different” events that fundamentally affect the definition of cause.

To fully understand the counterfactual concept of causation, special attention should be paid to the case of *overdetermination*. If there exist in a given situation multiple agents (or sets of agents), each of which is sufficient for a particular effect given the state of the world, then nothing can be said to have

caused the effect by the counterfactual definition. The counterfactual absence of any particular agent or set would not change the outcome. (Note that this means that a sufficient agent may not always be a “sufficient cause”, though the latter terminology is commonly used without regard for the presence of other sufficient agents.) This complicates practical implications and defies intuition in some contexts, and so often each of the multiple agents is called a cause, as it would be if the others were absent in a given situation. Altering the formal definition to allow for overdetermination complicates it substantially, and so it is often just left implicit that the counterfactual thought exercise about one agent can be carried out as if coincidentally present sufficient causes were absent. The “causal pie” model, below, helps illustrate some of these concepts.

The counterfactual definition of causation is not universally accepted, most often challenged on the grounds that invoking what happens in an impossible state of the world is troublesome. Alternative definitions often focus on a consistent pattern of one event following another, as in the penultimate sentence of the above passage from Hume. However, such proposed alternative definitions are usually regarded as indistinguishable from the counterfactual definition in a practical sense (since neither the counterfactual nor the universality of a pattern can be observed, and thus both are empirically based on only a finite number of observed events that form some pattern). This is evidenced by the unacknowledged equation of the two by Hume and others who quote him. When causation is accepted as fundamentally nonobservable, then any definition must incorporate something nonobservable. Given this, practical observers sometimes treat the definition as purely a matter of intuition, and refer to the counterfactual or other conceptualizations as models (i.e., simplifications that allow for extraction of information about the underlying construct), and not the definition *per se*. Throughout this chapter, the widely accepted and heuristically useful counterfactual conceptualization will be implicitly used; nothing depends on whether it is considered the definition or a model.

Empirical Implications

To observe causation, we would like to observe both the actual and counterfactual states of the world.

Since that is not possible, we usually try to observe two groups of potential objects of causation (e.g., populations of people) that are as identical as possible with respect to their response to an agent of interest, with the agent present in one group but not the other. For identical molecules in a chemistry experiment, we comfortably assume that the populations are identical. When people are the object of causation, it is difficult to justify that assumption; even sequentially comparing the same person with and without the agent, in situations where this is possible, violates the *ceteris paribus* condition due to changes in other variables over time.

Observing different groups of people, some of whom have an “*exposure*” (i.e., the agent is present at a particular level of intensity) and some of whom do not, introduces the problem of *confounding*, which exists when the two groups differ in their frequency of the outcome, even without any effect of the agent. For example, people who choose to smoke have characteristics other than smoking that make them less healthy, on average, than those who do not; so a comparison of “nonsmoking smokers” (people who actually smoke, but observed in the counterfactual state of not having smoked) and actual nonsmokers would show a difference in health outcomes. A simple comparison of smokers to nonsmokers would, thus, imply that smoking is more unhealthy than it actually is. Statistical models that incorporate other variables to “control for” their effects may partially correct for this problem. These covariates are generally called *confounders*, though only some of them actually are, as discussed below in the discussion of causal diagrams (see **Confounding**). The emphasis on confounders, however, is the source of much confusion about the nature of causation and confounding: Confounding is often mistakenly thought of as the difference between groups (with respect to the outcome, independent of the agent) caused by the confounders, when actually it is merely the difference, without regard to its cause. The covariates are not so much confounders as they are “unconfounders” [2], variables included in a statistical model with the hope that they reduce confounding, and whatever confounding the agents measured by those variables actually *causes* requires an analysis as complicated as any other study of causation.

Another solution to the problem of confounding is to randomize a population into groups and intervene by assigning each group to have the agent or

not (e.g., a clinical trial that gives some people a drug and others a placebo – see **Randomized Controlled Trials**). This replaces systematic confounding due to unknown systematic differences that are associated with the distribution of the agent in the world, as it exists with random confounding which can be assessed statistically. The “random error” in a randomized trial is seldom recognized as confounding, but attention to the concepts of causation and confounding shows that it is confounding a random allocation will, on average, generate groups that, apart from the effect of the agent, have the same outcomes. However, the actual allocation in a particular case will likely produce groups that are not exactly this average, and thus fit the definition of confounding. Because this confounding is generated through a known random process, its distribution can be assessed and reported using familiar confidence intervals and test statistics. These statistics are often calculated and reported when there is no randomization, but there they only represent random sampling deviations from the unknown actual confounded association, not from the unconfounded association as in a randomized experiment.

Randomized experiments are often erroneously called the *gold standard* research design in health science and similar fields. This misconception appears partially due to the incorrect belief that a randomized experiment allows direct observation of causation, while nonrandomized observational studies do not. In fact, direct observation of causation remains impossible in randomized experiments; allocating an agent at random to two groups of people and contrasting their response in real-world settings is still only a proxy for knowing the response that would have occurred in a single group had their exposure status been changed and all else held equal. For this reason, the value of any study for assessing causation, regardless of design, must be judged based on how well it approximates the counterfactual contrast. Randomized experiments have the advantage that they make confounding follow a predictable statistical distribution, rather than being unknown and potentially very misleading. However, in other ways, randomized trials are an inferior approximation for the counterfactual, as compared to certain observational studies: they can have nonrandom lack of compliance with the assigned exposure or nonrandom subject loss, and the study population and version of the agent

under experimental conditions often differ from the counterfactual contrast we are actually interested in.

Categorizing Causal Response Types

Based on ideas from Miettinen [3], Greenland and Robins introduced four basic causal response types, with respect to a particular agent: “doomed”, “causative”, “preventive”, and “immune” [4]. These were defined in the context of epidemiology, but the concept easily generalizes. Individuals who are doomed will have the outcome whether the agent is present or not, and those who are immune will not have the outcome, regardless of exposure to the agent. Thus, in the given state of the world, for these two types of individuals, the agent is not a cause of the outcome. The causative types are those who have the outcome if the agent is present, but not otherwise, the standard definition of the agent as a cause. The preventive type includes those for whom the absence of the agent causes the outcome (which is the intuitive definition of an agent that prevents an outcome).

Causal response types can only be inferred or modeled, since empirical observation of an individual’s exposure and subsequent outcome status can only narrow the possible types to two, and distinguishing between them requires knowing a counterfactual. For example, if we observe that an individual was exposed to an agent and later developed the outcome, then the individual was either doomed or had a causative response, but we can only imagine what the outcome would have been had the individual not been exposed, because we cannot repeat the observation modifying the exposure while holding all else constant. Similarly, if we observe that an individual was not exposed to an agent and did not later develop the outcome, we cannot determine if the individual was immune or had a preventive response. Maldonado and Greenland tied this causal response classification to the counterfactual definition of cause and showed the role of these concepts in understanding basic epidemiologic concepts [5]. In particular, they illustrated that the aim in etiologic research should be to find two groups that are identical in terms of their responses to the agent (are “perfect substitutes”); with that achieved, the outcomes in one group are what would have been observed in the comparison group if the exposure distributions had been exchanged, and the difference is caused by the agent.

Simple Causal Relationships

Evolution favors the development of some intuitive recognition of causation in an organism with decision-making capacity, because of its obvious survival advantage. However, informal observation and formal research in cognitive psychology suggest that even in humans this intuition is quite limited. There tends to be an intuitive recognition of only the simplest causal relationships, and a creation of perceived causation where it does not exist (e.g., superstition). Formal models, including some presented below, are often needed (though not always sufficient) when intuition is misleading.

The simplest causal relationships involve a single agent that is necessary and sufficient for a particular outcome. This view is overly simplistic for the causation of most natural phenomena, and indeed, represents a case where the agent and outcome are in some sense part of the same phenomenon. It is worth mentioning primarily because this unrealistic relationship is often implicitly invoked in discussions of causation, usually introducing confusion or obscuring scientific challenges. For example, discussions of disease often ask whether *the* cause is environmental or genetic, as if one agent were necessary and sufficient and others irrelevant, even when evidence shows no known class of agents is either necessary or sufficient.

Slightly more complicated and more realistic causal relationships are those where an agent is either necessary or sufficient for producing a particular outcome, but not both. Sufficiency describes the simple causal relationships that we seldom think of as causation (e.g., if I grasp a coffee cup and raise my hand, the cup will rise), as well as worldly “laws” and physical forces (e.g., gravitational attraction). An example of a necessary agent can be found in infectious diseases, when the particular disease occurs only in the presence of the corresponding infectious agent. A careful examination shows that this (at best) borders on circularity, since even if every symptom of the given infectious disease were present, if the infectious agent could be shown to be absent we would declare that the particular infectious disease was absent, the necessity of the cause thus being created by definition. Nevertheless, if we do not push too hard on this concept, a necessary agent is a useful starting point for modeling causation, and is the basis for one of

the earlier attempts to assess causation in complicated worldly phenomena.

Henle–Koch Postulates

The Henle–Koch postulates for identifying causal agents of infectious diseases are based on a single-agent causal model [6]. These postulates, dating from 1884, are as follows:

1. The parasite occurs in every case of the disease in question. . .
2. It occurs in no other disease. . .
3. After being. . .isolated and. . .grown. . . it can produce the disease anew.

These postulates are useful for overcoming the view of causation as isomorphic (since the parasite is not necessarily sufficient for the disease). However, even for an infectious disease, the simple necessary cause model has flaws. The requirement that a parasite be specific to a single disease is sometimes misleading, particularly because a spectrum of severity of symptoms arguably represents different disease outcomes. More important from the perspective of causal modeling, when the parasite is not sufficient for disease to occur, the postulates offer no conceptualization of factors that make someone susceptible to the infection, which are also causes. Finally, chronic infections are now known to sometimes cause chronic diseases such as cancer, which can also be caused independently by other agents, and so infectious agents cause diseases for which they are not necessary.

Multifactorial Concepts of Cause

For most effects, there is no single sufficient causal agent, and a multiplicity of agents is required to form a sufficient combination. This creates substantial complications for identifying causal relations, many of which are not obvious without modeling. In theory, the complicated interaction of causes is well understood in both biological and social sciences. A 1960 textbook of epidemiologic methods presented a causal diagram called the *web of causation* to depict how multiple factors lead to disease [7]. An illustration like a web emphasizes the complexity of causal relationships that may not be obvious to many

observers. Simpler and more formalized depictions, such as those presented below, are more useful still, serving as models of the implications of causation.

Interaction of Causes and the Causal Pie Model

In 1976, Rothman presented what he later dubbed the causal pie model to illustrate that a given disease results from one or more sufficient sets of component causes and disease occurs when a person has been exposed to all components of a sufficient set [8, 9]. This model depicts each sufficient set of agents as forming a circle, with each agent depicted as a slice of pie (Figure 1). The specifics of the illustration contain no information beyond the list of component causes (e.g., the size of the pie slice does not matter), but it has been adopted as a standard presentation in epidemiology. In some presentations, one slice of the pie is labeled U, representing posited unknown cofactors that are necessary for the included known factors to be collectively sufficient, because there are relatively few causal mechanisms where the known agents are sufficient [10].

This simple depiction parsimoniously illustrates that the component causes in a sufficient set interact to produce an outcome, and thus demonstrates several important aspects of causation that are not intuitive to most observers. The simplest is that because multiple agents combine to cause a single effect, and changing the status of any one of them would eliminate the effect. This means that the presence of other agents alters whether a particular agent is a cause of an outcome in a particular instance. An agent actually causes a particular outcome that it has the potential to cause only if the *causal complement* (that is, all the other elements in a pie) for at least one pie is present. The observable degree of interaction between any two component causes depends on how many sufficient sets contain both of those agents and how often the other elements in those sets are all present.

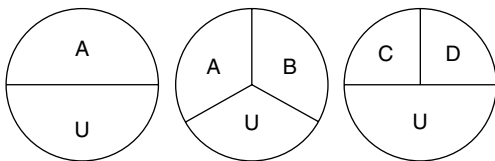


Figure 1 Causal pie diagrams

Conceptualizing sets of causal complements also allows prediction of *effect modification*, wherein an agent is expected to have different effects depending on the presence of other known variables (e.g., if we think that some of the causal complements for the agent, as eating meat as a cause of colon cancer includes being older, then we should expect the agent to cause the outcome in a smaller portion of younger people compared to older people). With this insight, it is possible to stratify a model based on an effect modifier or otherwise predict and account for the different effects of the same agent between or within studies.

Causal pies can be used to illustrate the concept of *attributable fraction*, the portion of all outcomes in a population that are caused by a particular agent (which can be determined by counting up the individuals for which (a) one or more causal pies are filled when the agent is present, but (b) no pies are filled if the agent is absent). This also illustrates that while the fraction of events attributable to a single-component cause cannot exceed 100%, there is no bound to the sum of these fractions across all causes (that is, it is to be expected that we will discover causes of a disease such that the proportion of cases of disease attributed to a cause, summed across all causes, is far more than 100% because multiple causes are usually identified for each case).

An additional lesson from the causal pie model is that the quantified measure of the agent–outcome association may be large or small and still represent causation. Quantitative estimates of causal relationships depend on how often causal complements are present (and thus how often an agent that is potentially a cause actually is a cause), as well as the prevalence of sufficient causal sets that do not include the agent in question. In concrete terms, the probability that an individual’s smoking will cause him to get oral cancer changes with the presence of complementary agents; there is a causal pie that includes smoking (plus unknown factors) and another that includes smoking and heavy drinking (plus unknown factors), and thus whether smoking is causal sometimes depends on whether the individual drinks heavily. However, the risk ratio for the population (i.e., the multiplicative increase in the probability of the outcome) also depends on the rate of oral cancer that exists when smoking is absent, since a higher baseline rate in the absence of the agent lowers the relative risk. So in the West, where oral cancer is quite rare

in nonsmokers, smoking increases the baseline risk by a factor of about 10, but elsewhere in the world, where the baseline incidence is as much as 10 times higher, the same absolute risk from smoking may only represent a twofold relative risk. However, both of these relative risks represent causation, and may even represent the same number of cases caused by smoking in a population of a given size.

As noted above, whether changing the timing of an event makes it a different event is a matter of subjective definition. Another element of timing that can be recognized using the causal pie model is the induction period, the time between a component cause becoming present and the outcome event. This period depends on how long it takes to complete the sufficient set. For example, for disease that results from the combination of chronic infection and a period of immunosuppression, the infection may be present for a period until a bout of immunosuppression occurs, that period being the induction period; the induction period for the immunosuppression, in that case, is zero (because it is assumed that the two agents form a sufficient set and the other agent is already present). This sequence should not be confused with a series of events such that X causes Y and then Y causes Z, which cannot be effectively illustrated in a model like the pie that is only an unordered list.

Probability and Causation

Conceptualizing causation in an empirical scientific context requires the use of probability. Because we cannot generally observe all determinants of particular events such that we can predict their future occurrence with certainty, there is generally some degree of uncertainty regarding the outcome of known or suspected causal agents. In the context of the causal pie, if the status of a causal complement for an agent is only known probabilistically, then whether an agent is a cause or not can only be known probabilistically. If an agent is absent and we consider the counterfactual state of it being present, only the probability of the outcome that would result can be known. Thus, in a population, the expected value of the frequency of the outcome (and its statistical variance) under the counterfactual presence of the agent can be assessed.

Empirically, the reasoning flows in the opposite direction: The frequency or proportion of individuals with an outcome is observed in two groups with

different status for the agent, and the probability that the agent causes the outcome for an average individual is inferred to be the estimated proportion of the group for which it is causal. Often, this measure of effect is expressed in terms of the increase in the probability of the outcome caused by the agent. Since the outcome either occurs or does not, for a given individual, the estimated increase in the probability of an outcome is the average effect at the population level, when the causative types are added to the doomed types. Each individual with the outcome is still either causative or doomed, and we cannot know which, but we can estimate the probability based on the relative sizes of those groups. Recognizing this can eliminate some confusion that results from simplistic concepts of causation among people who are not experts in scientific reasoning. For example, physicians often declare that a particular patient's cancer was caused by a particular exposure, presumably because they have learned that the exposure increases the probability of getting that cancer; what they fail to understand is that the patient in question might have been doomed with respect to that agent.

Note that inherent in this discussion is a conceptualization of probability as subjective, empirically estimated based on frequency, but stemming from observer ignorance. Under this conceptualization, for each individual an agent is either causal or not, so probabilities of causation between zero and one exists only in the assessment of an observer who does not know the individual's true type. Causation as a deterministic process, as portrayed in the causal pie model or the conceptualization of causal response types, can translate into empirical models that incorporate probability by equating the probability that an agent causes an outcome in a given instance with the probability that the causal complement is present.

Causal Pathway Diagrams

Diagrams of hypothesized causal pathways typically use arrows connecting boxes or words to illustrate assumed relations between causal factors and specific outcomes. Such diagrams contain information that is absent in models that just list component causes, like causal pies. Causal pathway models show the sequence and direction of hypothesized relations among variables and are particularly useful

for distinguishing between causation and noncausal association, as well as how proximal a cause is (whether it causes the outcome by affecting other posited causal agents, or without such intermediates), and for identifying variables that should be measured and controlled to obtain unconfounded estimates. However, such diagrams usually lack information that a pie model provides about interactions of causes and about why a measure of association might be large or small.

Causal pathway diagrams include formal, mathematically analyzable diagrams as well as informal illustrations, such as webs of causation. Figure 2 portrays hypothesized relations between an exposure, an outcome, and other variables. This type of diagram demonstrates whether other factors are confounders, intervening factors, or exist on a separate pathway from the agent in question (in the latter case, the diagrams usually provide no way to distinguish elements of the causal complement from independent sufficient causes). The example contains clear information about the hypothesized causal pathway, though as drawn it does not show whether a relationship is positive/causal or negative/protective (plus and minus signs are sometimes added to an arrow to indicate this), or strong or weak. Agent “exposure” is a cause of the outcome in the figure because it causes a change in agent “intervening factor”, which in turn causes a change in “health outcome”. If we acted to hold “intervening factor” constant, regardless of the status of “exposure”, then “exposure” would cease to be a cause of the outcome. “Confounder” is a cause of the outcome, and still would be if we held the “intervening factor” constant because it would no longer be a cause *via* “exposure”, but it would still be a cause *via* the direct pathway.

The most useful mathematical causal modeling using arrow-and-box models is based on directed acyclic graphs (DAGs), which impose intuitive rules that allow more precise mathematical analysis. In particular, the lines connecting entries are single-headed arrows (in the case of causal diagrams, this

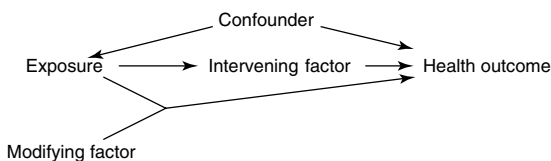


Figure 2 A casual pathway diagram

means they are all hypothesized causal relationships) and the arrows form no circular paths. These rules are intuitive, given the standard conceptualization of causation in which cause must precede effect in time, and time flows monotonically. (Quasi-circularities, such as positive feedback loops, can be portrayed in DAGs by incorporating time into the causes and effects in the graph; e.g., a weakened immune system at time 1 causes disease at time 2, which causes a weakened immune system at time 3.) The mathematics for using DAGs to determine which variables must be controlled to obtain unconfounded effect estimates (and which must *not* be controlled) produces relatively easy-to-apply rules [11, 12]. DAGs allow the identification of confounders, some of which fit the intuitive definition of a factor that causes confounding (e.g., the confounder in Figure 2, which causes both the exposure and the outcome, thereby causing the population with the exposure to have an increased outcome frequency for reasons other than the effect of the exposure), and some of which do not, but should be controlled statistically because the actual causal confounders are not known or perfectly measured. Thus, the graphs can serve as aids for constructing quantitative models for the estimation of causal effects [13].

Such models have the potential to substantially improve the epistemology of epidemiology and other observational social sciences, though they are generally overlooked despite their obvious usefulness. For example, researchers often conduct statistical analyses by constructing regression models that include every measured variable that might be related to the agent or effect of interest (perhaps dropping those that fail to meet a statistical significance test that actually tells us nothing about whether they actually reduce confounding). Having included in the regression model all available variables from the dataset, researchers often claim that they have controlled for confounding and readers naively believe that confounding is eliminated. However, thinking through the construction of a causal pathway diagram, incorporating background knowledge, often reveals additional hypothesized confounders that were not included in the model. Analyzing hypothesized patterns of causation using a DAG often reveals that adjusting for some of the variables included in the model, such as intervening factors, will make the effect estimate less accurate rather than more so,

something that statistical tests of associations within a dataset cannot reveal [14].

Embellishments of causal pathway diagrams beyond DAGs provide further insights. Predicting the effect of an agent through a complicated pathway can be accomplished by adding indicators of the strength of posited causal relationships along the path. Effects of an agent through multiple pathways can be illustrated to show whether different pathways reinforce the same effect or counteract one another. The different manifestations of effect modifiers and confounders in the diagrams also illustrate how these types of variables are quite different, even though they are often confused when all potential agents are merely thought of as other variables in a dataset.

In general, explicitly modeling causal pathways, based on existing knowledge about the phenomenon being studied that is outside the data being analyzed, and then designing statistical models based on hypothesized pathways [13], rather than on naive statistical significance tests or variable selection algorithms, can improve qualitative assessments of whether an agent is causal and quantitative estimates of its effect.

Inferring Causation from Observed Association

Ultimately, no matter how we define, logically analyze, and model causation, scientific analysis is concerned with inferring it about the physical world, in the context of general agreement that we cannot directly observe it. Strictly speaking, under the counterfactual definition (and under alternative inductive definitions), we can never know if an inference of causation is correct (especially in a particular case), and thus never know for sure whether the inference is right. However, at a practical level, modern science is built on the notion of “theories”, which are rules about how the world works that generally include causal relationships corresponding to hypotheses that have stood the test of time, having been supported consistently by empirical evidence and fitting with other widely accepted theories. The terminology of causal inference implicitly recognizes that such worldly conclusions can never take the form of proof or laws, as exist in axiomatic systems.

Worldly causal inference can be based on hypothesizing a causal relationship, usually motivated

by observations, and then assessing whether its implications are empirically borne out. This is typically coupled with trying to rule out explanations for the observation that do involve the hypothesized causal relationship. (Perhaps a more common practice is merely observing an empirical association in a dataset and declaring it to be causal, but this is an example of the naive simplistic views of causation mentioned above.) Although attempts to systematize the empirical process for inferring causation have been proposed, defining a method for inferring causation is quite similar to defining “the” scientific method, a construct that modern philosophy of science generally dismisses as nonexistent. Just as the scientific method is often accepted to be, roughly, as whatever it is that good scientists have found to be useful to achieve their goals, good practice seems to better define proper causal inference than does any attempt to provide a recipe for it.

However, though there is no accepted recipe, there are widely accepted ingredients, which are based on hypothetico-deductive reasoning [15], methods for ruling out alternative explanations for an observed association, and other analysis, including

- replication of results to show that an association is systematic rather than an accident of one dataset (this should include observation of the association under different circumstances where the hypothesized causal relationship should still hold);
- observation of changes in the association across populations with different distributions of hypothesized effect modifiers;
- observation that the association is absent for populations that lack the hypothesized causal complements;
- observation that the association disappears when a hypothesized intervening factor in the causal pathway is held fixed, and similar observations of predicted effects of changing other variables;
- observation of the predicted effects of controlling for confounders and other covariates of the agent and outcome, possibly including randomized trials;
- analysis of confounding and measurement errors, based on studying hypothesized sufficient sets or causal pathways that do not include the agent in question, to rule them out as explanations for apparent effects of the agent; and,

- perhaps most important, assessment of how well the populations that are being compared, with and without the agent, approximate the true counterfactual contrast.

This list reflects how implications of a given hypothesis are sometimes not completely obvious, and thus the above models and formalizations are often quite useful for directing scientific inquiry. The empirical predictions of a hypothesis are not clear without envisioning causal pathways, and the causal implications of previous observations are similarly obscure.

One class of tools that are sometimes presented as being useful for inferring causation, but which are not supported by either axiomatic logic or empirical evidence of usefulness, are commonly called *causal criteria* [16]. These are lists of considerations, such as strength of association, consistency across studies, and scientific plausibility or coherence, that supposedly help distinguish between causal and noncausal associations [17]. These tools are popular in health sciences, though not in similar social sciences [18]. Such lists consist primarily of commonsense considerations that should be part of the *ad hoc* process of scientific thinking that is sketched above. As such, they are potentially useful, given the ironic rarity of common sense.

It is generally accepted that no set of criteria could provide indisputable evidence of causation; it is easy to construct examples that show that any proposed condition can be met in the absence of causation [10]. Furthermore, as has been pointed out elsewhere [18, 19], these lists do not actually constitute criteria (i.e., there is no metric for determining if a particular condition has been met), let alone provide a demonstrated method for drawing more accurate conclusions. Though the commonsense advice seems likely to improve rather than hurt inference, there is no systematic evidence that inferences informed by such lists are more likely to stand the test of time or otherwise prove to be better than those that are not. When causal “criteria” are invoked by authors for the assessment of evidence, it is done inconsistently [20], and it appears authors often “deploy” [21] causal criteria to rhetorically bolster their conclusions about the presence or absence of causation, rather than employ them to aid inference.

The individual whose name is most closely associated with causal criteria is Austin Bradford Hill

for his list of causal considerations [22], often called the *Bradford Hill Criteria*, despite Hill’s own statement that there are no “‘hard-and-fast rules of evidence’ for causation” [p. 299]. Hill clearly described his list as “considerations”, not criteria, and expressed caveats about their limits. Hill’s address that presented his considerations actually contains a useful mix of epistemic approaches to inferring causation, and using that information to inform decisions, when muddling through a complicated and highly incomplete science [18]. He alluded to most of the “ingredients” that appear in the above list of methods for inferring causation, some of which are captured in lists of “criteria” and some of which are not, but did not imply that he (or anyone else) could offer a complete recipe for inferring causation in complex social sciences.

It is worth noting that because inferred causation cannot be definitive, all decisions that are based on such inference are based on uncertainty, though they still must be made. Hill [22, p. 300] noted that conclusions of causation are often quite uncertain but this “does not confer upon us a freedom to ignore the knowledge that we already have, or to postpone the action it appears to demand at a given time”. Optimal practical application of causal conclusions requires assessments of how certain the conclusion is, along with assessment of the costs and benefits of actions given the accuracy of the conclusion (using methods from economics known as *policy analysis*); efforts to infer causation while ignoring policy analysis are of little value to the goal, in Hume’s words, of regulating future events by their causes. Careful consideration of what causation is and the limits of how well we can know it provides tools for formal analysis and suggests methods for inference, but also immunity to the paralysis that can be caused by the pursuit of certainty.

References

- [1] Hume, D. (1993). *An Enquiry Concerning Human Understanding*, Harvard Classics Volume 37, Copyright 1910 P.F. Collier & Son, Released to Public Domain August 1993, at <http://18th.eserver.org/hume-enquiry.html#7.1>.
- [2] Dawid, A.P. (2007). Counterfactuals, hypotheticals and potential responses: a philosophical examination of statistical causality, in *Causality and Probability in the Sciences*, *Texts in Philosophy Series*, F. Russo &

- J. Williamson, eds, College Publications, London, Vol. 5, pp. 503–532.
- [3] Miettinen, O. (1982). Causal and preventive interdependence. Elementary principles, *Scandinavian Journal of Work, Environment and Health* **8**, 159–168.
- [4] Greenland, S. & Robins, J.M. (1986). Identifiability, exchangeability, and epidemiological confounding, *International Journal of Epidemiology* **15**, 413–419.
- [5] Maldonado, G. & Greenland, S. (2002). Estimating causal effects, *International Journal of Epidemiology* **31**, 421–421.
- [6] Evans, A.S. (1978). Causation and disease: a chronological journey, *American Journal of Epidemiology* **108**, 249–258.
- [7] MacMahon, B., Pugh, T.F. & Ipsen, J. (1960). *Epidemiologic Methods*, Little, Brown, Boston.
- [8] Rothman, K.J. (1976). Causes, *American Journal of Epidemiology* **104**, 587–592.
- [9] Rothman, K.J. (2002). *Epidemiology; An Introduction*, Oxford University Press, Oxford.
- [10] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Edition, Lippencott-Raven Publishers, Philadelphia.
- [11] Pearl, J. (2000). *Causality*, Cambridge, New York.
- [12] Greenland, S., Pearl, J. & Robins, J.M. (2001). Causal diagrams for epidemiologic research, *Epidemiology* **12**, 313–320.
- [13] Greenland, S. & Brumback, B. (2002). An overview of relations among causal modelling methods, *International Journal of Epidemiology* **31**, 1030–1037.
- [14] Robins, J.M. (2001). Data, design, and background knowledge in etiologic inference, *Epidemiology* **12**, 313–320.
- [15] Popper, K. (1959). *The Logic of Scientific Discovery*, Routledge, New York.
- [16] Weed, D.L. (1995). Causal and preventive inference, in *Cancer Prevention and Control*, P. Greenwald, B.S. Kramer & D. Weed, eds, Marcel Dekker, pp. 285–302.
- [17] Susser, M. (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology, *American Journal of Epidemiology* **133**, 635–648.
- [18] Phillips, C.V. & Goodman, K.J. (2006). Causal criteria and counterfactuals; nothing more (or less) than scientific common sense, *Emerging Themes in Epidemiology* **3**, 5.
- [19] Phillips, C.V. & Goodman, K.J. (2004). The missed lessons of Sir Austin Bradford Hill, *Epidemiologic Perspectives and Innovations* **1**, 3.
- [20] Weed, D.L. & Gorelic, L.S. (1996). The practice of causal inference in cancer epidemiology, *Cancer Epidemiology, Biomarkers and Prevention* **5**, 303–311.
- [21] Poole, C. (2001). Causal values, *Epidemiology* **12**(2), 139–141.
- [22] Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**, 295–300.

Related Articles

Association Analysis

Compliance with Treatment Allocation

Epidemiology as Legal Evidence

CARL V. PHILLIPS AND KAREN J. GOODMAN

Censoring

Censoring is commonly used to denote a type of sampling scheme or to describe a type of incomplete data. By censored data, we mean that a random variable of interest is only partly known, instead of being observed exactly. In many applications of risk assessment, the random variable is the time to some event such as the occurrence of infection, tumor, or a particular risk event under study. A typical example of censored data occurs in tumorigenicity experiments in which occurrence or prevalence rates of some tumors are of interest [1–3]. In this situation, the occurrence of tumor in study animals often cannot be directly observed while they are still alive, and the only way to detect these tumors is to euthanize (sacrifice) the animals or wait for their natural death. As a consequence, we know only whether a tumor is present or absent at the time of death or sacrifice, we do not know the exact time that the tumor arose. Censoring is also common in medical studies in which the response variable is time to a certain event or failure time [4].

It is worth noting that another type of sampling scheme or incomplete data that is related to censoring is truncation, which is described in a different article (see **Truncation**) and means that a random variable is observed only if it is within a certain window. To give an idea of the two types of incomplete data and to compare them, consider a simple example: let T be the age of an animal at which a lung tumor occurs. Suppose that the animal is under the study from birth and the study stops at the age C (e.g., $C = 2$) of the animal at which the animal is still lung tumor free. Then the only information available about T is $T \geq C$ and we say that T is right censored. That is, we have censoring. On the other hand, suppose that the study starts when the animal is at age U (e.g., $U = 1.5$) and the study includes only those subjects that are lung tumor free at the beginning of the study. Then the animal is included in the study only if $T \geq U$ and we say that T is truncated at U . That is, we have truncation. Note that with truncation, the variable T could be exactly observed (e.g., $T = 2.5$), while in the presence of censoring, the exact value of T (e.g., $T = 2.5$) is not known. It is apparent that censoring and truncation can occur individually or together.

In general, there are several types of censoring, including right censoring, left censoring, interval censoring, and double censoring. In the following sections, we describe each of these in detail and briefly discuss commonly used inference models and procedures.

Right Censoring

Right censoring is perhaps the most common censoring in practice and the one that has been studied the most in the literature [4]. Consider a study of some response variables that represent times to certain events. By right-censored data, we usually mean that the variable is either observed exactly or known only to be greater than another variable, commonly referred to as the *censoring variable*. This can occur, for example, when study subjects drop out of the study for reasons unrelated to the variable under study or when the study has to stop. In this situation, if the event under study has not occurred yet, then the only information available about the variable is that it is greater than the time at which the subject dropped out or the study was stopped. Right-censored data arising from n independent subjects are commonly represented in the form $\{x_i = \min(T_i, C_i), \delta_i = I(X_i = T_i)\}_1^n$, where the T_i 's denote the response variables of interest, and the C_i 's the censoring variables. Thus, the likelihood function can be written as

$$L = \prod_{i=1}^n [f(x_i)]^{\delta_i} [1 - F(x_i)]^{1-\delta_i} \quad (1)$$

assuming that T_i and C_i are independent, where f and F denote the pdf (probability density function) and cdf (cumulative density function) of the T_i 's, respectively.

There exists a great deal of literature on statistical analysis of right-censored data; some excellent books on this topic include Cox and Oakes [2], Kalbfleisch and Prentice [1], Klein and Moeschberger [4], and Lawless [5]. For inference about right-censored data, typically, three types of problems have been extensively investigated: nonparametric estimation of a distribution or survival function, nonparametric comparison of several distributions or survival functions, and regression analysis. For the nonparametric estimation, the most famous procedure is the Kaplan–Meier estimate initially proposed in Kaplan and Meier [6]. It has a closed form expression, and its properties have been well studied.

For nonparametric comparison of survival functions, a number of procedures have been developed, and the most commonly used one is perhaps the log-rank test [4]. The log-rank test statistic is the summation of the observed numbers of the events minus the expected numbers of the events. In terms of regression analysis of right-censored data, the first model that one learns is the proportional hazards or Cox model proposed in Cox [7]. It specifies that the hazard function of T_i at time t , the conditional density function of T_i given that T_i is greater than or equal to t , has the form $\lambda(t|Z_i) = \lambda_0(t) \exp(Z_i'\beta)$, where $\lambda_0(t)$ denotes the baseline hazard function, Z_i is the covariate vector associated with subject i , and β denote regression parameters. For inference about this model, Cox [8] proposed a partial-likelihood approach that has been proved to be very effective and useful, and generalized to many complex situations [4].

Left Censoring

Left censoring can be seen as the opposite of right censoring in terms of the structure. A left-censored observation usually means that the time to the event of interest is known only to be smaller than a specified value [6]. A common reason for left censoring is that the event of interest has occurred before the study began or before the subject entered the study. Another common reason is the inability to measure the variable of interest when it is below a certain level, for example, when the value is below than the sensitivity of the measuring device. An adverse event related to a drug sometimes can be determined only when the severity of the event exceeds a certain degree. Little research has specifically addressed the left censoring because it does not occur as often as right censoring, and it can be treated similar to right censoring or as a special case of interval censoring.

Interval Censoring

Interval censoring is a censoring mechanism that is more general than either the left- or the right-censoring mechanisms, and it has been attracting more and more attention [9]. Interval-censored data usually refer to the incomplete data in which the time to the event of interest is observed or known only to fall within an interval or some window of time. In

other words, one knows only the range within which the event occurred, but not the exact occurrence time. A typical example of interval-censored data is given in a follow-up study in which each study subject is observed or monitored periodically, and their observation times may differ from subject to subject. In this case, the variable of interest is known only to have some value within the interval given by the last observation time at which the event had not yet occurred and the first observation time at which the event has occurred. If the interval contains either a single point or infinity, then the interval-censored data reduce to right-censored data. On the other hand, if the interval contains either a single point or zero, the data reduce to left-censored data.

Another important special case of interval-censored data is the so-called current status data [10]. In this situation, each subject is observed only once to obtain the status of the occurrence of the event of interest at the observation time. In other words, the observation of the time to the event is either left or right censored. Many tumorigenicity experiments produce this type of data for the tumor onset time [11]. Cross-sectional data provide another example of current status data [12]. Note that for the first case, current status data result from the inability to measure the variable directly and exactly, and for the second case, they are observed because of the study design. Current status data are sometimes referred to as *case I interval-censored data*, and the general case is referred to as *case II interval-censored data* [9].

As before, let T denote the response variable of interest or the time to the event of interest, and let $F = \Pr(T \leq t|X)$ be its cdf, given a vector of covariates X . Suppose that observed data can be represented by $\{I_i, X_i\}_{i=1}^n$, where $I_i = (L_i, R_i]$ is the interval known to contain the unobserved time associated with the i th subject, and X_i are the covariates associated with subject i . Let $\{s_j\}_{j=0}^{m+1}$ denote the unique ordered elements of $\{0, \{L_i\}_{i=1}^n, \{R_i\}_{i=1}^n, \infty\}$, and α_{ij} be the indicator of the event $(s_{j-1}, s_j] \subseteq I_i$. Then the full-likelihood function is proportional to

$$\begin{aligned} L &= \prod_{i=1}^n \{F(R_i|X_i) - F(L_i|X_i)\} \\ &= \prod_{i=1}^n \left[\sum_{j=1}^{m+1} \alpha_{ij} \{F(s_j|X_i) - F(s_{j-1}|X_i)\} \right] \quad (2) \end{aligned}$$

assuming that the mechanism generating I_i is independent of the response variable.

For the one-sample problem associated with interval-censored data, a few algorithms have been developed to obtain the nonparametric maximum-likelihood estimate of a distribution function. The simplest one is the self-consistency algorithm proposed by Turnbull [11], which can be seen as an application of the EM (expectation-maximization) algorithm. This approach is easy to implement, but is known to have a slow convergence rate. The one-step EM algorithm proposed by Rai and Matthews [13] is not only self-consistent and faster, but also avoids boundary problems. Some alternatives that converge faster than the self-consistency algorithm include the convex minorant algorithm introduced in Groeneboom and Wellner [12] and a hybrid algorithm developed by Wellner and Zhan [14]. All the above algorithms are iterative and, in fact, there is no closed form for the estimate.

For regression analysis of interval-censored data, a number of methods have been developed independently or as generalizations of the approaches for right-censored data. Among regression models used to describe interval-censored data, as with right-censored data, the proportional hazards model is perhaps the one most commonly used [14]. Other models that have been investigated include the accelerated regression model [15], the additive hazards model [16], the proportional odds model [17], the logistic model [18] (see **Logistic Regression**), and the linear transformation model [19]. See Sun [8] for inference about these models and statistical approaches for other related problems such as nonparametric comparison of treatments based on interval-censored data.

Double Censoring

To this point, we have defined the response variable as the duration from time zero to an event. A more general situation that occurs in practice is that the response variable represents the time between two related events. Thus, double censoring occurs when the observations of the occurrences of both events suffer censoring discussed above [20]. In this case, the analysis is more complicated because the likelihood function involves not only the distribution of the response variable of interest but also the distribution of the originating event that defines the response variable. Following De Gruttola and Lagakos [21], who

first investigated this type of censoring and proposed a self-consistency algorithm for estimation of the distribution function, many authors have considered the inference about doubly censored data. See Sun [22] for a review of the related literature.

Other Censorings

The discussion so far has focused on univariate censoring or censoring of one response variable. Sometimes, a study may involve several response variables and all of them suffer censoring, either of the same type or of different types [22]. Also, in the above discussion, it was assumed that censoring is independent of the random mechanism generating the response variable of interest. In some situations, this may not be true, i.e., the censoring variable and the response variable may be related [23]. Another type of censoring that was not discussed above occurs when a response variable of interest is observed exactly if it belongs to an interval and is left or right censored if it is to the left or right of the interval. This type of censoring is also sometimes referred to as *double censoring* in the literature [23].

Acknowledgments

This research was supported in part by the Grant CA21765 from the National Institute of Health and by the American Lebanese Syrian Associated Charities (ALSAC). The authors thank Dr. Angela McArthur for critically scientific editing. The authors are also grateful to the editor and a referee for providing many helpful suggestions that significantly improved the manuscript.

References

- [1] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Edition, John Wiley & Sons, New York.
- [2] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*, Chapman & Hall, London.
- [3] Rai, S.N. & Matthews, D.E. (1995). Analysis of incomplete data using stochastic covariates. *The Canadian Journal of Statistics* **23**, 29–42.
- [4] Klein, J.P. & Moeschberger, M.L. (2003). *Survival Analysis*, Springer-Verlag, New York.
- [5] Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.
- [6] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

- [7] Cox, D.R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society, Series B* **34**, 187–220.
- [8] Sun, J. & Kalbfleisch, J.D. (1993). The analysis of current status data on point processes, *Journal of the American Statistical Association* **88**, 1449–1454.
- [9] Rai, S.N., Sun, J. & Hunt, D. (2002). Analysis of occult tumour trial data with varying lethality, in *Recent Advances in Statistical Methods*, Y.P. Chaubey, ed, World Scientific Publishing, London, pp. 233–251.
- [10] Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective (with discussion), *Journal of the Royal Statistical Society, Series A* **154**, 371–412.
- [11] Turnbull, B.W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data, *Journal of the Royal Statistical Society, Series B* **38**, 290–295.
- [12] Groeneboom, P. & Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*, DMV Seminar, Band 19, Birkhauser, New York.
- [13] Rai, S.N. & Matthews, D.E. (1993). Improving the EM algorithm, *Biometrics* **49**, 587–591.
- [14] Wellner, J.A. & Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data, *Journal of the American Statistical Association* **92**, 945–959.
- [15] Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data, *Biometrics* **42**, 845–854.
- [16] Betensky, R.A., Rabinowitz, D. & Tsiatis, A.A. (2001). Computationally simple accelerated failure time regression for interval censored data, *Biometrika* **88**, 703–711.
- [17] Lin, D.Y., Oakes, D. & Ying, Z. (1998). Additive hazards regression with current status data, *Biometrika* **85**, 289–298.
- [18] Huang, J. & Rossini, A.J. (1997). Sieve estimation for the proportional odds failure-time regression model with interval censoring, *Journal of the American Statistical Association* **92**, 960–967.
- [19] Sun, J. (1997). Regression analysis of interval-censored failure time data, *Statistics in Medicine* **16**, 497–504.
- [20] Zhang, Z., Sun, L., Zhao, X. & Sun, J. (2005). Regression analysis of interval-censored failure time data with linear transformation models, *The Canadian Journal of Statistics* **33**(1), 61–70.
- [21] De Gruttola, V.G. & Lagakos, S.W. (1989). Analysis of doubly-censored survival data, with application to AIDS, *Biometrics* **45**, 1–11.
- [22] Sun, J. (2002). Statistical analysis of doubly interval-censored failure time data, in *Handbook of Statistics: Survival Analysis*, N. Balakrishnan & C.R. Rao, eds, Elsevier, North Holland, pp. 105–122.
- [23] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer-Verlag, New York.

Further Reading

- Cai, T. & Cheng, S. (2004). Semiparametric regression analysis for doubly censored data, *Biometrika* **91**, 277–290.
- Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer-Verlag, New York.
- Zhang, Z., Sun, L., Sun, J. & Finkelstein, D.M. (2007). Regression analysis of failure time data with informative interval censoring, *Statistics in Medicine* **26**, 2533–2546.

Related Articles

Hazard and Hazard Ratio Relative Risk

SHESH N. RAI AND JIANGUO SUN

Change Point Analysis

Change point analysis refers to the statistical method of characterizing changes in the underlying nature of a response variable over a specified range of values of a predictor variable. As an example, one of the most oft-cited change point examples from the literature consists of a historical data set containing the length of time intervals (in days) between coal mining disasters with more than 10 men killed during the period 1851–1962 in Great Britain. The data set was first introduced by Maguire *et al.* [1] and subsequently corrected and updated by Jarrett [2]. A plot of the length in days for each successive interval is shown in Figure 1.

For this example, the response variable is the length of the intervals between accidents and the predictor variable is time. One might ask if there is evidence that the average interval length has changed significantly over time. Considering Figure 1, there appears to be periods between the 50th and 60th intervals, the 70th and 80th intervals, and the 100th and 109th intervals where the average time between such accidents greatly increases. Over these time periods, the occurrence of annual *risk* of an accident appears to decrease significantly. A change point analysis, for this example, might consist of estimating the exact times such changes in the mean level occurred as well as developing estimates for the mean level before and after each estimated change point. Of course, the end goal of the analysis might be to

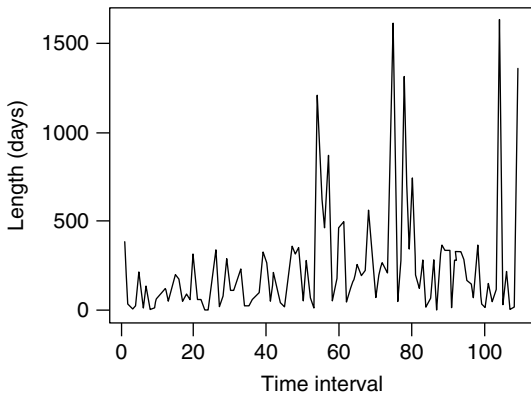


Figure 1 A plot of the interval length in days between successive coal mining disasters

relate the estimated times where the changes occurred to real-world events. Examples of statistical methods used to perform a change point analysis for this coal mining data may be found in West and Ogden [3] and Raftery and Akman [4]. Since the coal mining data is collected over time, there is also a suite of statistical methods in the area of *intervention analysis* designed to model this type of changing behavior.

The above example illustrates potential abrupt changes in the mean interval length over time. These abrupt changes in the response are frequently referred to as *discontinuous* change points. This implies that some underlying functional governing the response is a discontinuous function of the predictor variable. The simplest mathematical model with a single discontinuous change point in the mean level is given by

$$E(Y(x)) = \begin{cases} \mu & \text{for } x < \theta \\ \mu + \beta & \text{for } x \geq \theta \end{cases} \quad (1)$$

where the response variable $Y(x)$ has an expected value of μ for values of the predictor variable, x , less than the change point θ and a mean level $\mu + \beta$ for values of x greater than or equal to θ . In this case, β represents the *size* of the discontinuity at the change point. The values μ , β , and θ must be estimated to fully specify the model. In addition to point estimates for these quantities, confidence intervals that reflect the uncertainty in the estimates should also be computed. If a confidence interval for β contains the value zero, then the entire change point model may be statistically insignificant, meaning that there is no strong evidence of a change in the mean response. A model such as the one above might be appropriate when there is an introduction of a new factor influencing the response variable, which has an immediate sustained impact on the response. In the context of the coal mining data, such a factor might be the implementation of new safety procedures by the coal mining industry.

In many settings, the above model may not be appropriate as abrupt changes in the response are not expected. This is most often the case in biological settings where, for example, an organism may be expected to withstand exposure to a certain amount of exposure to a toxic substance before it begins to gradually deteriorate. Perhaps the simplest mathematical model for this type of behavior is called

the *onset-of-trend model*, which is given by

$$E(Y(x)) = \begin{cases} \mu & \text{for } x < \theta \\ \mu + \beta(x - \theta) & \text{for } x \geq \theta \end{cases} \quad (2)$$

where the expected value of the response $Y(x)$ is constant at μ for x less than θ but then either increases or decreases from this value depending on the sign of β as x increases beyond θ . The continuous nature of the above mean response allows for a gradual change in the mean rather than an abrupt change. West and Kodell [5] fit the onset-of-trend model to data on the body weight gain (or loss) in grams for Fischer 344 rats treated with aconiazide over a 14-day period in dose groups of 0, 100, 200, 500, and 750 mg kg⁻¹ body weight. Figure 2 shows the raw data for this example along with the fitted change point model where the estimated change in mean weight gain is estimated to occur at a dose of approximately 350 mg kg⁻¹ body weight. The 95% lower confidence limit on the change point was computed to be 242.69 mg kg⁻¹ body weight. If this lower limit had been below zero (the lower limit of the observed dose predictor), the change point model might be statistically insignificant.

The above models are examples of models for changes in mean response over the predictor variable. Change point analysis can be conducted for other

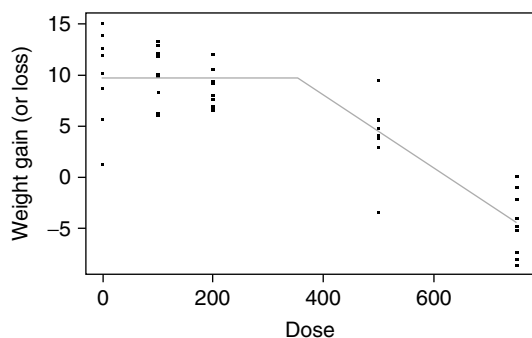


Figure 2 The onset-of-trend model for the aconiazide data

underlying characteristics of the response variable as well. For example, Hsu [6] and Inclan and Tiao [7] develop statistical methods to identify shifts in the variance of a response variable rather than for the mean response. In addition to models for other response characteristics, one may also develop much more complicated functional models relating the response characteristic to the predictor variable. This underlying functional relationship may be estimated by either specifying a complicated parametric model like a nonlinear function or by using nonparametric methods as described by Loader [8]. The fitting and analysis of these more complicated change point models typically requires the help of an experienced statistician.

References

- [1] Maguire, B.A., Pearson, E.S. & Wynn, A.H.A. (1952). The time intervals between industrial accidents, *Biometrika* **38**, 168–180.
- [2] Jarrett, R.G. (1979). A note on the intervals between coal-mining disasters, *Biometrika* **66**, 191–193.
- [3] West, R.W. & Ogden, R.T. (1997). Continuous-time estimation of a change-point in a poisson process, *Journal of Statistical Computation and Simulation* **56**, 293–302.
- [4] Raftery, A.E. & Akman, V.E. (1986). Bayesian analysis of a poisson process with a change-point, *Biometrika* **73**, 85–89.
- [5] West, R.W. & Kodell, R.L. (2005). Changepoint alternatives to the NOAEL, *Journal of Agricultural Biological and Environmental Statistics* **10**, 197–211.
- [6] Hsu, D.A. (1977). Tests for variance shift at an unknown time point, *Applied Statistics* **26**, 279–284.
- [7] Inclan, C. & Tiao, G.C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance, *Journal of the American Statistical Association* **89**, 913–923.
- [8] Loader, C.R. (1996). Change point estimation using nonparametric regression, *The Annals of Statistics* **24**, 1667–1678.

RONNIE WEBSTER WEST

Chernobyl Nuclear Disaster

The Incident

On April 26, 1986, Unit 4 of the Chernobyl nuclear power plant, located 100 km from Kiev in Ukraine, exploded. The explosion was caused by a combination of flaws in the reactor's design and a series of human errors during a test of the plant's ability to provide enough electrical power to operate the reactor core cooling system between a loss of main station power and the start up of emergency power. After the explosion, graphite and other materials caught fire. The Chernobyl nuclear disaster caused the death of 30 power plant workers, evacuation of thousands of people from their homes, and contamination of people, animals, plants, and soil over a wide area.

The fire burned for 10 days and released radioactive gases (see **Radioactive Chemicals/Radioactive Compounds**), condensed aerosols, and a large amount of fuel particles primarily in the former Soviet Union, but this was also witnessed throughout the Northern Hemisphere. Total ionizing radiation^a released has been estimated at 14 EBq (or 1.4×10^{18} Bq, where 1 Bq equals 1 atomic disintegration per second), including the release of 1.8 EBq of iodine-131, 0.085 EBq of cesium-137, 0.01 EBq of strontium-90, and 0.003 EBq of plutonium radionuclides [1].

The duration and high altitude reached by the Chernobyl radioactive plume resulted in approximately 200 000 km² of Europe receiving levels of cesium-137 in excess of 37 kBq m⁻². About 70% of that area was in the most affected countries of Belarus, the Russian Federation, and Ukraine (Figure 1). Deposition of radioactivity throughout these countries was variable, but was increased in areas where it was raining when the radioactive plume passed. Heavier fuel particles were deposited within 100 km of the power plant, but lighter, more volatile fission products, such as cesium and iodine, were able to travel greater distances [2].

Dose

Dose is a complex concept in radiation biology (Figure 2). Measurement of biological dose starts out

with the *equivalent* dose, which takes into consideration the biologic potency of different types of radionuclides. The *effective* dose is then obtained by multiplying the *equivalent* dose to various tissues and organs by a weighing factor appropriate to each and then summing up the products. The Sievert (Sv) is the quantitative risk that a person will suffer adverse health effects from a particular exposure to radiation.

On the basis of the availability of quantitative estimations of radiation doses, five exposed populations can be identified: (a) workers who responded first to the explosion and fire ("emergency workers"); (b) workers involved in the cleanup and recovery phase from 1986 through 1990 ("recovery operation workers" or "liquidators"); (c) inhabitants of contaminated areas in Belarus, the Russian Federation, and Ukraine ("evacuees"); (d) inhabitants of contaminated areas not evacuated; and (e) inhabitants of areas outside Belarus, the Russian Federation, and Ukraine [3].

Exposed Populations

Emergency Workers

The most exposed were emergency responders – Chernobyl plant staff and security guards, firefighters, and local medical aid personnel. During the night of the accident, about 600 workers were on site at the Chernobyl power plant. These workers were mainly exposed to external irradiation (whole-body γ irradiation and β irradiation to skin) from the radioactive cloud, the fragments of the damaged reactor core scattered over the site, and radioactive particles deposited on the skin. During the fire, measured exposure rates near the reactor core were hundreds of gray^b per hour.

Acute health symptoms were noted in 237 highly exposed emergency workers and a diagnosis of acute radiation syndrome (ARS) was confirmed in 134 of them. Since the emergency workers' individual dosimeters (or film badges) were massively overexposed to radiation during the accident, they could not be used to estimate the individual doses of γ irradiation received. Instead, complex biological dosimetry processes were used to estimate external dose in the 134 hospitalized workers.

The results of this dosimetry showed that 41 workers had mild ARS. All of them received whole-body doses from external γ irradiation in the range



Figure 1 Map of area affected by the explosion at the Chernobyl nuclear power plant [Source: <http://www.sheppardsoftware.com/images/Europe/factfile/chernobyl1.jpg>]

of 0.8–2.1 Gy. All 41 survived. Ninety three workers received higher γ doses. Of these 93 workers, 50 workers received doses between 2.2 and 4.1 Gy and only 1 died from ARS; 22 workers received between 4.2 and 6.4 Gy and 7 died; and 21 workers (13 of whom received bone-marrow transplantations) received doses between 6.5 and 16 Gy and 20 died. Skin doses from β irradiation were evaluated in eight treated workers and were found to be in the range of 400–500 Gy [1].

A total of 28 emergency workers died of ARS within the first four months after the accident. For reference, the lethal dose that would kill 50% of a γ radiation exposed population within 30 days is

about 4.5 Gy. For these severely affected workers, more than 20 radionuclides were detectable in whole-body γ measurements, but only radioiodines (iodine-131 and iodine-132) and radiocesiums (cesium-134 and cesium-137) were present in more than negligible amounts [1].

Recovery Workers or “Liquidators”

An estimated 600 000 workers – 360 000 civilians, and 240 000 soldiers – were involved in recovery and cleanup activities from 1986 to 1990. These workers were also known as *liquidators*. Recovery activities mainly involved decontaminating the

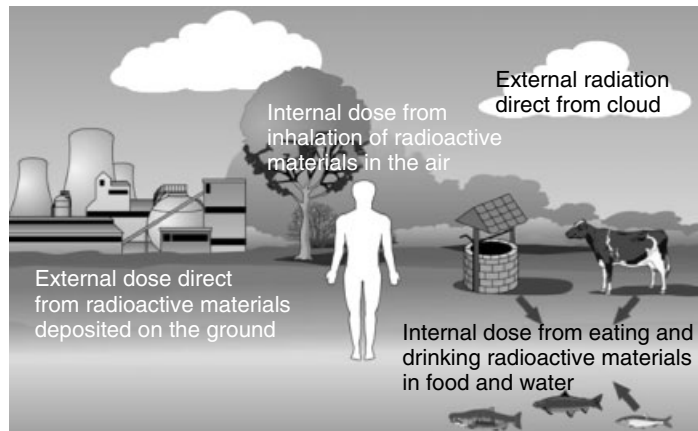


Figure 2 Sources of exposure to radioactive materials [Source: <http://www.greenfacts.org/fr/tchernobyl/images/figure-2-exposure.jpg>]

reactor block, reactor site, and roads surrounding the power plant and constructing a containment structure for the reactor known as the *sarcophagus* or *shelter object*, in addition to a settlement for plant personnel and various radio-waste repositories [4].

Among recovery workers, the most highly exposed include 5000 military personnel who worked on the roof of the reactor shortly after the accident, 1200 helicopter pilots who dumped material on the reactor to contain the fire, and others involved in risky emergency operations during the first weeks after the explosion and fire [5].

Estimates of γ irradiation doses to recovery workers, in general, show that the mean *effective* dose decreased from year to year from about 170 mSv in 1986, 130 mSv in 1987, 30 mSv in 1988, to 15 mSv in 1989 and that the average effective dose from external γ irradiation to recovery operation workers in the years 1986–1989 was about 113 mSv [1]. For reference, national regulations in the Soviet Union established a maximum allowable dose limit of 250 mSv in 1981. From 1989, the annual dose limit for civilian workers was set much lower at 50 mSv.

In addition to penetrating external radiation, recovery workers also received nonpenetrating skin doses from β irradiation and thyroid doses from internal irradiation. Unprotected skin doses were several times greater than γ doses. Thyroid doses during the first weeks after the explosion may also have been substantial.

Evacuees

Beginning within 48 h after the explosion and continuing through late 1986, about 116 000 inhabitants of villages in Belarus, the Russian Federation, and Ukraine were evacuated from the “Chernobyl exclusion zone (CEZ)”, which measures 30 km in all directions from the damaged Chernobyl reactor. Releases from radionuclides during the fire accounted for most of the exposure of the general public of which iodine-131 (half-life 8 days) and cesium-137 (half-life 30 years) were the most significant [1].

The average effective dose from external irradiation for 30 000 evacuees from Pripjat, a town only 3 km from the reactor, which was evacuated within 48 h, was estimated to be 17 mSv with a range of 0.1–380 mSv. Thyroid doses received by evacuees varied with their age (i.e., the very young were at greater risk), place of residence (i.e., the closer you lived to the reactor, the greater risk your risk), and the date of evacuation (i.e., the earlier the evacuation, the less was the risk). For Pripjat, the population-weighted average thyroid dose is estimated to have been 0.17 Gy with a range of 0.07 Gy for adults to 2 Gy for infants aged 3 years or less [1].

Estimates are that evacuation averted a collective external exposure of approximately 8260 man Sv in the 116 000 inhabitants evacuated [1]. A “man Sievert” refers to a quantity obtained by multiplying the average *effective* dose by the number of people exposed to a given source of ionizing radiation.

Inhabitants of Contaminated Areas in Belarus, the Russian Federation, and Ukraine not Evacuated

Areas of the countries of Belarus, the Russian Federation, and Ukraine are considered “contaminated” on the basis of measured cesium-137 levels. The known background level of cesium-137 from atmospheric weapons testing by the former Soviet Union in the cold war era was corrected for radioactivity decay that would have occurred in the ensuing years and the “excess” radioactivity was considered due to Chernobyl. On the basis of this type of retrospective radiologic risk estimation, any area with a radiation level of 37 kBq m^{-2} was specified as an area contaminated by the Chernobyl accident. Using this method, about 3% of the former Soviet Union is considered to have been contaminated by cesium-137 deposition densities greater than 37 kBq m^{-2} . Chiefly, these areas have been designated (a) *Central* (within 100 km of the reactor to the west and northwest), (b) *Gomel–Mogilev–Bryansk* (200 km north–northeast of the reactor), and (c) *Kaluga–Tula–Orel* (500 km northeast of the reactor in the Russian Federation) [6].

The total number of people living in areas in the three countries with cesium-137 deposition 37 kBq m^{-2} or greater in 1986 was about 5 million. During the late 1980s, about 272 000 people continued to live in areas with cesium-137 deposition levels in excess of 555 kBq m^{-2} (i.e., “strict control areas”). By 1995, outmigration, as opposed to death, from these highly contaminated areas had reduced the population to about 193 000 inhabitants [1].

Early on, the most important pathways of exposure to inhabitants of the contaminated areas came from *internal* exposure, through ingestion of cow’s milk and other foodstuffs contaminated with iodine-131, cesium-134, cesium-137, as well as *external* exposure from deposition of short-lived radionuclides (tellurium-132, iodine-131, barium-140, rubidium-103 and cerium-144, and others) and long-lived radionuclides (cesium-134 and cesium-137) [7]. Now, external exposure from cesium-137 deposited on the ground and in contaminated foodstuffs is of greatest importance while that from strontium-90 (half-life of 28 years) is of lesser importance.

The average effective dose from cesium-134 (half-life of 2 years) and cesium-137 received during the first 10 years after the accident by inhabitants of

contaminated areas is estimated to be about 10 mSv. Lifetime effective doses are predicted to be about 40% greater than those doses received during the first 10 years following the accident [1].

Inhabitants Outside Belarus, the Russian Federation, and Ukraine

Radioactive materials of a volatile nature like iodine and cesium released during the accident spread throughout the Northern Hemisphere. Radiation doses to populations outside Belarus, Russia, and Ukraine were relatively low and showed large differences depending on whether rainfall occurred during passage of the radioactive plume from Chernobyl. In fact, the Chernobyl nuclear disaster was first detected in April 1986 in the West by Swedish scientists, who were puzzled as to why their radiation monitors were getting activated.

Reliable radiation dose estimation information for populations other than those of Belarus, the Russian Federation, and Ukraine is relatively scant. Thyroid cancer studies in populations of Croatia, Greece, Hungary, Poland, and Turkey have estimated an average thyroid dose range from 1.5 to 15 mGy [8]. Studies of leukemia in populations from Bulgaria, Finland, Germany, Greece, Hungary, Turkey, Romania, and Sweden have reported an average bone-marrow dose range from about 1 to 4 mGy [8].

Late Health Impacts

Thyroid cancer in children exposed at a young age to radiation from Chernobyl has been widely accepted as causally associated with radiation exposure from Chernobyl [9]. Thyroid cancer cases per million population began rising in Belarus and Ukraine from 1990 onward, and then later in the Russian Federation. In Belarus, childhood thyroid cancer incidence rose by a factor of 100, by 200 in the Gomel region, by 7 in Ukraine, and by 8 in the Russian regions of Bryansk and Kaluga [2, 9]. The total number of thyroid cancers diagnosed in the three areas during 1986–2002 among children who were exposed to Chernobyl radiation and who were between the ages of 0–17 years is 4837 cases [9]. Cancer risk for children older than 10 years at the time of exposure seems to have reached a plateau and since 1995 the incidence rate has been decreasing for those 5–9 years

old at the time of the Chernobyl accident, while increasing for those younger than 5 years in 1986 [1]. There is still uncertainty about the existence of a causal relationship for those exposed as adults, as the adult thyroid gland is much less radiosensitive than in childhood [9].

Other than childhood thyroid cancer, no other health effects were considered to be related to Chernobyl exposure. Recently, though, data on emergency workers from the Russian National Medical-Dosimetry Registry (RNMDR) indicates some increase in morbidity and mortality caused by leukemia and solid organ cancer. Ongoing studies of noncancer risks such as cataracts, cardiovascular disease, immune system health effects, reproductive health effects, and mental and behavioral health effects are being regularly evaluated by the World Health Organization [9]. However, separating the effect of radiation from other potential factors in the causation of any late health effect seen is not a simple task. The lack of reliable individual quantitative radiation dose estimates for many in the exposed populations, the low doses of radiation outside the exposed group of emergency workers, and the existence of many other potential confounding factors makes epidemiologic detection of causal relationships between Chernobyl exposure and a particular health effect difficult to discern from the background levels of these diseases [7]. Outside of those exposed as young children and highly exposed emergency and recovery workers who are generally recognized, based on individual estimates, to be at increased risk of adverse health effects from Chernobyl-generated radiation exposure, other populations worried about adverse health effects “do not need to live in fear about adverse health effects from the Chernobyl nuclear disaster” [1].

Environmental Impacts

Large areas within Belarus, the Russian Federation, and Ukraine are still excluded from agricultural uses; and, in an even larger area, crop production, dairy production, and farm animal production, are subject to strict controls. Measurable contamination is expected to persist for 300 more years based on a factor of 10 half-lives for cesium-137. Exclusion and controls operative during the first 10 years after the accident in agricultural areas outside of Belarus,

the Russian Federation, and Ukraine have generally been lifted. Water for drinking and recreation, as well as aquatic environments for fish, have not been seen as a major contributor to total exposure from the accident despite cesium-contaminated fish being found in lakes in Sweden. Deposition in forests continues to be a problem owing to the efficient filtering characteristics of trees. For example, the trees in a forest near Chernobyl were killed by radiation and had to be destroyed and treated as radioactive waste (so-called red forest) [4].

Residual Risks

The “sarcophagus” constructed in May through November 1986 was aimed at quick environmental containment of the damaged reactor, reduction of radiation levels on the site, and prevention of further radionuclide release. However, the containment structure was built very rapidly in 1986 and has degraded since that time. Roof leaks have caused corrosion of the containment unit’s structural members and a worst-case scenario predicts a criticality incident involving leaked water interacting with the melted nuclear fuel and collapse of the sarcophagus with resuspension of radioactivity into the atmosphere. Plans are under way for a new safer containment structure, together with dismantling the existing “sarcophagus”, removing the highly radioactive fuel and decommissioning the damaged reactor [6]. Other residual risks include groundwater and atmospheric contamination from radioactive wastes stored in repository sites within the CEZ, partly buried in trenches and partly conserved in containers [6].

Most of the biologically significant radioactive isotopes released from the Chernobyl nuclear disaster had short half-lives and they have decayed away. Natural decontamination processes have reached an equilibrium and the decrease in contamination will be mainly owing to radioactive decay. Over the next hundreds to thousands of years, the only radionuclide of importance will be plutonium-239 (half-life of 24 110 years) and americium-241 (half-life of 432 years) [1].

International Collaboration

The Chernobyl nuclear disaster was considered by the former Soviet Union to have been purely an internal

matter. More than 20 years later, new thinking has emerged, as well as openness on the part of Belarus, the Russian Federation, and Ukraine, which has generated robust international collaborations. Chernobyl is now considered to be an international risk issue with several United Nations agencies and other international bodies issuing periodic assessments of the risks from Chernobyl. This new era of international collaboration has led to a productive effort to assess and manage the ongoing risks to human health and the environment from the Chernobyl nuclear disaster.

End Notes

^a. Ionizing radiation – radiation that has enough energy to remove tightly bound electrons from atoms creating an “ion pair” made up of a free electron with a negative charge and the remaining atom with a net positive charge, both of which can cause tissue destruction.

^b. Gray – a gray (Gy) refers to *absorbed* dose of ionizing radiation or the quantity of energy delivered by ionization per unit of tissue mass.

References

- [1] United Nations Scientific Committee on the Effects of Atomic Radiation (2000). *Sources and Effects of Ionizing Radiation*, United Nations Scientific Committee on the Effects of Atomic Radiation, 2000 Report to the General Assembly (UNSCEAR 2000), *Annex J: Exposure and Effects of the Chernobyl Accident*, New York, at http://www.unscear.org/unscear/en/publications/2000_2.html.
- [2] Bard, D., Verger, P. & Hubert, P. (1997). Chernobyl, 10 years after: health consequences, *Epidemiologic Reviews* **19**, 187–204.
- [3] Chernobyl’s Legacy (2006). *Health, Environmental and Socio-Economic Impacts and Recommendations to the Governments of Belarus, the Russian Federation and Ukraine*, The Chernobyl Forum: 2003–2005 (Second revised version), International Atomic Energy Agency, Vienna.
- [4] Nuclear Energy Agency (NEA). (2002). *Chernobyl: Assessment of Radiological and Health Impacts: 2002 Update of Chernobyl: Ten Years On*, Organization for Economic Cooperation and Development.
- [5] Boice, J.D. (1997). Leukemia, chernobyl and epidemiology, *Journal of Radiological Protection* **17**, 129–133.
- [6] International Atomic Energy Agency (IAEA) (2006). *Environmental Consequences of the Chernobyl Accident and their Remediation: Twenty Years of Experience*, Report of the Chernobyl Forum Expert Group ‘Environment’, International Atomic Energy Agency, Vienna.
- [7] United Nations Scientific Committee on the Effects of Atomic Radiation (1988). *Sources, Effects and Risks of Ionizing Radiation*, Report to the General Assembly (UNSCEAR 1988), with annexes, United Nations Sales Publication E.88.IX.7, New York.
- [8] Sali, D., Cardis, E., Sztanyik, L., Auvinen, A., Bairakova, A., Dontas, N., Grosche, B., Kerekes, A., Kusic, Z., Kusoglu, C., Lechpammer, S., Lyra, M., Michaelis, J., Petridou, E., Szybinski, Z., Tominaga, S., Tulbure, R., Turnbull, A. & Valerianova, Z. (1996). Cancer consequences of the Chernobyl accident in Europe outside of the former USSR: a review, *International Journal of Cancer* **67**, 343–352.
- [9] World Health Organization (2006). *Health Effects of the Chernobyl Accident and Special Health Care Programmes*, Report of the UN Chernobyl Forum, B. Bennett, M. Repacholi & Z. Carr, eds, Expert Group “Health”, Geneva.

JOHN HOWARD

Clinical Dose–Response Assessment

Much of the science and practice of medicine boils down to the clinician's question of how to treat the patient that seeks his help. After investigating relevant factors and making diagnosis, the question is often which pharmaceutical agent to prescribe and at which dose. The choice of dose is a matter of striking a balance between the risk for insufficient therapeutic effect and the risk for side effects. One simple example is thrombotic prophylaxis; a drug that inhibits the blood's ability to coagulate will decrease the risk of blood clot formation, but at the same time increase the risk of bleeding. Blood clots and bleeding in the brain may both cause stroke and death. Since different patients differ in severity of disease, vulnerability, and individual preferences, the best dose may not be the same in all patients.

We start by considering the dose decision when there are good models describing the risks. This problem can be straightforward when positive and negative effects can naturally be measured on the same scale (section titled "Risk–Benefit"). However, in most situations different outcomes have to be weighted differently (section titled "Choosing the Relative Weights"). Furthermore, informed dose decisions rely on the *estimation*, based on limited data, of dose–response for the relevant effects (section titled "Estimating Dose–Response Curves"). This is sometimes complicated by complete lack of good clinical data for the drug in focus. It may be needed to rely on other sources of information to assess the risks (section titled "Good Data Do Not Always Exist"). An important question is how clinical experiments can be designed to give the best possible data upon which assessments can be based (section titled "Designing Dose-Finding Trials").

Risk–Benefit

Risk can be defined in different ways. One common definition of risk is the probability of a (specified) negative event. We can, for example, say that the risk of drug X causing an allergic reaction is less than 1 in 10 000. Risk may alternatively refer to both the probability of an event and the consequence of

this event. This latter interpretation is used in this article, where we combine the risk for unintended side effects with the risk for insufficient therapeutic effect. It is convenient to define (net) risk as the expected value of the net consequences for a specific individual or, alternatively, for a population of patients. We are interested in the relation dose–risk and, in particular, we are concerned with finding the best dose, that is, the dose that minimizes the risk.

There are often multiple effects of a drug, and the severity of side effects is often difficult to compare with the benefits. To simplify things, however, assume that the key positive and negative effects of a drug relates to mortality. Assume also that the dose–response relations for both effect and safety are known. Figure 1 shows an example of the relation between dose and deaths risks. In this example, the optimal dose is 9 mg and the resulting minimal possible mortality is 8.5%, which could be compared with the mortality of 10% at dose zero, that is, with placebo treatment. Doses above 27 mg result in higher total risks and are therefore considered more harmful than with no pharmaceutical treatment.

Implicitly, we have assumed that a death due to lack of effect is exactly as serious as a death due to an adverse effect. This is not obviously the case, perhaps for both good and bad reasons. One complication is that the time of death may be different; the positive effect may be a long-term reduction in mortality, while safety problems will occur very rapidly in some patients. A clear example, not related to dose finding, is that surgery may be associated with an almost instantaneous risk, while a successful operation may reduce the disease-related risk for many years. In such cases, the treatment option giving later deaths would be preferred if the total risk of dying from the disease or the side effect is the same for the different treatments. Giving that they are about to die, patients may also prefer to know about that in advance and get some time to prepare. Even if the consequences are identical to the patient, treating physicians and regulators may be more concerned about side effects than lack of treatment effect.

Choosing the Relative Weights

In general, a patient may experience a number of different health outcomes, and it is not sufficient only to consider mortality. Different outcomes have

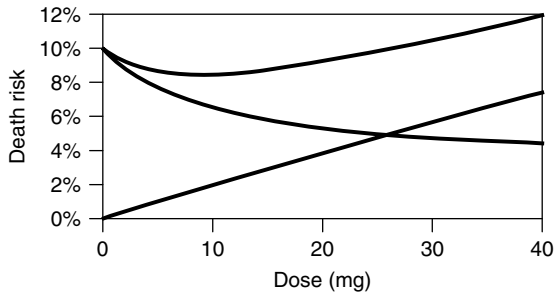


Figure 1 Risk of death due to the disease (decreasing curve), due to adverse side effects (increasing curve), and total risk (U-shaped curve), as functions of the dose

dissimilar severity and should therefore be weighted differently. It may be possible to formulate the overall risk, r , as a weighted combination of the outcome probabilities,

$$r = \sum_{i=1}^m w_i \cdot P(\text{outcome } i) \quad (1)$$

where $w_1, \dots, w_m$ are the relative weights, reflecting the severity of the outcomes.

Consider an example where there are three relevant dichotomous variables, independent of each other. The intended effect of the pharmaceutical drug is to reduce the probability of the first outcome. However, there are two potential side effects, one mild and common, and one rare but serious. Figure 2 shows examples of probabilities for the three different outcomes, as functions of the dose. The relative weights of the outcomes are set to 1 : 0.1 : 5. That is the occurrence of the severe side effect is considered 5 times more serious than the first outcome, which the pharmaceutical drug intends to prevent. This outcome, in its turn, is considered 10 times more serious than the event of a mild side effect. The three risk components, $w_k \cdot P(\text{outcome } k)$, are displayed in Figure 3, together with the total risk. The dose minimizing this risk is considered optimal.

As mentioned in the section titled “Risk–Benefit”, it is complicated to compare deaths and deaths. What then about comparing a decreased risk of stroke with a common occurrence of headache and a minimal probability of a life-threatening allergic reaction? The choice of weights is difficult and often very subjective. However, it may be argued that this approach, explicitly weighting different risk

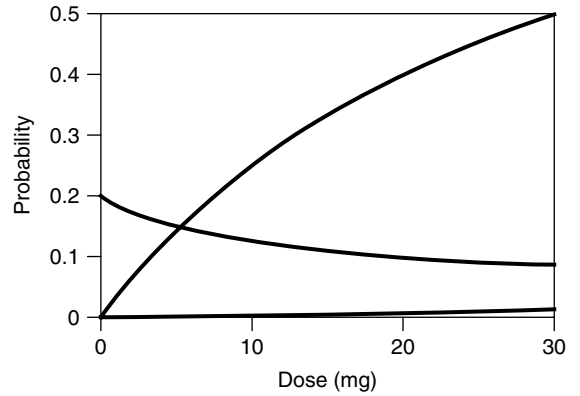


Figure 2 Probabilities of three different outcomes, as functions of dose. Intended risk reduction (decreasing curve), mild side effect (increasing high-probability curve), and severe side effect (increasing low-probability curve)

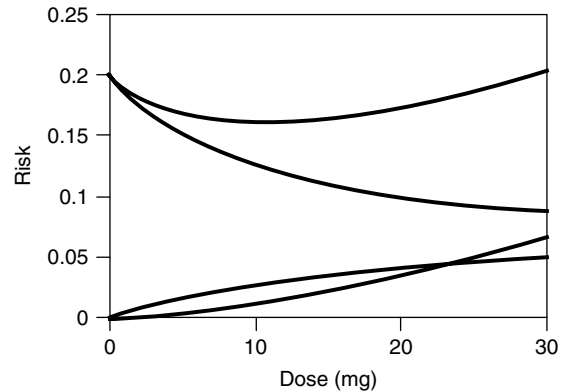


Figure 3 Risks components for three different outcomes as well as total risk (U-shape curve), as functions of dose

components, is more rational than its alternatives and also that it facilitates clear communication. A choice of dose is often inevitable, and this will implicitly mean a weighting of the risk components.

The choice of weights is discussed in general decision analysis literature [1, 2] and in a health-care setting [3, 4]. In the health-care sector, so-called quality-adjusted life years (QALYs) are sometimes used to weight together different health outcomes, especially in health economic evaluations. This concept can be very useful and facilitate a rational allocation of limited health-care resources. However, the formulation quality-adjusted life years may be perceived as offensive by laypersons. We can also see

reasons to be critical of how the concept is sometimes used. For example, the value of a life year with a certain disease is often estimated by asking the healthy public. Some health states may even be given negative values! It is occasionally suggested that individual patient preferences should formally be incorporated in a decision analysis; Protheroe *et al.* [5] gives an example from thrombotic prophylaxis. This approach has clear limitations [4]. Small probabilities are difficult to understand and evaluate for the patient, and the mere mentioning of a very rare but serious side effect may cause unnecessary worries.

Similar entities have different names in different contexts. The terms *utility*, from economics [6] and decision theory [7], or “(clinical) utility index” are often used for a weighted score of health outcomes [8]. QALYs and other preference measurement scales are often used in health economics [9]. In statistical theory the term *loss* is used [10]. The basic ideas are the same, and what we call risk can be interpreted as expected loss or the expectation of negative utility. The expected utility theory underlying much of the reasoning is described in an axiomatic way by French and Ríos Insua [11] and in a medical decision-making context by Parmigiani [4]. We have assumed the risk function to be linear, and the relevant outcomes to be independent and dichotomous. In general, the health value for a patient may be a nonlinear function of variables of different types. The risk is then the expectation of this value.

In this article we generally consider a total risk, reflecting both beneficial and adverse effects. As it is not straightforward to determine the relative importance of different outcomes, however, it is common that effect and safety are analyzed separately. Different requirements are often imposed on these aspects, and the therapeutic interval is defined to be the range of doses that fulfill both these requirements. Assume that the effect and side effect are monotonically increasing with dose. The minimum effective dose (MED) is the lowest dose for which the effect is at least a certain target value. Similarly, the maximal tolerable dose (MTD) is the highest dose for which the frequency of side effects (of a well-defined severity level) is not higher than an acceptable limit [12]. In case MED exceeds MTD, no doses are considered both safe and efficacious. In this case, and also often when the therapeutic window (MED, MTD) is narrow, the drug is not considered viable.

Estimating Dose–Response Curves

Up to now we have not considered the uncertainty in the dose–response curves. In practice, however, these curves always have to be estimated on the basis of a finite amount of data. Given a parametric model, the estimation of dose–response is a simple application of statistical regression theory [13–15]. However, the choice of model is an important issue. One should bear in mind George Box’s statement that “all models are wrong; some are useful”. When assessing risks, we are often interested in small probabilities, and the choice of model may then be especially critical for the conclusions.

Starting, however, with a continuous variable, y , a commonly used model for the dependence of the dose, d , is the $E_{\max}$ model [8, 16, 17]:

$$y(d) = \frac{E_0 + E_{\max} d^\gamma}{(d^\gamma + \text{ED}_{50}^\gamma) + \varepsilon} \quad (2)$$

where ε is the residual error, often assumed to be normally distributed with zero mean. At least three of the four parameters are readily interpretable. The mean effect at dose zero, that is placebo, is E_0 . The maximal placebo-adjusted effect is $E_{\max}$. Half of the maximal placebo-adjusted effect is achieved at a dose of ED_{50} , which therefore is a measure of the potency of the drug. Finally, the fourth parameter, γ , is related to the slope of the dose–response curve. The $E_{\max}$ model is quite flexible for monotonic dose–response curves. Nonmonotonic responses are possible, but are typically the result of a combination of effects going in opposite directions, as in Figure 1. One drawback with the full (sigmoidal) $E_{\max}$ model is that it has as many as four parameters, which are difficult to estimate with limited data on few doses. A number of models with fewer parameters, for example, the linear, loglinear and three-parameter $E_{\max}$ model (with $\gamma = 1$), can be seen as special cases of, or approximations to the four-parameter $E_{\max}$ model. Model selection may be part of the analysis, especially in exploratory settings, but usually the model is fixed in the study protocol and the resulting analysis is checked for robustness.

The logistic model (*see Logistic Regression*) is the most common model relating a covariate x to the probability p of an event:

$$p(x) = \frac{\exp(\alpha + \gamma x)}{(1 + \exp(\alpha + \gamma x))} \quad (3)$$

This model is rather similar to the $E_{\max}$ model. Assume that $\gamma > 0$, which is the case for side effects. Using the logarithm of the dose $\ln(d)$ as the argument x , the model can be written as

$$p(d) = \frac{d^\gamma}{(d^\gamma + \exp(-\alpha/\gamma)^\gamma)} \quad (4)$$

Using the reparametrization $\exp(-\alpha/\gamma) = \text{ED}_{50}$, we see that this expression is the same as that in the $E_{\max}$ model for the mean of $y(d)$ when $E_0 = 0$ and $E_{\max} = 1$. In some situations, there is a natural background incidence of an adverse event, so that $p(0) > 0$. Not all patients may be susceptible to a potential side effect, and $p(d)$ may be limited by a lower constant than 1. A four-parameter analog to equation (2) can then be motivated,

$$p(d) = \frac{p_0 + p_{\max} d^\gamma}{(d^\gamma + \text{ED}_{50}^\gamma)} \quad (5)$$

where $p_0 + p_{\max} < 1$. The probit model [18] is a common alternative to the logistic model. The models give similar results when having the same argument x , as long as the probability is not close to 0 or 1. However, extrapolations to small probabilities are not robust, and these two models may lead to considerably different predictions. Generally, the models used should not be trusted blindly.

It is wise to question assumptions and check how robust the conclusions are. As the dose–risk curve is usually flat around the optimum, small deviations from the best dose are often acceptable. The estimation uncertainty should also be reflected in the overall risk assessment. By combining the statistical uncertainties for the risk components, the uncertainty for the total risk may be deduced. Consequently, the uncertainty in the optimal dose can be investigated. The Bayesian approach [4, 19] uses probability distributions for the model parameters, and Bayesian probabilities for different outcomes are calculated by integrating over these distributions (*see Bayesian Statistics in Quantitative Risk Assessment*).

Good Data Do Not Always Exist

The risks associated with a new pharmaceutical agent become better understood over time. When developing a drug, animal and *in vitro* data give the basis for early risk assessments and for the choice of doses for the first studies in man (early phase I). Later on,

experiences from increasing human exposure justify drug administration for longer time in increasingly larger and more vulnerable groups of patients. The total risk level, in terms of numbers and severity of side effects, is considerably larger in late clinical trials (phase III) and for marketed drugs, than in small phase I trials. However, the risk for the exposed individual may be more significant for healthy volunteers in phase I than for patients in great need of a treatment. The risks should therefore be put into a context that also reflects the benefits for the individual. This is an essential component in analyzing whether a trial design is ethical [20].

In practice, the risks for healthy volunteers and patients in early clinical development are seldom quantified. Preclinical information is used together with rules of thumb and the experts' qualitative understanding. From preclinical dose-ranging trials, one often estimates a dose giving no toxic effects in a species, for example, mouse. Interspecies scaling is used to translate this dose to a corresponding dose in humans. This scaling may be based, for example, on body weight or plasma concentrations. Finally, a safety margin is added [21, 22]. Toxicological experts lead this process and determine a dose limit that is regarded as sufficiently safe for human administration. This process has generally worked well for many years, and severe side effects have been very rare in healthy volunteer studies. However, a first time in man study in London in 2006, resulted in severe adverse events in six healthy volunteers. This scandal has emphasized risks in early development and has lead to reports commissioned by the UK Secretary of State for Health [22] and the Royal Statistical Society [23]. Although not commonly done, it may be possible to quantify human risks based on preclinical data [23]. However, such predictions will rely heavily on the assumptions made. When assessing cancer risks, toxicologists often use what is regarded as a conservative assumption that the risk is proportional to the dose at low doses. This assumption is probably overly conservative for most pharmaceutical side effects [24]. It may be possible to predict some side effects (and their frequencies) early, on the basis of the pharmacological mode of action or similarities with already marketed pharmaceuticals.

During later clinical development and postmarketing, clinical data exist upon which risk assessments can be made. However, the dose–response data may be limited, and there is always a possibility that rare

side effects exist even if they have not been observed. Suppose that a clinical program has exposed as many as $N = 10\,000$ patients to the new pharmaceutical, without observing any lethal side effects. A standard confidence interval for the probability of a lethal side effect is then ranging from 0 to approximately $3/N$, that is, 3 in 10 000. For many pharmaceutical agents, treating less severe diseases, this upper confidence limit corresponds to an unacceptable risk. The solution is not to require that the clinical program size is made a magnitude larger, which would lead to unacceptable costs. What is needed is a qualitative risk assessment (are there any indications that severe side effects may occur?) and a monitoring of adverse events in clinical practice. In principle, it may be possible to translate partly subjective expert opinions about the relation between dose and risks of unseen events to quantified values through a process of elicitation [25], but the value of this for practical decision making is unclear. A discussion about estimating risks that cannot be observed directly can be found in [26].

Dose–response is not routinely addressed explicitly in postmarketing and late stage clinical trials. However, Food and Drug Administration representatives have suggested that inclusion of more than one dose in phase III could be useful for dose determination. Also, late stage trial data and postmarketing data may be combined with information from earlier phases and may occasionally motivate a change in the recommended therapeutic dose.

Designing Dose-Finding Trials

When designing a dose-finding trial, or any other clinical trial, the team has to consider both ethics and efficiency. “When considering the design of a dose-ranging study, we must first consider patient safety” to quote [27]. It is almost tautological that only ethical trials should be proposed by the sponsor, and only such trials should be approved in the mandatory review by ethical committees. The question is what constitutes an ethical trial. Much has been written about this topic and there is broad consensus that informed consent from the patient (or his guardian) is one requirement. However, it is generally not possible to lay all responsibility on the patient. It is part of the physician’s job to digest and evaluate relevant medical information; the decision cannot be

handed over solely to the patient. The most important ethical codex for clinical research is the Declaration of Helsinki [28], which states “It is the duty of the physician in medical research to protect the life, health, privacy, and dignity of the human subject. Physicians should cease any investigation if the risks are found to outweigh the potential benefits”. Many authors have been discussing “clinical equipoise” as a motivation for the ethics of a trial [29]. The definition is rather vague and may lead to an unclear distribution of responses. The treating physician has to take personal responsibility for the optimal treatment of his or her patient. It can therefore be argued that the physician has to evaluate whether it is better for the patient not to take part in a trial, and receive standard treatment, rather than to be randomized to one of the trial arms. This evaluation may apply a quantitative risk assessment, as described above [20, 30]. The same approach may be used by trial sponsors and ethical committees when proposing a design and evaluating a proposed design.

Within the class of designs that are considered ethical, the sponsor will search for as efficient a design as possible. A guiding principle is to use a wide dose range. It is obvious, however, that the same design will not be optimal for all dose-finding trials. First, the ethical constraints have to be acknowledged. Secondly, the design has to be placed in the full drug development context, including what is known about effect and safety before the trial, what the precise objectives of the study are, operational possibilities and restrictions, regulatory concerns, and the plans for later trials in a confirmatory phase. A good start for the design work is to model the available data [31]. As the reason for conducting an experiment is to generate information and reduce uncertainty, a critical part of the modeling work is to understand and quantify how uncertain different model components are, before the experiment, *cf.* [23]. It is also critical to consider carefully what the objectives of the trial should be, and how these objectives and the data generated in the trial would help decide whether to proceed to the confirmatory phase; and in that case, how this phase should be designed [32]. Dose-finding trial objectives are unfortunately often formulated in vague terms as e.g., “to investigate the dose–response relationship”. Unless such statements are translated into a clear quantitative objective, it is not possible to optimize the trial design. The resulting design may happen

to be reasonable, but the lack of transparency and analysis may also lead to suboptimal and unethical designs.

When there exist an agreed weighting of different outcomes (see sections titled “Risk–Benefit” and “Choosing the Relative Weights”), the main objective of a dose-finding trial is, naturally, to find the dose that minimizes the overall risk. Strictly speaking, the optimal dose is not the same in all patients, but a good design to estimate the dose that optimizes the population-average of the risk is likely to be a design that also generates useful information for individual dose optimization. To simplify the problem, assume that we have models with known uncertainty for effect and safety, a well-defined value function of these aspects, and are to choose one single dose, d^* , after the dose-finding trial. Given a proposed design and a proposed method to find the estimate d^* , it is straightforward to simulate the trial results [33] and calculate the corresponding risk. The algorithm first simulates one possible truth, that is, it selects model parameters reflecting the uncertainty in them. Thereafter, the result of the proposed trial is simulated, and the dose d^* is estimated on the basis of the simulated trial data. The risk is then calculated based on d^* and the simulated model parameter values. This simulation of risk is iterated a large number of times and the average simulated risk is calculated. Non-adaptive designs having the same sample size can often be compared directly so that the design that has lowest simulated risk is optimal within the class. The average risk will, however, decrease with the sample size. Thus, the reduced risk has to be contrasted against the sample size, cf. the literature about optimal sample size [34]. Adaptive designs, including group-sequential approaches, are generally more efficient than designs with fixed sample size. The gain, in terms, for example, of expected sample size giving the same average risk, should be contrasted to the cost and complexity of adaptations [32]. Zohar and O’Quigley [35] suggest an adaptive design for dose finding in oncology.

The books by Ting [12] and Chevret [36] deal with adaptive and nonadaptive dose-finding studies. When the risk cannot be readily defined, the objective of the trial may have to be formulated in other terms. The goal may, for example, be to find the dose giving a certain effect [32] or to estimate the dose for which the derivative of the effect with respect to $\ln(\text{dose})$ is a specified value. Phase III may be

used to fine-tune the dose; not just one dose, but two or three are included in these trials. Inclusion of more than one dose is especially useful when phase II gives limited safety data. It can be expected that the safety profile will be considerably better understood based on phase III data, and the optimal dose should therefore be chosen after receiving this information. A current hot research topic is seamless phase II + III trials, combining dose finding with a confirming phase in one single study [37]. The International Conference on Harmonisation (ICH) [38] gives a regulatory perspective on dose–response.

References

- [1] Raiffa, H. (1968). *Decision Analysis; Introductory Lectures on Choices Under Uncertainty*, McGraw-Hill, New York.
- [2] Goodwin, P. & Wright, G. (1998). *Decision Analysis for Management Judgment*, 2nd Edition, John Wiley & Sons, New York.
- [3] Hunink, M.G.M., Glasziou, P.P., Siegel, J., Weeks, J., Pliskin, J., Elstein, A. & Weinstein, M. (2001). *Decision making in health and medicine: Integrating evidence and values*, Cambridge University Press, Cambridge.
- [4] Parmigiani, G. (2002). *Modeling in Medical Decision Making*, John Wiley & Sons, Chichester.
- [5] Protheroe, J., Fahey, T., Montgomery, A.A. & Peters, T.J. (2000). The impact of patients’ preferences on the treatment of atrial fibrillation: observational study of patient based decision analysis, *British Medical Journal* **320**, 1380–1384.
- [6] von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.
- [7] Raiffa, H. & Schlaifer, R. (1961). *Applied Statistical Decision Theory*, John Wiley & Sons, New York, reprinted 2000.
- [8] Rowland, M. & Tozer, T.N. (1995). *Clinical Pharmacokinetics*, 3rd Edition, J.B. Lippincott, Philadelphia.
- [9] McDonough, C.M. & Tosteson, A.N.A. (2007). Measuring preferences for cost-utility analysis, *Pharmacoeconomics* **25**, 93–106.
- [10] Lehmann, E.L. (1986). *Testing Statistical Hypothesis*, 2nd Edition, John Wiley & Sons, New York.
- [11] French, S. & Ríos Insua, D. (2000). *Statistical Decision Theory*, Arnold Publishing, London.
- [12] Ting, N. (ed) (2006). *Dose Finding in Drug Development*, Springer Science, New York.
- [13] van Belle, G., Fisher, L.D., Heagerty, P.J. & Lumley, T. (2004). *Biostatistics; A Methodology for Health Sciences*, 2nd Edition, John Wiley & Sons, Hoboken.
- [14] McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*, John Wiley & Sons, New York.

- [15] Lindsey, J.K. (1996). *Parametric Statistical Inference*, Oxford University Press, Oxford.
- [16] Gabrielsson, J. & Weiner, D. (2001). *Pharmacokinetic and Pharmacodynamic Data Analysis, Concepts and Applications*, 3rd Edition, Swedish Pharmaceutical Press, Stockholm.
- [17] MacDougall, J. (2006). Analysis of dose-response studies-Emax model in *Dose Finding in Drug Development*, N. Ting, ed, Springer Science, New York.
- [18] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, John Wiley & Sons, New York.
- [19] Spiegelhalter, D.J., Abrams, K.R. & Myles, J.P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluations*, John Wiley & Sons, Chichester.
- [20] Burman, C.-F. & Carlberg, A. (2008). Future challenges in the design and ethics of clinical trials, submitted for inclusion in *Clinical Trial Handbook*, S. Gad, ed, John Wiley & Sons, New York.
- [21] Salsburg, D. (2006). Dose finding based on preclinical studies, in *Dose Finding in Drug Development*, N. Ting, ed, Springer, New York.
- [22] ESG (Expert Scientific Group on Phase One Clinical Trials) (2006). Final Report. Department of Health, London, at <http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH.063117> (access Apr 2007).
- [23] RSS (Royal Statistical Society) (2007). *Report of the Working Party on Statistical Issues in First-in-Man Studies*, London.
- [24] Gaylor, D.W. (1998). Extrapolation, low-dose, in *Encyclopedia of Biostatistics*, P. Armitage & T. Colton, eds, John Wiley & Sons, Chichester.
- [25] O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, R., Garthwaite, P., Jenkinson, D., Oakley, J. & Rakow, T. (2006). *Uncertain Judgements; Eliciting Experts' Probabilities*, John Wiley & Sons, Chichester.
- [26] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Edition, Lippincott Williams & Wilkins, Philadelphia.
- [27] Dmitrienko, A., Fritsch, K., Hsu, J. & Ruberg, S. (2007). Design and analysis of dose-ranging clinical studies, in *Pharmaceutical Statistics using SAS; A Practical Guide*, A. Dmitrienko, C. Chuang-Stein & R. D'Agostino, eds, SAS Press, Cary.
- [28] WMA (World Medical Association) (2000). *Declaration of Helsinki*, at <http://www.wma.net/e/policy/b3.htm> (access Apr 2007).
- [29] Freedman, B. (1987). Equipoise and the ethics of clinical research, *The New England Journal of Medicine* **317**, 141–145.
- [30] Lilford, R.J. (2003). Ethics of clinical trials from a Bayesian and decision analytic perspective: whose equipoise is it anyway? *British Medical Journal* **326**, 980–981.
- [31] Burman, C.-F., Hamrén, B. & Olsson, P. (2005). Modelling and simulation to improve decision-making in clinical development, *Pharmaceutical Statistics* **4**, 47–58.
- [32] Bornkamp, B., Bretz, F., Dmitrienko, A., Enas, G., Gaydos, B., Hsu, C.-H., König, F., Krams, M., Liu, Q., Neuenschwander, B., Parke, T., Pinheiro, J., Roy, A., Sax, R. & Shen, F. (2007). White paper of the PhRMA PISC working group on adaptive dose-ranging designs, *Journal of Biopharmaceutical Statistics* November 2007.
- [33] Holford, N.H.G. (2000). Simulation of clinical trials, *Annual Review of Pharmacology & Toxicity* **40**, 209–234.
- [34] Pezeshk, H. (2003). Bayesian techniques for sample size determination in clinical trials: a short review, *Statistical Methods in Medical Research* **12**, 489–504.
- [35] Zohar, S. & O'Quigley, J. (2006). Identifying the most successful dose (MSD) in dose-finding studies in cancer, *Pharmaceutical Statistics* **5**, 187–199.
- [36] Chevret, S. (ed) (2006). *Statistical Methods for Dose-Finding Experiments*, John Wiley & Sons, Chichester.
- [37] Maca, J., Bhattacharya, S., Dragalin, V., Gallo, P. & Krams, M. (2006). Adaptive seamless phase II/III designs – background, operational aspects, and examples, *Drug Information Journal* **40**, 463–473.
- [38] ICH (International Conference on Harmonisation) (1994). *Dose-Response Information to Support Drug Registration*, at <http://www.fda.gov/Cder/guidance/iche4.pdf> (access Apr 2007).

Related Articles

Cost-Effectiveness Analysis

Detection Limits

Dose–Response Analysis

CARL-FREDRIK BURMAN

Coding: Statistical Data Masking Techniques

The rapid expansion of government and company databases combined with an ever growing use of the web and mobile devices has led to concomitant increase in sensitive and confidential personal data in the public domain. The US Government relies on the use of such data to identify and track terrorists and their activities. In an effort to ensure domestic security, the Homeland Security bill supports the development and implementation of advanced information technologies (e.g., data mining) for aggregating and analyzing data (*see* **Environmental Performance Index**) available both in public and private domains [1, 2]. The most recent example is the Automated Targeting System (ATS), the Department of Homeland Security (DHS) database that contains information on US citizens and foreign travelers, that is used for creating a terrorist risk profile for anyone crossing the US border [2]. The ATS data originate from multiple sources, various federal and commercial databases, and airlines' passenger name record data. The information from ATS may be shared with other DHS agencies, and airport and airlines operators. Do such data sharing and mining activities provide effective means for identification of terrorists without posing infringements on privacy of nonterrorists?

Most of the criticism of homeland security has focused on utility of data mining algorithms (*see* **Risk in Credit Granting and Lending Decisions: Credit Scoring**) in this context, and the potential high false positive rate (*see* **Microarray Analysis; Public Health Surveillance**), i.e., misidentification of individuals and their behaviors as terrorists [2–4]. Data mining algorithms require either aggregating multiple sources into a single data warehouse or mining multiple distributed databases. In both cases, record linkage methodologies and privacy protection methods play an important role [4, 5]. Traditionally, the government agencies have maintained data confidentiality by sharing data in two ways. In the *restricted access* setting only approved personnel can access data at the designated data centers. Owing to the presence of inaccurate and incomplete data, restricted access does not preclude unintended information disclosure and misidentification since most mining algorithms are based on statistical models and

assumptions that data are free of errors. An alternative method is to add a layer of privacy by changing data in some way. Allowing open access to such *masked data* offers more flexibility for mining distributed databases.

Coding broadly refers to altering the data in some way. In computer science literature, it may refer to encryption, compression, or error correction. These are relevant for cryptography and cryptanalysis, and for data and homeland security. In statistical confidentiality literature, coding encompasses various data masking techniques. These techniques generally introduce bias and variance to data. The question is how much and what kind of bias and variance is introduced and how this influences the utility of data mining algorithms and precision of identification, that is the false positive error rate. This article gives a brief overview of statistical disclosure limitation techniques for data masking, pointing out the likelihood of their use and potential impacts for data privacy. For more details on disclosure limitation and these techniques, see [6–8], and for related privacy preserving data mining techniques, see [4, 5, 9].

Statistical Disclosure Limitation and Data Masking

In statistical disclosure limitation, masking methods introduce bias and variance to data in order to minimize identity and attribute disclosure while trying to retain sufficient information needed for proper statistical inference. The goal of disclosure limitation is to examine and manage a trade-off between data utility and disclosure risk [8, 10, 11]. Data utility is a measure of usefulness of a dataset for an intended analyst. Disclosure risk measures the degree to which a dataset and its released statistics reveal sensitive information. This notion is probabilistic; as released data accumulate in the public domain, the probability of uniquely identifying members of the population increases. Vulnerabilities of public databases are evident in daily news, and they point out that neither a simple removal and replacement of identifiers nor encryption by itself is a sufficient mode of protection. An article in New York Times recently professed that a woman's identity was easily uncovered from an on-line AOL's database [12], while the Chicago Tribune claimed an easy identification of CIA employees on the basis of a web search [13].

It is, therefore, necessary to include additional levels of privacy protection.

Data masking involves transforming an $n \times p$ (cases by variables) original data matrix Z through pre- and postmultiplication and the possible addition of noise. That is, $Z \rightarrow AZB + C$, where A is a matrix that operates on the n cases, B is a matrix that operates on the p variables, and C is a matrix that adds perturbations or noise. Matrix masking can be applied to either microdata or table of counts, and includes a variety of standard approaches to disclosure limitation: (a) recodings (e.g., rounding and thresholding), (b) cell suppression, (c) data swapping, (d) perturbation, and (e) sampling (e.g., releasing subsets of data or variables – deleting rows or columns of Z), and simulation (e.g., adding rows to Z).

Rounding and *recoding* are the most commonly used methods and refer to aggregating high-risk variables into coarser values and categories. Income and money transactions are among the most data-mined characteristics and activities in counterterrorism searches. However, rounding such sensitive values to the nearest \$1000 may produce inconsistent data, i.e., the marginal totals may be hard to preserve. If rounded data are then recoded into intervals, these intervals may be biased because the distribution within an interval may not be uniform. *Blurring* is a related method, where a sensitive value in the microdata file is replaced by an average based on a group of records. The intervals may be produced *via top coding*, *bottom coding*, and/or *global recoding*. Top coding recodes the right tail of the distribution with the last k values of a variable into a new category. For example, grouping of all income values above \$200,000 is often done on US Census data. Bottom coding recodes the lower end of the distribution, while global recoding is applied to whole dataset. After tabulating the data, the cell frequencies may still be considered sensitive by some threshold rule. Agencies typically require at least three to five respondents in a cell. If this is the case, the categories may be further collapsed into a smaller set of categories, which lowers the data utility and increases the risk of false positives.

Suppression can be applied to actual microdata values, but is most often associated with suppressing cells in higher-dimensional tables. *Cell suppression* is a technique that has been widely used by statistical agencies where sensitive cells (and possibly others)

are not released. This method has been criticized for lack of generalization and statistical content for the higher-dimensional tables, and potentially false inferences. The suppressions introduce “holes” in the data and nonrandom missingness, thus creating difficulty in performing correct record linkages and statistical analyses.

Data perturbation includes methods such as *adding random noise* to the original data, and reporting the altered data or a table. In order to reduce the disclosure risk, we can add a random noise to a sensitive numerical value (e.g., add a value of a normally distributed random variable to income). A related technique for count data should provide consistency between perturbed and original tables. In *random* and *controlled rounding*, every cell is a multiple of some rounding base. In a *controlled tabular adjustment* technique, the original sensitive value in a table is replaced by another “safe but sufficiently close” value. The problem with these methods is that they change marginal distributions, which may produce faulty statistical correlations and pattern matching in data mining algorithms. Controlled rounding is designed to preserve marginal totals but, like other methods, works well only for two-way tables.

Data swapping can be considered a type of perturbation. This method chooses random pairs of respondents and exchanges a fraction of their data. The US Census Bureau uses data swapping for census tabulations, but typically swaps a small percent of data as it may impair correlations. In this setting, a search on only a few characteristics such as income, level of education, and frequency of travel, for example, can easily lead to misclassification of individuals. Another criticism of this approach is that it does not provide a general method for performing data swaps for k -way tables. Fienberg and McIntyre [6] describe swapping as a type of cell perturbation that involves trading values of some variables while holding others fixed, and generalization for k -way tables.

Most recently, agencies have shifted their focus to *sampling and simulation* techniques. *Fully synthetic* data methods use Bayesian methodology and release multiple-imputed datasets instead of the original ones. The probability of uniquely identifying an individual in such data is extremely low, but for counterterrorism purposes this will produce a high number of false positives. In related *partially synthetic* data methods, released data are composed of

the original nonsensitive and sensitive units that are replaced in a released dataset by synthetic values. Raghunathan *et al.* [14] describe methods for drawing valid inferences from such datasets. The US Census Bureau is currently working on implementing these and related methods, and on building synthetic datasets.

Related methods for contingency tables utilize Markov bases and Markov chain Monte Carlo sampling algorithms. These methods replace the original table by a random draw from the exact distribution of the tables under a log-linear model whose sufficient statistics are released marginal totals (see [9], and various articles in [7]). The released table can be considered an instance of a fully synthetic or a partially synthetic dataset produced with a special type of data swapping. The agency can also opt for releasing *partial information* by releasing only the relevant margins instead of a full table. To assess disclosure risk for the original table, these methods use undirected graph theory in connection with log-linear models as a framework to compute bounds on cell entries given a set of released margins. They also generalize the Bonferroni–Fréchet–Hoeffding bounds for cell entries in k -way contingency tables given such partial information. Alternatively, these bounds can be used as recoding intervals for released data providing another form of partial or fragmentary data release. Fienberg and Slavkovic [9] extend these results to bounds, simulations, and samplings from the posterior distribution of tables that satisfy other constraints (e.g., conditional probabilities and odds ratios), and tie them to mining of association rules.

Some argue that “selective revelation” techniques may significantly reduce privacy concerns [4]. A selective revelation is a method that attempts to separate “transactional” data (e.g., ticket purchase, money transfers, etc.) from identity, and reveals information incrementally. Since the bounds are calculated by accounting for the collective on the basis of partially released information, a potential exists to dynamically update the safe releases, and incrementally reveal identities as purported by selective revelation.

The importance of simulation/sampling methods is that we can generate distributions of all possible tables given the margins and other statistics, and hence establish a probabilistic framework for disclosure assessment and statistical inference. Risk of identification in such datasets is minimal. The utility is tied to the family of posterior predictive

distributions from which data have been drawn. However, what effects these methods have on record linkage and privacy preserving data mining methods is still unclear; that is, how will information from various databases be combined, and how will this affect the false positive error rate and profiling?

Conclusions

This article has outlined some data masking techniques as a way of recoding data to protect privacy. Most of the aforementioned masking methods introduce bias and variance to data, and create sparse, partial, and fragmentary data. Such data present methodologically hard statistical problems and can cause errors in naive data mining algorithms whose main goal is to uncover patterns in data. Along with the low prevalence of terrorists in a large pool of nonterrorists, these methods exacerbate the already present issues of producing many false positives and posing infringements on privacy.

Effective search for terrorists without infringement on privacy is almost impossible. It is a “needle in a haystack problem” as (a) the volume of available data is enormous and (b) the relevant records are often hidden within vast amounts of irrelevant data [15]. The solution involves complex interplay between confidentiality promises made to the respondents and the government’s need for fighting terrorism. It requires work on the optimization problem of maximizing available (“correct”) information needed for identification of terrorists, but minimizing the false positives and the risk of undesirable data disclosure. There has to be a careful consideration of the data context and analysis goals. This is a dynamic system, and as new technologies emerge, the current privacy practices (e.g., statistical, legal, ethical) are challenged and require multidisciplinary work. Every method requires a careful cost–benefit analysis, a decision-theoretic solution, and merging of different tools from statistics, privacy preserving data mining, and cryptography and cryptanalysis.

Acknowledgment

The research reported here was supported in part by NSF grant SES-0532407 to the Department of Statistics, Pennsylvania State University. The article has benefited from author’s discussion with S. Fienberg and R. Schilder.

References

- [1] Office of Homeland Security (2002). *The National Strategy for Homeland Security*, <http://www.whitehouse.gov/homeland/book/index.html>.
- [2] Seifert, J.W. (2007). *CRS Report to Congress RL31798 – Data Mining and Homeland Security: An Overview*, <http://www.fas.org/sgp/crs/homesecc/RL31798.pdf>.
- [3] Taipale, K.A. (2003). Data mining and domestic security: connecting the dots to make sense of data, *Columbia Science & Technology Law Review* **5**, <http://www.stlr.org/cite.cgi?volume=5&article=2>.
- [4] Fienberg, S.E. (2006). Privacy and confidentiality in an e-commerce world: data mining, data warehousing, matching and disclosure limitation, *Statistical Science* **21**(2), 143–154.
- [5] Agrawal, R., Evfimievski, A., & Srikant, R. (2003). Information sharing across private databases, *Proceedings 2003 ACM SIGMOD International Conference on Management of Data*, ACM Press, New York, pp. 86–87.
- [6] Fienberg, S.E. & McIntyre, S.E. (2005). *Report on Statistical Disclosure Limitation Methodology*, Statistical Policy Working Paper 22, <http://www.fcsm.gov/working-papers/spwp22.html>.
- [7] Domingo-Ferrer, J. & Torra, V. (eds) (2004). *Privacy in Statistical Databases – PSD'2004*, Lecture Notes in Computer Science No. 3050, Springer-Verlag, pp. 58–72.
- [8] Doyle, P., Lane, J., Theeuwes, J. & Zayatz, L. (eds) (2001). *Confidentiality, Disclosure and Data Access. Theory and Applications for Statistical Agencies*, Elsevier, New York.
- [9] Fienberg, S.E. & Slavkovic, A.B. (2005). Preserving the confidentiality of categorical statistical databases when releasing information about association rules, *Journal of Data Mining and Knowledge Discovery* **11**, 155–180.
- [10] Duncan, G.T. & Stokes, S.L. (2004). Disclosure risk vs. data utility: the R–U confidentiality map as applied to topcoding, *Chance* **17**(3), 16–20.
- [11] Trottini, M. (2003). *Decision Models for Data Disclosure Limitation*, Ph.D. Thesis, Department of Statistics, Carnegie Mellon University.
- [12] Barbaro, M. & Zeller Jr, T. (2006). A face is exposed for AOL Searcher No. 4417749, *The New York Times* August 9, <http://www.nytimes.com/2006/08/09/technology/09aol.html?r=2&ei...epag&adxnlnx=1155326605-/1FEV853bC3qmt0ZgsF3hw&pagewanted=print&oref=slogin>.
- [13] Crewdson, J. (2006). Internet blows CIA cover, *Chicago Tribune* March 12, <http://www.chicagotribune.com/news/nationworld/chi-060311ciamain-story,1,123362.story?coll=chi-news-hed>.
- [14] Raghunathan, T.E., Reiter, J.P. & Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation, *Journal of Official Statistics* **19**, 1–16.
- [15] The National Academies Committee on Science and Technology for Countering Terrorism (2002). *Making the Nation Safer: The Role of Science and Technology in Countering Terrorism*, The National Academy Press, Washington, DC.

ALEKSANDRA B. SLAVKOVIC

Cohort Studies

Cohort studies provide the most comprehensive approach for evaluating overall patterns of health and disease. Every epidemiologic study should ideally be based on a particular population's experience over time. The cohort design has the advantage, relative to other study designs, that ideally it includes all of the relevant person-time experience of the population under study, whereas other study designs, such as case-control and cross-sectional studies, involve sampling from that experience. In general, cohorts are defined by the nature of their exposures. We illustrate the design and analytical features of cohort studies using examples primarily from studies of occupational and environmental exposures, since these are the exposures most commonly involved in quantitative risk assessment. Nevertheless, the basic principles also apply to other types of cohort studies.

Basic Cohort Design

In a cohort study, an investigator may follow a population into the future, i.e., a *prospective cohort study*, or may follow an historical cohort to the present, i.e., an *historical cohort study* (also known as *retrospective cohort study*). Some studies combine features of both prospective and historical designs.

For studies investigating environmental exposures, a cohort is usually based on a particular community (or a sample of the community) that is followed (usually prospectively) over time. This usually requires a special survey to be conducted at the start of the follow-up period. Follow-up may then involve either repeated surveys, or linkage with routine records (e.g., hospital admissions and mortality records). For example, Grandjean *et al.* [1] studied a cohort of 1022 consecutive singleton births in the Faroe Islands, and assessed prenatal methylmercury exposure from the maternal hair mercury concentration; 917 of the children then underwent detailed neurobehavioral examination at about age 7 years.

Alternatively, a cohort can be constructed from routine population records for the entire population of a particular area, without contacting each person individually. For example, Magnani *et al.* [2] studied incident cases of pleural malignant mesotheliomas in the town of Casale Monferrato during

1980–1991. The population “at risk” was estimated from population statistics for the local health authority area. The analysis excluded incident cases for which there was some suggestion of occupational or paraoccupational exposure in order to estimate the general population risks for nonoccupational exposure from living in the vicinity of the plant.

Cohorts may also be constructed on the basis of occupational exposures, (*see Occupational Cohort Studies*) i.e., the cohort members are defined according to where they work rather than where they live. One option is to include “all workers who worked for at least 1 month in (a particular) factory at any time during 1969–1984” (e.g. [3]). Perhaps a more common option is to include workers from multiple plants operated by different companies, but engaged in the same industrial processes. For example, Demers *et al.* studied 27 464 men employed by 14 sawmills for 1 year or more between 1950 and 1995. International studies, which combine occupational populations from similar facilities located in different countries, are extensions of this approach (e.g. [4]). The advantage of restricting the cohort to workers from one facility, is that characterization of exposures may be more consistent and precise when work history and environmental data are obtained from a single source. However, pooling cohort members from multiple facilities may be necessary to increase study size, especially when outcomes of interest are rare. When a specific exposure is of interest, cohorts can be pooled based on their shared exposure patterns in multiple workplace or environmental settings. The international collaborative historical cohort study of workers exposed to phenoxy herbicides, which focused primarily on soft tissue sarcomas and non-Hodgkin's lymphoma [5], is a good example.

A *fixed cohort* is one in which the cohort is restricted to workers employed (or people living in the community) on some given date. By contrast, a *dynamic cohort* includes all workers who would be enumerated in a fixed cohort and workers hired subsequently (or people who moved to the community subsequently). Fixed and dynamic cohorts are sometimes referred to as *closed* or *open* cohorts, respectively. The choice of a fixed or dynamic cohort may be dictated by the availability of data or by special exposure circumstances that occurred during some particular (usually brief) time period. Ott and Zober's [6] study of a group of workers exposed

to dioxin during an industrial accident at a German trichlorophenol manufacturing plant, or studies of A-bomb survivors [7], are examples where a fixed cohort would be preferred – although in the latter instance, concern about possible effects on offspring of those exposed has also led to the establishment of a dynamic cohort of their children who were born after the event [8].

In community-based cohorts, comparisons are usually made internally between study participants exposed and those not exposed to a particular risk factor, e.g., high *versus* low arsenic intake in well water (see **Water Pollution Risk**) [9]. In studies of occupational populations, an internal comparison may also be possible, e.g., by comparing workers with high carbon black exposure to those with low carbon black exposure [10]. However, in some instances this may not be possible because good individual exposure information is not available or because there is not sufficient variation in exposure within the population (e.g., because everyone who worked in the factory had high exposure). In this situation, an external comparison may be made, e.g., with national or regional death rates or cancer registration rates.

For practical reasons, very few cohort studies address lifelong exposure and disease experience. Instead, cohort studies are usually restricted to time periods for which such information can be obtained; thus health effects that occur outside of this time period (e.g., those with a long latency period) may not be detected. Additionally, an inappropriate choice of comparison groups (e.g., national populations compared to worker cohorts) can lead to confounding. The healthy worker effect, which is a manifestation of the latter type of bias, poses a major problem for occupational cohort studies involving external comparisons [11]. Several approaches for reducing bias caused by the healthy worker effect are available. One strategy is to compare disease rates among the workers in the cohort with rates in an external reference population also consisting of employed (and retired) workers, but such information is rarely available. Alternatively, the reference population might be selected from among employees in an industry (or community) other than the one under study. For example, in their study of cardiovascular diseases among viscose rayon factory workers, Hernberg *et al.* [12] selected as a comparison group workers at a paper mill not exposed to the suspected etiologic factor, carbon disulfide.

Another approach is to use an internal reference population, consisting of some segment of the full study cohort, usually assumed to be nonexposed to the hazardous agents of concern [13]. This approach may not eliminate bias from the healthy worker effect, but will usually reduce it [14]. Stratification on employment status (active *versus* inactive) [15] and on length of follow-up [14] may also help minimize healthy worker effect bias in internal comparison analyses. An additional advantage of using an internal reference population is that similarity of data quality is anticipated for all groups compared.

Cohort Enumeration and Follow-up

Cohort Enumeration

In a community-based cohort study, enumeration of the cohort usually requires an initial community survey, followed by prospective follow-up. The enumeration of an historical occupational cohort usually requires assembly of personnel employment records containing dates of first and last employment and the various jobs held, with their associated dates. In occupational studies, short-term transient workers may be excluded since they are often difficult to trace, and may have atypical lifestyles that make them noncomparable to longer-term workers [16]. For example, Boffetta *et al.* [17] found markedly elevated mortality rates from external causes of death (e.g., motor vehicle accidents), but only slight differences for cancer and ischemic heart disease, among workers employed for less than 1 month compared to longer-term workers in two international cohorts of workers in the man-made vitreous fiber and reinforced plastics industries. In dynamic community-based studies of environmental exposures, short-term residents may be excluded for the same reasons.

Follow-up

In some instances, particularly in community-based studies, follow-up may involve regular contact with the study participants, including repeated surveys of health status. Perhaps more commonly, follow-up may be done by routine record linkage. For example, study participants may be followed over time by linking the study information with national death records, or disease incidence records (e.g., a national

cancer registry), as well as with other record systems (e.g., social security records, drivers license records) to confirm vital status in those who are not found to have died during the follow-up period.

Although most developed countries have complete systems of death registration, and it is easy, in theory, to identify all deaths in a particular cohort, this may not be so straightforward in practice. For example, many countries do not have national identification numbers and record linkage may have to be done on the basis of name and date of birth using record linkage programs to identify “near matches”. A further problem is that some countries do not have national death registrations, and these may be done on a regional or state basis instead, making it necessary to search multiple registers. Moreover, just because someone has not been identified in death records, this does not mean that they are still alive and “at risk” since they may have emigrated or may not have been identified in death registrations for some other reason. It is therefore desirable to confirm that they are alive using other record sources such as drivers license records, voter registrations, social security records, etc. [11].

Table 1 summarizes the various sources of data on vital status, cause of death, and disease incidence. If coding the cause of death is not performed routinely by national or regional health statistics agencies, then a copy of each death certificate should be obtained and coding should be performed by a nosologist trained in the rules specified by the International Classification of Diseases (ICD) volumes compiled by the World Health Organization (<http://www.who.int/classifications/icd/en/>). Revisions to the ICD have been made over the years, and changes in coding for some causes may influence the mortality findings of the study. If a comparison is being made with national mortality rates, then it is important that the cohort study deaths should be coded using the same ICD revision as was used for national data during the same time period. Death certificates in most countries record underlying and contributing causes of death. The underlying cause is generally of most interest in occupational mortality studies, although information on contributing causes and other significant medical conditions may be useful for identifying some health outcomes (e.g., diabetes, hypertension, pneumoconiosis) [18, 19].

Data on disease incidence, rather than mortality, can be obtained provided that there are

Table 1 Sources of vital status and disease occurrence^(a)

Source	Data supplied
Social Security Administration (US)	Vital status, date and location of death
National Death Index (US)	Date, location, cause of death
Motor vehicle bureaus	Alive status inferred from license or citation issuance
Voter registration lists	Alive status, inferred
Credit bureau listings	Alive status, inferred
National Office of Pensions and Insurance (UK)	Vital status, location, date of death
Population-based disease registries (Sweden, Finland, among others)	Vital status, cause of death, incidence of specific diseases, location, date of occurrence
Vital statistics bureaus	Death certificates, birth certificates
National insurance and benefits offices	Alive status (inferred)
Company medical and insurance claims	Death certificates, illness and injury occurrence
Unions and professional societies	Death certificates, illness and injury occurrence

^(a) Reproduced from [11]. © Oxford University Press, 2004

population-based disease registers or that special incidence surveys have been conducted on the workforce. National registries are only available in a few countries, but there are a number of population-based regional cancer registries around the world [20], such as the Surveillance, Epidemiology, and End Results (SEER) cancer registries in the United States [21].

In a cohort mortality study, there are two types of participants for whom vital status information may be missing [11]. The first are participants of unknown vital status. One should include all of the available information in the analysis by counting person-years of observation for unknowns up until the dates of last contact, typically the dates of termination from the industry (or the dates of leaving the community). The assumption in this approach is that there is no systematic bias resulting from the loss of information on the subsequent period, i.e., that the association between exposure and disease in

the “missing data” is the same as that in the cohort as a whole.

The second category of study participant with missing information comprises those with deaths of unknown cause. This situation usually arises when death certificates cannot be located, but may also occur when cause of death information on some death certificates is judged by the nosologist to be inadequate for coding. The usual approach is to create an unknown cause of death category because this approach does not require unverifiable assumptions about the distribution of cause of death among the unknowns.

Data Analysis

Person-Time Analyses

An understanding of the analytic techniques used in cohort studies requires an appreciation of the distinction between disease risks and rates. A *risk* is the probability of developing or dying from a particular disease during a specified time interval. Thus, if we conduct a 10-year follow-up on a cohort of 1000 people, and we have no losses to follow-up, and we observe 30 deaths, then the 10-year risk of death is 30/1000 people.

A second approach is to compute the number of newly occurring, or *incident*, cases during the time interval of follow-up per number of *person-years of observation*. This quantity is a disease *rate* [22]. The value of computing rates, rather than risks, in cohort studies is that rates take into account the person-time of observation, which is likely to vary between cohort members when there are losses to follow-up, when participants die from other diseases than the outcome under study, or when the cohort is dynamic.

Computation of person-time data has become standard practice in dynamic occupational cohort studies since the 1950s [11]. Each subject alive at the beginning of follow-up contributes person-years of observation until he or she either develops or dies from the disease of interest, dies from another cause, or the end of follow-up occurs. When a minimum employment (or residency) duration of x years is imposed in a cohort (or community-based) study, then follow-up for each participant should be started either at the date at which x years of service (or residency) had been attained, or at the date of cohort enumeration, whichever occurred later. The reason

for this shift in follow-up commencement date is that, by virtue of the cohort inclusion criterion, it is known that each participant survived for at least x years. Thus, his first x years are effectively free from risk. Failure to account for these x years as “risk free” would tend to underestimate disease rates in the cohort. Thus, if a study involves follow-up for the period January 1, 1980 to December 31, 2005, and a person started work (or moved into the area) on July 7, 1993, and there is a 1-year minimum duration of employment (or residency) criterion, then follow-up for this participant would commence on July 7, 1994 – unless they had left work (or moved) in the meantime, in which case they would be excluded from the study, unless they subsequently returned and eventually met the minimum employment (or residency) criterion.

Similarly, follow-up for each cohort member finishes when they die, develop the outcome of interest (in an incidence study), are lost to follow-up, or the overall study follow-up finishes, whichever date is earliest. Thus, if a study involves follow-up for the period January 1, 1980 to December 12, 2005, and a person died, developed the outcome of interest (in an incidence study), or was lost to follow-up on July 7, 2003, then follow-up for this participant would finish on this date.

Thus, length of follow-up and length of employment (or length of residency) are not usually identical – someone may have accumulated many years of employment (or residency) if they commenced work (or moved to the area) before the study follow-up commenced.

Stratified Analysis

The epidemiologic method for summarizing person-years across age and calendar years is to stratify the cohort’s collective person-years into a cross-classified distribution [11]. Age and calendar year are the two most important stratification variables because many diseases are strongly age-related, and population rates (even within narrow age categories) vary over time [23]. Length of employment, age at first employment, time since first employment, and employment status (active/inactive) can also influence disease rates, primarily as confounders or effect modifiers of the associations between health outcomes and the exposures of interest. Person-years accumulation is therefore accomplished by summing

each cohort member's contribution to every attained category, where age and calendar year strata jointly specify the categories.

To illustrate the method of stratification, consider a worker for whom follow-up commenced on January 1, 1987, at which time he/she was exactly 28 years. He/she would contribute 2 person-years to the 25–29 year age and 1985–1989 calendar year jointly classified stratum, then 1 person-year to the 30–34 year age and 1985–1989 calendar year stratum, then 4 person-years to the 35–39 and 1990–1994 stratum, and so on. The process of stratification of person-years can be extended to also include other factors in addition to age and calendar year (e.g., ethnicity, socioeconomic status, duration of follow-up, cumulative exposure) by apportioning person-years into a K -dimensional matrix, where K represents the number of stratification variables. The numbers of person-years in any cell of such a matrix will diminish as the number of variables and associated categories increase. Fortunately, there are sophisticated computer programs that can accommodate person-years counting for these more complex, yet typical, situations [11].

Stratum-specific incidence or mortality rates are computed by dividing the cohort's total number of cases (deaths) by the number of person-years. However, when studying rates for many different diseases among a cohort, the stratum-specific rates may be highly unstable because of small numbers of cases or person-years in individual strata, even when the cohort is relatively large.

Table 2 gives the general layout for data from a cohort study. This is the data layout for the *crude* table when there is no stratification made on age, calendar year, or any other potentially confounding variable. The same basic data layout is used for each stratum in a stratified data presentation; here, the cell entries (a , b , N_1 , N_0 , M , and T) would be indexed by subscripts such as i , j , k , etc., denoting levels of the various stratification factors. For example, if stratification were performed on age at 5 levels and calendar year at 3 levels, then there would

be $5 \times 3 = 15$ age/calendar year-specific strata. Age might be indexed by $i = 1, 2, 5$, and calendar year by $j = 1, 2, 3$. Thus, the data table for the second age-group and first calendar year stratum would be a_{21} , b_{21} , and so forth.

The standard measure of effect in cohort studies is the rate ratio, i.e., the ratio of the incidence rate (or mortality rate) in the exposed (study) population to that in the nonexposed (comparison) population. The two approaches that can be taken to summarize rates across strata of a potential confounder (e.g., age), while maintaining the unique information contained within strata (stratum-specific rates), are *standardization* and *pooling* [22].

The basic feature of a *standardized* rate (SR) is that it is a weighted average of stratum-specific rates. A general expression for an SR is given by

$$SR = \frac{\sum_i W_i I_i}{\sum_i W_i} \quad (1)$$

where i indexes the strata; the W_i are the stratum-specific weights; and the I_i are the stratum-specific incidence (or mortality) rates. The ratio of standardized rates (RR_s) is expressed as:

$$RR_s = \frac{\sum_i W_i (a_i / N_{1i})}{\sum_i W_i (b_i / N_{0i})} \quad (2)$$

Note that the numerator of expression (2) is the SR in the exposed population, and the denominator is the SR in the reference population.

In computing a standardized mortality ratio (SMR), the W_i are taken from the confounder distribution of the exposed population (i.e., $W_i = N_{1i}$), and expression (2) reduces to:

$$SMR = \frac{\sum_i a_i}{\sum_i N_{1i} (b_i / N_{0i})} \quad (3)$$

The SMR is thus the ratio of the sum of the observed cases in the exposed population, relative to the sum of the expected numbers in the exposed population, where the expected numbers are based on rates in the reference population.

Table 2 Data layout for a cohort study using rates and person-years

	Exposed	Nonexposed	Total
Cases	a	b	M
Person-years	N_1	N_0	T

Approximate confidence interval estimation for the SMR (CI_{SMR}) can be obtained from the following formula [22]:

$$CI_{SMR} = e^{\ln(SMR) \pm Z(SD(\ln(SMR)))} \quad (4)$$

where $SD(\ln(SMR))$ is the standard deviation of the natural logarithm of the SMR, defined as under:

$$SD(\ln(SMR)) = 1/Obs^{0.5} \quad (5)$$

Obs is the numerator of the SMR and Z is the standard normal deviate specifying the width of the confidence interval (e.g., $Z = 1.96$ for a 95% interval).

The alternative method is exemplified by the standardized rate ratio (SRR) [24]. One variant of this involves taking weights from the confounder distribution of the reference population (i.e., $W_i = N_{0i}$), in which case, expression (2) becomes

$$SRR = \frac{\sum_i N_{0i}(a_i/N_{1i})}{\sum_i b_i} \quad (6)$$

The SRR is therefore the ratio of the number of expected cases in the reference population, based on rates in the exposed group, to the number of observed cases in the reference population.

Another general method of summary rate ratio estimation is *pooling*, which involves computing a *weighted average of the stratum-specific rate ratios* (rather than the ratio of weighted averages of stratum-specific rates, as in expression (2)). In this case, RR_s takes the form:

$$RR_s = \frac{\sum_i W_i(a_i/N_{1i})/(b_i/N_{0i})}{\sum_i W_i} \quad (7)$$

The most common choice of weights is that given by the Mantel–Haenszel [25] method which uses weights of $b_i N_{1i}/T_i$ [22]. With these weights, expression (7) becomes:

$$RR_{M-H} = \frac{\sum_i a_i N_{0i}/T_i}{\sum_i b_i N_{1i}/T_i} \quad (8)$$

An approximate standard error for the natural log of the rate ratio is given by Greenland and Robins [26]

$$SE = \frac{[\sum M_{1i} Y_{1i} Y_{0i}/T_i^2]^{0.5}}{[(\sum a_i Y_{0i}/T_i)(\sum b_i Y_{1i}/T_i)]^{0.5}} \quad (9)$$

Thus, an approximate 95% confidence interval for the summary rate ratio is then given by

$$CI_{RR} = RRe^{\pm 1.96SE} \quad (10)$$

When the rate ratio is constant across all strata of the confounder, all three estimators (SMR, SRR, and RR_{M-H}) give the same result. This can be shown by assuming that the rate for the exposed group is equal to some multiple, M , of the rate in the reference population, i.e., $a_i/N_{1i} = M(b_i/N_{0i})$. Substitution into expression (2) (for the SMR or SRR) or expression (5) (for RR_{M-H}) yields: $RR_s = M$, in each instance.

The analytic methods based on stratified analyses that have been presented above are adequate for many occupational and environmental epidemiology studies, particularly those in which a simple exposed *versus* nonexposed classification is used. However, stratified analyses may not be feasible if there are multiple exposure categories or more than two or three confounders; it is then necessary to use multiple regression. In a cohort study, the appropriate model form is *Poisson regression* which is an extension of the simple analysis of rates (SMR, SRR, or RR_{M-H}) described above [27]. This is described in detail in standard texts such as Rothman and Greenland [22].

Strategies of Data Analysis

The first research question addressed in a cohort study is whether or not the rates of various diseases and injuries observed for the study cohort are different from rates found in a comparison population, which is presumed to be nonexposed to the agent(s) of concern. Table 3 summarizes the findings of a study by Wellmann *et al.* [10] of mortality in German carbon black workers employed for at least 1 year at a carbon black manufacturing plant, between 1960 and 1998. Follow-up was for the period 1976–1998. It shows that the cohort experienced an excess risk of mortality from all causes with 332 deaths observed, compared with 275.5 expected ($SMR = 1.20$, 95%

Table 3 Standardized mortality ratios (SMR) for all cause mortality and lung cancer mortality in German carbon black workers^(a)

Cause of death	Observed	Expected	SMR	95% CI
All causes	332	275.5	1.20	1.08–1.34
Lung cancer	50	23.0	2.18	1.61–2.87

^(a) Reproduced from [10]. © BMJ Publishing Group Ltd, 2006

CI 1.08–1.34); there was also a specific excess of lung cancer deaths, with 50 observed compared with 23.0 expected ($SMR = 2.18$, 95% CI 1.61–2.87).

Subcohort Analyses

The second level of cohort data analysis is a comparative examination of disease rates between subcohorts defined on the basis of their exposure experience, provided that job and/or exposure level data are available or can be reconstructed for the cohort. The simplest measure of exposure is the number of years of employment in the industry. An example of this approach, from the study of Wellmann *et al.* [10] is shown in Table 4. Person-years (and deaths) were accumulated by categories of years of employment, with study participants moving through the various categories over time (just as they are moved through age-groups and calendar periods). This involved an internal Poisson regression analysis in which those person-years and deaths which occurred in the 1–5 years of employment category were taken as the reference (there were no person-years accumulated for those with less than 1 year of employment as this was an inclusion criteria for the study).

The next level of analysis involves obtaining some information on actual exposures rather than simply the number of years that the study participants were exposed. In community-based studies, exposure estimates may be based on street address and length of residence, or may be based on personal exposure measurements. For example, Bertazzi *et al.* [28] studied the population of the Italian town of Seveso which was exposed to dioxin (see **Health Hazards Posed by Dioxin**) following a factory accident in 1976. Three contamination zones were delimited (Zones A, B, and R) based on their proximity to the plant and whether they fell under the natural fallout path of the chemical cloud. Thus, study

participants were categorized according to where they were living at the time of the accident, but this categorization was later confirmed by measurements of dioxin Tetrachlorodibenzo-*p*-dioxin (TCDD) concentrations in soil and in the blood of selected residents.

In occupational studies, characterization of workers into groups defined on the basis of process division in the plant or similarity of jobs and tasks is especially valuable when exposure intensity data are limited or nonexistent. The simplest approach is to categorize each worker into only one grouping, e.g., their longest held job. However, this can sacrifice information on other jobs/exposures. The solution is to make full use of all the person-time information available [11], by carrying out a separate analysis for each job category. The simplest such analysis is a comparison of the disease rates of workers who have ever worked in a particular job category with those of workers who have never worked in that category. Each analysis thus involves all of the person-time experience and cases (or deaths) in the study. An example of this approach, from the study of Wellmann *et al.* [10] is shown in Table 4. Separate analyses were conducted for employment (for at least 1 year) in “lamp or furnace black”, “gas black”, or the “preparation plant”; each of these analyses involved all of the deaths (and all of the person-years) in the cohort, and the data were classified according to whether each participant (at the time of the person-year in question) had accumulated, or not accumulated, at least 1 year of employment in the relevant job category. Once again, this involved an internal analysis in which those people who had not accumulated at least 1 year of employment for the particular job title (at the time of the person-year in question) were taken as the reference.

A refinement on this approach is to rank job categories with respect to exposure intensity for the agent(s) of interest [11] (or areas of residency which are similarly ranked, as in the case of the Seveso study [28]). *Ordinal rankings* can be made either on the basis of informed judgment or with reference to exposure measurements. Under such a scheme, jobs are grouped according to polychotomous ordinal rankings, such as “low”, “moderate”, and “high.” This is illustrated for the study by Wellmann *et al.* [10] in Table 4 for the category of “heavily exposed jobs”.

Table 4 Poisson regression analyses of all cause mortality and lung cancer mortality in German carbon black workers^(a)

Exposure category	All causes			Lung cancer		
	Deaths	RR	95% CI	Deaths	RR	95% CI
Years of employment						
1–5	82	1.00 ^(b)	0.53–1.11	12	1.00 ^(b)	0.40–2.74
5–10	45	0.76	0.56–10.9	9	1.05	0.34–2.09
10–20	96	0.78	0.43–0.85	16	0.84	0.33–2.81
>20	109	0.61	0.76–0.95	3	0.96	0.66–1.43
Lamp or furnace black						
No	206	1.00 ^(b)	–	28	1.00 ^(b)	–
Yes	126	1.14	0.92–1.43	22	1.48	0.84–2.61
Gas black ^(c)						
No	256	1.00 ^(b)	–	37	1.00 ^(b)	–
Yes	76	0.91	0.70–1.18	13	10.7	0.57–2.02
Preparation plant ^(c)						
No	311	1.00 ^(b)	–	46	1.00 ^(b)	–
Yes	21	0.79	0.51–1.23	4	1.05	0.37–2.95
Heavily exposed jobs ^(c)						
No	123	1.00 ^(b)	–	16	1.00 ^(b)	–
Yes	209	1.00	0.80–1.25	34	1.29	0.71–2.34
Average carbon black exposure						
0–1	73	1.00 ^(b)	–	12	1.00 ^(b)	–
1–2	75	0.92	0.67–1.27	14	1.06	0.49–2.28
2–3	140	1.00	0.75–1.33	21	0.97	0.47–1.98
>3	44	0.98	0.67–1.42	3	0.39	0.11–1.39
Cumulative carbon black exposure						
0–5	66	1.00 ^(b)	–	10	1.00 ^(b)	–
5–10	41	0.73	0.49–1.08	10	1.07	0.44–2.59
10–20	46	0.73	0.50–1.07	9	0.84	0.34–2.10
20–40	79	0.89	0.63–1.26	12	0.70	0.29–1.70
>40	100	0.61	0.43–0.85	9	0.28	0.11–0.72

^(a) Reproduced from [10]. © BMJ Publishing Group Ltd, 2006^(b) Reference category^(c) At least 1 year's employment

Cumulative Exposure

When quantified exposure data are available for all or most workers, either by means of linkage of job and work area data to employment history information or, less often, from personal monitoring, it will be possible to define subcohorts on the basis of maximum exposure intensity or with respect to cumulative exposure levels. Many occupational diseases are strongly related to *cumulative exposure* which is simply the summed products of exposure

intensities and their associated durations. Typical examples of cumulative exposure in occupational studies are milligrams of dust per cubic meter $\times$ years, fibers per milliliter $\times$ years, and parts per million $\times$ years for chemical exposures. In a community-based study, cumulative exposure might be based on estimated exposure levels (e.g., levels of arsenic in well water) $\times$ years of residency.

An occupational example of this approach, from the study of Wellmann *et al.* [10] is shown in Table 4. Study participants were classified according

to their average annual carbon black exposure, and this was combined with information on years of employment to yield estimates of cumulative carbon black exposure. Once again, this involved an internal analysis in which study participants were moved through exposure categories over time, and those person-years (and deaths) which occurred in the category of 0–5 units of cumulative carbon black exposure were taken as the reference. The analyses indicated that, although there were overall excesses of all causes mortality and lung cancer mortality in this cohort, these were not associated with average or cumulative levels of carbon black exposure.

Latency and Induction Time

For all diseases or injuries, there are requisite time lags between exposure to etiologic factors and clinical manifestation. In the extreme case of an acute episode, such as an electrocution following exposure to high voltage or eye irritation from a noxious gas, the time lag is effectively nil because the response is virtually instantaneous. However, many occupationally related diseases are delayed effects of exposure, where the time lag may range from hours or days (e.g., pulmonary edema) to decades (e.g., cancer).

Latency is the term often applied to the time interval between exposure and disease manifestation or recognition. However, Rothman [29] defines latency as the period from disease initiation to manifestation. This definition is consistent with the idea that there is some period of time when the disease exists in a “hidden” state in an individual. A related concept is *induction time*, which can be defined as the period

from first exposure to an agent or collection of agents to disease initiation [29]. The induction time will of course vary, depending on the nature of the exposure and disease, as well as on individual responses to exposure. Many occupational and environmental exposures can be considered to be chronic, extending over periods of years or decades, but may also be intermittent within the period of exposure. Intermittent exposures, while relevant for induction of some diseases, will not be discussed here so as to simplify matters. Thus, while we cannot usually estimate induction time, we can determine the interval from exposure onset to disease manifestation or recognition. Rothman [29] refers to this interval, which is the sum of induction and latency times, as the *empirical induction time*.

Several methods for taking into account induction and latency have been applied [11]. The most valid and appropriate technique is to *lag exposures* by some assumed latency interval (y), such that a worker's current person-year at risk is assigned to the exposure level (either intensity, duration, or most commonly, cumulative exposure) achieved y years earlier [30, 31]. Consider an example using cumulative exposure as the exposure index. If by age 40 a worker had accumulated 25 cumulative exposure units, and by age 50 his cumulative exposure was 45 units, then under an assumed latency interval of 10 years, his person-year at risk for age 50 would be assigned to the 25 cumulative exposure unit level. Exposure lagging has the particular advantage of including the entire enumerated cohort and all members' complete exposure histories, unlike the other

Table 5 Poisson regression analyses of cumulative occupational arsenic exposure and circulatory disease mortality, by 0, 10, and 20-year lag periods^(a)

Cumulative arsenic exposure ($\mu\text{g As}/\text{m}^3\text{-years}$)	0-year lag		10-year lag		20-year lag	
	<i>RR</i>	95% confidence interval	<i>RR</i>	95% confidence interval	<i>RR</i>	95% confidence interval
<750 ^(b)	1.0 ^(b)	—	1.0 ^(b)	—	1.0 ^(b)	—
750–1999	0.9	0.6–1.3	1.1	0.8–1.5	1.0	0.7–1.3
2000–3999	0.9	0.6–1.3	1.2	0.9–1.6	1.2	0.9–1.6
4000–7999	1.0	0.7–1.4	1.1	0.8–1.6	1.3	1.0–1.8
8000–19 999	1.0	0.7–1.4	1.5	1.0–2.0	1.6	1.2–2.1
>20 000	1.0	0.7–1.5	1.3	0.9–2.0	1.5	1.0–2.3

^(a) Reproduced from [32]. © Oxford University Press, 2000

^(b) Reference category

methods described above which sacrifice information either by subject exclusion or exposure truncation.

Table 5 gives an example of exposure lagging from a study by Hertz-Picciotto *et al.* [32] of circulatory disease mortality in a cohort of 2082 white males employed for 1 year or more in a copper smelter in Tacoma, Washington during 1960–1964. Follow-up covered the period 1940–1976. The use of lagged exposures increased the relative risks of circulatory disease mortality for all levels of exposure compared with the baseline level, supporting the hypothesis that cumulative arsenic exposure increases circulatory disease mortality.

An extension of the lagging approach to latency analysis is to estimate effects related to exposures that occur during discrete-time periods, or *time windows* [31]. Thus, for example, it might be considered that, for some particular health outcome, both the most recent and most distant exposures were probably less related than exposures during a critical time window. A good illustration would be exposure effects that are linked with birth defects during particular periods of fetal development [33]. The analysis of time windows is performed in a similar manner as lagging, except exposures during the assumed etiologically unrelated recent and distant periods are discounted. The analysis should maintain the temporal relation with follow-up, such that person-years are distributed appropriately to the moving time window of interest [34].

Discussion

We have described methods for assembling and following community-based and occupational cohorts, and for making disease rate comparisons with external and internal reference populations. These comparisons permit evaluations of the cohort's overall disease experience and associations with various exposures. What may not be evident from the discussion is the scope of work required to conduct an occupational or environmental cohort study successfully. These studies typically involve the efforts of epidemiologists, environmental scientists, occupational hygienists, biostatisticians, computer programmers, and clerical staff. The magnitude of the effort will depend on the size of the cohort studied and on the volume and complexity of the exposure data to be assimilated. For example, vital status tracing and

cause of death ascertainment for a cohort of 5000 people is seldom completed in less than 1 year, and a thorough reconstruction of exposures and analysis of data might require another 1–2 years. These comments pertain primarily to historical cohort studies. Prospective cohort studies are usually substantially more expensive and time consuming.

There are, however, particular advantages to cohort studies. First, by design, they include all or as many members of the dynamic cohort as can be identified and traced. There is a statistical precision advantage to this approach, but more importantly, cohort studies offer the broadest available picture of the health experience of a workforce because rates for multiple health outcomes can be examined. More intensive investigations of exposure–response associations are best done for a selected subset of health outcomes of *a priori* interest and those discovered to occur at excessive rates in the cohort.

The second, less obvious, advantage of cohort studies is that the process of enumerating a cohort makes the investigator more aware of the particular characteristics of the population under study, e.g., the ethnic, gender, and social class groups most heavily represented, and the subgroups of the cohort most completely traced. Other study designs that include only samples of the cohort do not offer as complete an assessment of the exposure and health experience of a population. Finally, phenomena characteristic of occupational populations, such as bias resulting from the healthy worker effect, are examined most directly in cohort studies.

Acknowledgments

Funding for Neil Pearce's salary is from a Programme Grant from the Health Research Council of New Zealand. Harvey Checkoway contributed to this manuscript during a Visiting Scientist Fellowship at the International Agency for Research on Cancer.

References

- [1] Grandjean, P., Weihe, P., White, R.F. & Debes, F. (1998). Cognitive performance of children prenatally exposed to "safe" levels of methylmercury, *Environmental Research* 77, 165–172.
- [2] Magnani, C., Terracini, B., Ivaldi, C., Botta, M., Mancini, A. & Andrion, A. (1995). Pleural malignant mesothelioma and nonoccupational exposure to asbestos

- in Casale-Monferrato, Italy, *Occupational and Environmental Medicine* **52**, 362–367.
- [3] Mannelte, A.T., McLean, D., Cheng, S., Boffetta, P., Colin, D. & Pearce, N. (2005). Mortality in New Zealand workers exposed to phenoxy herbicides and dioxins, *Occupational and Environmental Medicine* **62**, 34–40.
- [4] Kogevinas, M., Becher, H., Benn, T., Bertazzi, P.A., Boffetta, P., Bueno-de-Mesquita, H.B., Coggon, D., Colin, D., Flesch-Janys, D., Fingerhut, M., Green, L., Kauppinen, T., Littorin, M., Lynge, E., Mathews, J.D., Neuberger, M., Pearce, N. & Saracci, R. (1997). Cancer mortality in workers exposed to phenoxy herbicides, chlorophenols, and dioxins. An expanded and updated international cohort study [comment], *American Journal of Epidemiology* **145**, 1061–1075.
- [5] Saracci, R., Kogevinas, M., Bertazzi, P.A., Bueno de Mesquita, B.H., Coggon, D., Green, L.M., Kauppinen, T., L'Abbe, K.A., Littorin, M., Lynge, E., Mathews, J.D., Neuberger, M., Osman, J., Pearce, N.E. & Winkelmaam, R. (1991). Cancer mortality in workers exposed to chlorophenoxy herbicides and chlorophenols, *Lancet* **338**, 1027–1032.
- [6] Ott, M.G. & Zober, A. (1996). Cause specific mortality and cancer incidence among employees exposed to 2,3,7,8-TCDD after a reactor accident, *Occupational and Environmental Medicine* **53**, 606–612.
- [7] Preston, D.L., Kato, H., Kopecky, K.J. & Fujita, S. (1987). Studies of the mortality of a-bomb survivors. 8. cancer mortality, 1950–1982, *Radiation Research* **111**, 151–178.
- [8] Izumi, S., Suyama, A. & Koyama, K. (2003). Radiation-related mortality among offspring of atomic bomb survivors: a half-century of follow-up, *International Journal of Cancer* **107**, 292–297.
- [9] Smith, A.H., Marshall, G., Yuan, Y., Ferreccio, C., Liaw, J., von Ehrenstein, O. & Steinmaus, C. (2005). Childhood exposure to arsenic in water in Chile and increased mortality from chronic pulmonary disease, *Epidemiology* **16**, S120.
- [10] Wellmann, J., Weiland, S.K., Neiteler, G., Klein, G. & Straif, K. (2006). Cancer mortality in German carbon black workers 1976–1998, *Occupational and Environmental Medicine* **63**, 513–521.
- [11] Checkoway, H., Pearce, N. & Kriebel, D. (2004). *Research Methods in Occupational Epidemiology*, 2nd Edition, Oxford University Press, New York.
- [12] Hernberg, S., Partanen, T., Nordman, C.-H. & Sumari, P. (1970). Coronary heart disease among workers exposed to carbon disulfide, *British Journal of Industrial Medicine* **27**, 313–325.
- [13] Gilbert, E.S. (1982). Some confounding factors in the study of mortality and occupational exposure, *American Journal of Epidemiology* **116**, 177–188.
- [14] Pearce, N., Checkoway, H. & Shy, C. (1986). Time-related factors as potential confounders and effect modifiers in studies based on an occupational cohort, *Scandinavian Journal of Work Environment and Health* **12**, 97–107.
- [15] Steenland, K. & Stayner, L. (1991). The importance of employment status in occupational cohort mortality studies, *Epidemiology* **2**, 418–423.
- [16] Collins, J.F. & Redmond, C.K. (1976). The use of retirees to evaluate occupational hazards, *Journal of Occupational Medicine* **18**, 595–602.
- [17] Boffetta, P., Sali, D., Kolstad, H., Coggon, D., Olsen, J., Andersen, A., Spence, A., Pesatori, A.C., Lynge, E., Frentzel-Beyme, R., Chnag-Claude, R.J., Lundberg, I., Biocca, M., Gennaro, V., Teppo, L., Partanen, T., Welp, E., Saracci, R. & Kogevinas, M. (1998). Mortality of short-term workers in two international cohorts, *Journal of Occupational and Environmental Medicine* **40**, 1120–1126.
- [18] Rushton, L. (1994). Use of multiple cause of death in the analysis of occupational cohorts – an example from the oil industry, *Occupational and Environmental Medicine* **51**, 722–729.
- [19] Steenland, K., Nowlin, S., Ryan, B. & Adams, S. (1992). Use of multiple-cause mortality data in epidemiologic analyses: US rate and proportion files developed by the National Institute for Occupational Safety and Health and the National Cancer Institute, *American Journal of Epidemiology* **136**, 855–862.
- [20] Parkin, D.M., Whelan, S.L., Ferlay, J., Teppo, L. & Thomas, D.B. (eds) (2003). *Cancer Incidence in Five Continents*, IARC, Lyon, Vol. VIII.
- [21] Ries, L.A.G., Kosary, C.L. & Hankey, B.F. (eds) (1998). *SEER Cancer Statistics Review, 1973–1995*, US National Cancer Institute, Bethesda.
- [22] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Edition, Lippincott Williams & Wilkins, Philadelphia.
- [23] Pearce, N. (1992). Methodological problems of time-related variables in occupational cohort studies, *Revue d'Epidemiologie et de Sante Publique* **40**, S43–S54.
- [24] Miettinen, O.S. (1972). Standardisation of risk ratios, *American Journal of Epidemiology* **96**, 383–388.
- [25] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.
- [26] Greenland, S. & Robins, J.M. (1985). Estimation of a common effect parameter from sparse follow-up data, *Biometrics* **41**, 55–68.
- [27] Pearce, N., Checkoway, H. & Dement, J. (1988). Exponential models for analyses of time-related factors, illustrated with asbestos textile worker mortality data, *Journal of Occupational Medicine* **30**, 517–522.
- [28] Bertazzi, P.A., Consonni, D., Bachetti, S., Rubagotti, M., Baccarelli, A., Zocchetti, C. & Pesatori, A.C. (2001). Health effects of dioxin exposure: a 20-year mortality study, *American Journal of Epidemiology* **153**, 1031–1044.
- [29] Rothman, K.J. (1981). Induction and latent periods, *American Journal of Epidemiology* **114**, 253–259.
- [30] Gilbert, E.S. & Marks, S. (1979). An analysis of the mortality of workers in a nuclear facility, *Radiation Research* **79**, 122–148.

- [31] Checkoway, H., Pearce, N., Hickey, J.L. & Dement, J.M. (1990). Latency analysis in occupational epidemiology, *Archives of Environmental Health* **45**, 95–100.
- [32] Hertz-Picciotto, I., Arrighi, H.M. & Hu, S.W. (2000). Does arsenic exposure increase the risk for circulatory disease? *American Journal of Epidemiology* **151**, 174–181.
- [33] Sever, L.E. & Mortensen, M.E. (1996). Teratology and the epidemiology of birth defects: occupational and environmental perspectives, in *Obstetrics: Normal and Problem Pregnancies*, S.G. Gabbe, J.R. Niebyl & J.L. Simpson, eds, Churchill Livingstone, New York, pp. 185–213.
- [34] Pearce, N. (1988). Multistage modelling of lung cancer mortality in asbestos textile workers, *International Journal of Epidemiology* **17**, 747–752.

Related Articles

Case–Control Studies

Environmental Health Risk

NEIL PEARCE AND HARVEY CHECKOWAY

Collective Risk Models

Individual and collective risk models in actuarial science are often referred to as *models for aggregate claims* (see **Comonotonicity; Longevity Risk and Life Annuities; Reinsurance**) that arise from an insurance portfolio over a fixed period of time. If the insurance portfolio is closed, that is, the number of policies in the portfolio is fixed and no replacement is allowed, then the aggregate claims may be expressed as

$$S_n = X_1 + X_2 + \cdots + X_n \quad (1)$$

where n is the number of policies, and for $i = 1, 2, \dots, n$, X_i is the claim amount over the period from the i th policy. If there is no claim, $X_i = 0$. It is normally assumed that the claim amounts $X_1, X_2, \dots, X_n$ are mutually independent but they are not identically distributed. The above model is referred to as the *individual risk model*. The evaluation of the model involves the use of convolution techniques (see **Ruin Probabilities; Computational Aspects; Operational Risk Modeling**). In contrast with the individual risk model, the collective risk model does not make direct reference to the risk characteristics of the individual members of the portfolio when describing the aggregate claims experience of the whole portfolio itself. To model the total claims of a collection of risks over a fixed (usually prospective) period of time, the collective approach incorporates both claim frequency and claim severity components into the probability distribution of the aggregate.

To substantiate, let N be a counting random variable representing the number of claims experienced by the portfolio over the predetermined time period, and let Y_i be a random variable representing the amount of the i th claim to occur within this same period. Then the total or aggregate claims over the period is given by the random sum S where

$$S = Y_1 + Y_2 + \cdots + Y_N \quad (2)$$

with $S = 0$ if $N = 0$. For mathematical tractability, it is normally assumed that $\{Y_1, Y_2, \dots\}$ is a sequence of independent and identically distributed (iid) non-negative random variables independent of N . This assumption implicitly requires (among other things) that the time period of interest be short enough so

that temporal variation in claims experience may be ignored. By conditioning on N , it follows from equation (2) that the expected aggregate claim amount is given by

$$E(S) = E(N) E(Y) \quad (3)$$

where Y is a generic single claim amount random variable. Clearly, equation (3) is intuitively very reasonable, and the analogous result for the variance is

$$\text{Var}(S) = E(N) \text{Var}(Y) + \text{Var}(N) \{E(Y)\}^2 \quad (4)$$

For risk management and other technical purposes (e.g., stop-loss analysis) (see **Credit Risk Models; Optimal Stopping and Dynamic Programming**), it is of interest to evaluate the distribution of S . Again, conditioning on the number of claims, it follows that

$$\bar{G}(x) = \sum_{n=1}^{\infty} p_n \bar{F}^{*n}(x), \quad x \geq 0 \quad (5)$$

where $G(x) = 1 - \bar{G}(x) = \Pr(S \leq x)$, $p_n = \Pr(N = n)$, and $\bar{F}^{*n}(x) = \Pr(\sum_{i=1}^n Y_i > x)$. As an analytic tool, the Laplace–Stieltjes transform (LST) or probability generating function (pgf) is quite useful in this situation. Let

$$P(z) = E(z^N) = \sum_{n=0}^{\infty} p_n z^n \quad (6)$$

be the pgf of N , and the LST of S is given by

$$E(e^{-zS}) = P(E(e^{-zY})) \quad (7)$$

whereas (in the discrete case), the pgf of S is given by

$$E(z^S) = P(E(z^Y)) \quad (8)$$

In general, evaluation of equation (5) is quite complicated, but for some choices of $\{p_0, p_1, \dots\}$ and/or $F(x) = 1 - \bar{F}(x) = 1 - \bar{F}^{*1}(x)$, it may be simplified. If $F(x)$ is the distribution function (df) of a gamma or inverse Gaussian random variable, then $\bar{F}^{*n}(x)$ admits a simple form. In particular, if $F(x) = 1 - e^{-\mu x}$, then $\bar{F}^{*n}(x) = e^{-\mu x} \sum_{j=0}^{n-1} \frac{(\mu x)^j}{j!}$ and equation (5) becomes

$$\bar{G}(x) = e^{-\mu x} \sum_{j=0}^{\infty} \bar{P}_j \frac{(\mu x)^j}{j!}, \quad x \geq 0 \quad (9)$$

with $\bar{P}_j = \sum_{n=j+1}^{\infty} p_n$. In this case $E(e^{-zY}) = \mu/(\mu + z)$ and equation (7) becomes

$$E(e^{-zS}) = P\left(\frac{\mu}{\mu + z}\right) \quad (10)$$

Furthermore, if N has a zero-modified geometric distribution with $p_n = (1 - p_0)(1 - \phi)\phi^{n-1}$ for $n = 1, 2, \dots$, then $\bar{P}_j = (1 - p_0)\phi^j$ and equation (5) becomes $\bar{G}(x) = (1 - p_0)e^{-\mu(1-\phi)x}$. Suppose, more generally, that Y has a mixed Erlang distribution with $F(x) = 1 - e^{-\mu x} \sum_{k=1}^{\infty} q_k \sum_{j=0}^{k-1} (\mu x)^j / j!$ where $\{q_1, q_2, \dots\}$ is a counting measure. This df may approximate any continuous distribution on $(0, \infty)$ arbitrarily accurately (cf. [1, p. 163]), and is very tractable mathematically (e.g., [2]). One main reason for this mathematical tractability is that $E(e^{-zY}) = Q(\mu/(\mu + z))$ where $Q(z) = \sum_{k=1}^{\infty} q_k z^k$, and thus equation (7) becomes $E(e^{-zS}) = C(\mu/(\mu + z))$ with $C(z) = P(Q(z))$, which is of the same form as equation (10) in the exponential claims case, but with $P(z)$ replaced by $C(z) = P(Q(z))$. Conveniently, $C(z)$ is itself of the discrete compound from equation (8).

Another family of distributions for which explicit computational techniques are available in this situation is the class of phase-type distributions (see **Ruin Probabilities: Computational Aspects; Markov Modeling in Reliability**). See Asmussen and Rolski [3] and references therein for details.

For both theoretical and computational reasons, models for $\{p_0, p_1, \dots\}$ are typically chosen to be of parametric form. In this regard the Poisson distribution with $p_n = \lambda^n e^{-\lambda} / n!$ and $P(z) = \exp\{\lambda(z - 1)\}$ is of central importance. This assumption follows from modeling the stochastic process of the number of claims as a Poisson process. In the compound Poisson case, equation (7) becomes

$$E(e^{-zS}) = e^{\lambda\{E(e^{-zY}) - 1\}} \quad (11)$$

implying that the compound Poisson distribution is infinitely divisible (e.g., [4]). It is instructive to note that infinite divisibility, an important concept in probability theory, is also important in claims modeling in connection with business growth [5, p. 108]. Furthermore, the compound Poisson distribution is itself of central importance in the theory of infinite divisibility.

Compound Poisson distributions are closed under convolution. Thus, if S_i has a compound Poisson distribution with Poisson parameter λ_i and claims

size df $F_i(y)$ for $i = 1, 2, \dots, n$, and the S_i are independent, it follows that $S_1 + S_2 + \dots + S_n$ is again compound Poisson with Poisson parameter $\lambda = \lambda_1 + \lambda_2 + \dots + \lambda_n$ and “claim size” df $F(y) = \{\lambda_1 F_1(y) + \lambda_2 F_2(y) + \dots + \lambda_n F_n(y)\} / \lambda$. This closure property is itself the basis for approximation of the individual risk model, as now described.

Let the portfolio consist of n_{ij} individuals insured for amount i where $i \in \{1, 2, \dots, r\}$ with claim probability q_j where $j \in \{1, 2, \dots, m\}$. If $n_{ij} > 0$, then N_{ijk} is the number of claims for the k th of the n_{ij} individuals for $k \in \{1, 2, \dots, n_{ij}\}$. The total claims are thus $S = \sum_{i=1}^r \sum_{j=1}^m \sum_{k=1}^{n_{ij}} i N_{ijk}$ where $\sum_{k=1}^0 = 0$. Let N_{ijk} have pgf $P_j(z) = E(z^{N_{ijk}})$, and assuming independence of the N_{ijk} , it follows that S has pgf

$$E(z^S) = \prod_{i=1}^r \prod_{j=1}^m \{P_j(z^i)\}^{n_{ij}} \quad (12)$$

which holds even when $n_{ij} = 0$. For the individual risk model, $N_{ijk} \sim \text{Bin}(1, q_j)$, so that $P_j(z) = 1 - q_j + q_j z$. The compound Poisson approximation to the individual risk model involves the approximating assumption that N_{ijk} is Poisson distributed with mean $\lambda_j = \lambda_j(q_j)$. Thus, with $P_j(z) \approx e^{\lambda_j(z-1)}$, equation (12) yields

$$E(z^S) \approx \prod_{i=1}^r \prod_{j=1}^m e^{n_{ij}\lambda_j(z^i-1)} = \prod_{i=1}^r e^{\lambda_i(z^i-1)} \quad (13)$$

where $\lambda_i = \sum_{j=1}^m n_{ij}\lambda_j$. In turn,

$$E(z^S) \approx e^{\lambda\{E(z^Y)-1\}} \quad (14)$$

with $\lambda = \sum_{i=1}^r \lambda_i = \sum_{i=1}^r \sum_{j=1}^m n_{ij}\lambda_j$ and

$$\Pr(Y = i) = \frac{\lambda_i}{\lambda} = \frac{\sum_{j=1}^m n_{ij}\lambda_j}{\sum_{k=1}^r \sum_{j=1}^m n_{kj}\lambda_j}, \quad i = 1, 2, \dots, r \quad (15)$$

If the means are matched then $\lambda_j = q_j$, whereas if the probabilities at 0 are matched then $\lambda_j = -\ln(1 - q_j) > q_j$.

The compound Poisson distribution has many attractive properties, and in particular lends itself to

a simple recursive computational formula, commonly referred to as *Panjer's recursion* [6]. Thus, if

$$E(z^S) = e^{\lambda \left\{ \sum_{y=0}^{\infty} f_y z^y - 1 \right\}} \quad (16)$$

where $\{f_0, f_1, \dots\}$ is a counting measure, then the probability mass function of S may be computed recursively from the formula

$$\Pr(S = x) = \frac{\lambda}{x} \sum_{y=1}^x y f_y \Pr(S = x - y),$$

$$x = 1, 2, \dots \quad (17)$$

beginning with $\Pr(S = 0) = e^{-\lambda(1-f_0)}$. If Y has an absolutely continuous distribution, the distribution of S may be approximated arbitrarily accurately using equation (17) after suitably discretizing the distribution of Y (e.g., [5], Chapter 6).

For some applications (e.g., some lines of automobile insurance), the Poisson assumption may be inappropriate as a result of poor fit to count data (often in the right tail of the distribution), and alternatives may be considered. A computationally attractive family of counting distributions is that for which

$$p_n = \left(a + \frac{b}{n}\right) p_{n-1}; \quad n = 2, 3, 4, \dots \quad (18)$$

where a and b are model parameters not depending on n .

Members of this family, introduced by Sundt and Jewell [7], include the Poisson, negative binomial, binomial, logarithmic series, and extended truncated negative binomial with pgf

$$P(z) = \frac{\{1 - \beta(z - 1)\}^{-r} - (1 + \beta)^{-r}}{1 - (1 + \beta)^{-r}} \quad (19)$$

where $\beta > 0$ and $-1 < r < 0$. Since equation (18) does not involve p_0 , zero-modified versions of these distributions also satisfy equation (18). The ability to select p_0 arbitrarily greatly increases the range of applicability of these models to situations involving various shapes of distributions. See Johnson *et al.* [8] for further discussion of these discrete distributions. Also, from equation (19), $\lim_{\beta \rightarrow \infty} P(z) = 1 - (1 - z)^{-r}$, which is the pgf of the Sibuya distribution, and no distributions satisfy equation (18) other than zero-modified version of those mentioned in this paragraph [9]. See also Hess *et al.* [10].

If equation (18) holds and $E(z^S) = \sum_{n=0}^{\infty} p_n \left(\sum_{y=0}^{\infty} f_y z^y\right)^n$, then equation (17) generalizes for $x = 1, 2, \dots$, to the recursive formula

$$\Pr(S = x) = \frac{\{p_1 - (a + b)p_0\} f_x + \sum_{y=1}^x \left(a + b \frac{y}{x}\right) f_y \Pr(S = x - y)}{1 - a f_0} \quad (20)$$

beginning with $\Pr(S = 0) = P(f_0) = \sum_{n=0}^{\infty} p_n f_0^n$. For further details on the recursive technique, see Sundt [11].

Another extremely important family of distributions in claim count modeling is the family of mixed Poisson distributions with

$$p_n = \int_0^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} dB(\lambda); \quad n = 0, 1, 2, \dots \quad (21)$$

where $B(\lambda)$ is a df on $(0, \infty)$. The pgf associated with equation (21) is

$$P(z) = \int_0^{\infty} e^{\lambda(z-1)} dB(\lambda) = \tilde{b}(1 - z) \quad (22)$$

where $\tilde{b}(s) = \int_0^{\infty} e^{-s\lambda} dB(\lambda)$ is the LST associated with $B(\lambda)$. This importance is a result of both theoretical and practical considerations. From a theoretical viewpoint, heterogeneity in the Poisson claim rate in the portfolio of risks as described by the “structural” df $B(\lambda)$ has much appeal, and mixed Poisson distributions both fit well to many types of claims data and are computationally tractable. What is of central importance in this family is the negative binomial distribution with $B(\lambda)$ as a gamma df, but the Poisson-inverse Gaussian distribution with $B(\lambda)$ as an inverse Gaussian df is a convenient, heavily skewed alternative for which recursive techniques are applicable. More generally, Sichel’s distribution results when $B(\lambda)$ is a generalized inverse Gaussian df. See Grandell [12] for a thorough discussion of mixed Poisson distributions and processes.

Many other parametric models for N have been proposed and are discussed by Johnson *et al.* [8]. The use of recursive techniques with parametric claim count and claim size models as a method for handling deductibles and policy maximums may be found in

Klugman *et al.* [5]. A more advanced discussion is given by Rolski *et al.* [13].

There are various other approaches to evaluation of the total claims distribution (*see Enterprise Risk Management (ERM); Equity-Linked Life Insurance; Nonlife Loss Reserving*), numerical inversion of the LST equation (7), the fast Fourier transform in the discrete case (8), and simulation have also been used in various situations. For larger portfolios, parametric approximations such as the normal distribution based on central limit theoretic arguments may also be used, but the normal distribution itself typically understates the right tail of the aggregate distribution (*see Risk Measures and Economic Capital for (Re)insurers; Simulation in Risk Management; Extreme Value Theory in Finance*). Others include gamma, shifted or translated gamma, Haldane, Wilson–Hilfarty, normal power, and Esscher approximations. See Pentikainen [14] and references therein for further discussion of these methods.

There are also asymptotic formulas available for the extreme right tail, which are well suited for use in conjunction with recursive techniques. In the following text, the notation $a(x) \sim b(x)$, $x \rightarrow \infty$, is used to denote that $\lim_{x \rightarrow \infty} a(x)/b(x) = 1$. First, we assume that

$$p_n \sim Cn^\alpha \theta^n, \quad n \rightarrow \infty \quad (23)$$

where $C > 0$, $-\infty < \alpha < \infty$, and $0 < \theta < 1$. Embrechts *et al.* [15] showed that if there exists $\kappa > 0$ satisfying $E(e^{\kappa Y}) = 1/\theta$ and the distribution of Y is not arithmetic, then

$$\bar{G}(x) \sim \frac{Cx^\alpha e^{-\kappa x}}{\kappa \{\theta E(Ye^{\kappa Y})\}^{\alpha+1}}, \quad x \rightarrow \infty \quad (24)$$

In fact, C may be replaced by a function that varies slowly at infinity. The estimate (24) is often relatively accurate even for small x . A similar result holds if Y has a counting distribution.

An alternative formula to equation (24) often holds in cases where Y has no moment generating function and hence no $\kappa > 0$ may be found. If F is a subexponential df (examples include the Pareto types with regularly varying tails and the lognormal) and the pgf (6) has radius of convergence exceeding 1, then

$$\bar{G}(x) \sim E(N)\bar{F}(x), \quad x \rightarrow \infty \quad (25)$$

The resulting equation (25) is classical and is given by Embrechts *et al.* [16, p. 580], for example. Unfortunately, it is of limited practical use because typically both sides of equation (25) are extremely close to 0 before it becomes accurate. A related asymptotic formula holds when Y has a distribution from a class that includes the (generalized) inverse Gaussian distribution (e.g., [17], and references therein).

Lundberg-type bounds, which often complement the asymptotic formulas, are discussed by Rolski *et al.* [13, Section 4.5] and Willmot and Lin [18, Chapters 4–6].

Finally, we remark that the iid assumption typically made with respect to the claim size sequence $\{Y_1, Y_2, \dots\}$ may sometimes be relaxed. For example, Denuit *et al.* [19] and references therein, consider dependency. Also, the identically distributed assumption may be relaxed with little additional complexity in the case with mixed Poisson claims, a useful approach when incorporating inflation into the model (e.g., [18, Section 4.1]).

References

- [1] Tijms, H. (1994). *Stochastic Models—An Algorithmic Approach*, John Wiley & Sons, Chichester.
- [2] Willmot, G. & Woo, J. (2007). On the class of Erlang mixtures with risk theoretic applications, *North American Actuarial Journal* **11**(2), 99–118.
- [3] Asmussen, S. & Rolski, T. (1991). Computational methods in risk theory: a matrix-algorithmic approach, *Insurance: Mathematics and Economics* **10**, 259–274.
- [4] Steutel, F. & Van Harn, K. (2004). *Infinite Divisibility of Probability Distributions on the Real Line*, Marcel Dekker, New York.
- [5] Klugman, S., Panjer, H. & Willmot, G. (2004). *Loss Models—From Data to Decisions*, 2nd Edition, John Wiley & Sons, New York.
- [6] Panjer, H. (1981). Recursive evaluation of a family of compound distributions, *Astin Bulletin* **12**, 22–26.
- [7] Sundt, B. & Jewell, W. (1981). Further results on recursive evaluation of compound distribution, *Astin Bulletin* **12**, 27–39.
- [8] Johnson, N., Kemp, A. & Kotz, S. (2005). *Univariate Discrete Distributions*, 3rd Edition, John Wiley & Sons, New York.
- [9] Willmot, G. (1988). Sundt and Jewell’s family of discrete distributions, *Astin Bulletin* **18**, 17–29.
- [10] Hess, K., Liewald, A. & Schmidt, K. (2002). An extension of Panjer’s recursion, *Astin Bulletin* **32**, 283–297.
- [11] Sundt, B. (2002). Recursive evaluation of aggregate claims distributions, *Insurance: Mathematics and Economics* **30**, 297–322.

-
- [12] Grandell, J. (1997). *Mixed Poisson Processes*, Chapman & Hall, London.
- [13] Rolski, T., Schmidli, H., Schmidt, V. & Teugels, J. (1999). *Stochastic Processes for Insurance and Finance*, John Wiley & Sons, Chichester.
- [14] Pentikainen, T. (1987). Approximate evaluation of the distribution function of aggregate claims, *Astin Bulletin* **17**, 15–39.
- [15] Embrechts, P., Maejima, M. & Teugels, J. (1985). Asymptotic behaviour of compound distributions, *Astin Bulletin* **15**, 45–48.
- [16] Embrechts, P., Kluppelberg, C. & Mikosch, T. (1999). *Modelling Extremal Events – for Insurance and Finance*, Springer-Verlag, Berlin.
- [17] Embrechts, P. (1983). A property of the generalized inverse Gaussian distribution with some applications, *Journal of Applied Probability* **20**, 537–544.
- [18] Willmot, G. & Lin, X.S. (2001). *Lundberg Approximations for Compound Distributions with Insurance Applications*, Springer-Verlag, New York.
- [19] Denuit, M., Dhaene, J., Goovaerts, M. & Kaas, R. (2005). *Actuarial Theory for Dependent Risks – Measures, Orders, and Models*, John Wiley & Sons, Chichester.
- X. SHELDON LIN AND GORDON E. WILLMOT

Combining Information

Introduction: Meta-Analysis

An area of growing importance in risk assessment involves the combination of information from diverse sources relating to a similar end point. A common rubric for combining the results of independent studies is to apply a *meta-analysis*. The term suggests a move to incorporate and *synthesize* information from many associated sources; it was first coined by Glass [1] in an application of combining results across multiple social science studies. Within the risk analytic paradigm, the possible goals of a meta-analysis include consolidation of results from independent studies, improved analytic sensitivity to detect the presence of adverse effects, and construction of valid inferences on the phenomenon of interest based on the combined data. The result is often a pooled estimate of the overall effect. For example, it is increasingly difficult for a single, large, well-designed ecological or toxicological study to definitively assess the risk(s) of exposure to some chemical hazard (*see Ecological Risk Assessment*). Rather, many small studies may be carried out from which quantitative strategies that can synthesize the independent information into a single, well-understood inference will be of great value. To do so, one must generally assume that the different studies are considering equivalent endpoints, and that data derived from them will provide *exchangeable* information when consolidated. Formally, the following assumptions should be satisfied:

1. All studies/investigations meet basic scientific standards of quality (proper data reporting/collecting, random sampling, avoidance of bias, appropriate ethical considerations, fulfilling quality assurance (QA) or quality control (QC) guidelines, etc.).
2. All studies provide results on the same quantitative outcome.
3. All studies operate under (essentially) the same conditions.
4. The underlying effect is a fixed effect; i.e., it is nonstochastic and *homogeneous* across all studies. (This assumption relates to the exchangeability feature mentioned above.)

In practice, violations of some of these assumptions may be overcome by modifying the statistical model; e.g., differences in sampling protocols among different studies – violating assumption 3 – may be incorporated *via* some form of weighting to de-emphasize the contribution of lower-quality studies.

We also make the implicit assumption that results from all relevant studies in the scientific literature are available and accessible for the meta-analysis. Failure to meet this assumption is often called the *file drawer problem* [2]; it is a form of *publication bias* [3, 4] and if present, can undesirably affect the analysis. Efforts to find solutions to this issue represent a continuing challenge in modern quantitative science [5–7].

Combining P Values

Perhaps the most well-known and simplest approach to combining information collects together P values from $K \geq 1$ individual, independent studies of the same null hypothesis, H_0 , and aggregates the associated statistical inferences into a single, combined P value. Fisher gave a basic meta-analytic technique toward this end; a readable exposition is given in [8]. Suppose we observe the P values $P_k, k = 1, \dots, K$. Under H_0 , each P value is distributed as independently uniform on the unit interval, and using this Fisher showed that the transformed quantities $-2 \log(P_k)$ are each distributed as $\chi^2(2)$. Since their sum is then $\chi^2(2K)$, one can combine the independent P values together into an aggregate statistic: $X_{\text{calcd}}^2 = -2 \sum_{k=1}^K \log(P_k)$. The resulting combined P value is then $P_+ = P[\chi^2(2K) \geq X_{\text{calcd}}^2]$. Report combined significance if P_+ is less than some predetermined significance level, α . This approach often is called the *inverse χ^2 method*, since it inverts the P values to construct the combined test statistic.

One can construct alternative methods for combining P values; for instance, from a set of independent P values, $P_1, P_2, \dots, P_K$, each testing the same H_0 , the quantities $\Phi^{-1}(P_k)$ can be shown to be distributed as standard normal: $\Phi^{-1}(P_k) \sim \text{i.i.d. } N(0, 1)$, where $\Phi^{-1}(\bullet)$ is the inverse of the standard normal cumulative distribution function. We then sum the individual terms to find $\sum_{k=1}^K \Phi^{-1}(P_k) \sim N(0, K)$. Dividing by the corresponding standard deviation, $\sqrt{K}$, yields yet another standard normal random variable. Thus, the

quantity

$$Z_{\text{calcd}} = \frac{1}{\sqrt{K}} \sum_{k=1}^K \Phi^{-1}(P_k) \quad (1)$$

can be used to make combined inferences on H_0 . Owing to Stouffer *et al.* [9], this is known as the *inverse normal method*, or also *Stouffer's method*. To combine the P values into a single aggregate, calculate from Z_{calcd} the lower tail quantity $P_+ = \Pr[Z \leq Z_{\text{calcd}}]$, where $Z \sim N(0,1)$. Notice that this is simply $P_+ = \Phi(Z_{\text{calcd}})$. As with the inverse χ^2 method, report combined significance if P_+ is less than a predetermined α .

Historically, the first successful approach to combining P values involved neither of these two methods. Instead, it employed the ordered P values, denoted as $P_{[1]} \leq P_{[2]} \leq \dots \leq P_{[K]}$. Originally proposed by Tippett [10], the method took the smallest P value, $P_{[1]}$, and rejected H_0 if $P_{[1]} < 1 - (1 - \alpha)^{1/K}$. Wilkinson [11] extended Tippett's method by using the L th ordered P value: reject H_0 from the combined data if $P_{[L]} < C_{\alpha,K,L}$, where $C_{\alpha,K,L}$ is a critical point found using specialized tables [12]. Wilkinson's extension is more resilient to possible outlying effects than Tippett's method, since it does not rest on the single lowest P value. On the other hand, since it relies upon specialized tables it is also far less useful than the other methods discussed above.

Fisher's observation that a P value under H_0 is uniformly distributed can motivate a variety of statistical manipulations to combine independent P values. The few described above represent only the more traditional approaches. For a discussion of some others, see Hedges and Olkin [12].

Effect Size Estimation

While useful and simple, combined P values have drawbacks. By their very nature, P values are summary measures that may overlook or fail to emphasize relevant differences among the various independent studies [13]. To compensate for potential loss of information, one can directly calculate the size of the effect detected by a significant P value. For simplicity, suppose we have a simple two-group experiment where the effect of an external hazardous stimulus is to be compared with an appropriate control group. For data recorded on a continuous scale, the simplest

way to index the effect of the stimulus is to take the difference in observed mean responses between the two groups. When combining information over two such independent studies, it is common to standardize the difference in means by scaling inversely to its standard deviation. Championed by Cohen [14], this is known as a *standardized mean difference*, which for use in meta-analysis is often called an *effect size* [15].

Formally, consider a series of independent two-sample studies. Model each observation Y_{ijk} as the sum of an unknown group mean μ_i and an experimental error term ε_{ijk} : $Y_{ijk} = \mu_i + \varepsilon_{ijk}$, where $i = C$ (control group), T (target group); $j = 1, \dots, J$ studies, and $k = 1, \dots, N_{ij}$ replicates per study. (The indexing can be extended to include a stratification variable, if present; see [16].) We assume that the additive error terms are independently normally distributed, each with mean zero, and with standard deviations, $\sigma_j > 0$, that may vary across studies but remain constant between the two groups in a particular study. Under this model, each effect size is measured *via* the standardized mean difference

$$d_j = \frac{\varphi_j(\bar{Y}_{Tj} - \bar{Y}_{Cj})}{s_j} \quad (2)$$

where $\bar{Y}_{ij}$ is the sample mean of the N_{ij} observations in the j th study ($i = C, T$), s_j is the pooled standard deviation

$$s_j = \sqrt{\frac{(N_{Tj} - 1)s_{Tj}^2 + (N_{Cj} - 1)s_{Cj}^2}{N_{Tj} + N_{Cj} - 2}} \quad (3)$$

using the corresponding per-study sample standard deviations s_{ij} , and φ_j is an adjustment factor to correct for bias in small samples:

$$\varphi_j = 1 - \frac{3}{4(N_{Tj} + N_{Cj} - 2) - 1} \quad (4)$$

We combine the individual effect sizes in equation (2) over the J independent studies by weighting each effect size inversely to its estimated variance, $\text{Var}[d_j]$: the weights are $w_j = 1/\text{Var}[d_j]$. A large-sample approximation for these variances that operates well when the samples sizes, N_{ij} , are roughly equal and are at least 10 for all i and j is [12]

$$\text{Var}[d_j] \approx \frac{N_{Tj} + N_{Cj}}{N_{Tj}N_{Cj}} + \frac{d_j^2}{2(N_{Tj} + N_{Cj})} \quad (5)$$

With these, the weighted averages are

$$\bar{d}_+ = \frac{\sum_{j=1}^J w_j d_j}{\sum_{j=1}^J w_j} \quad (6)$$

Standard practice views a combined effect size as minimal (or “none”) if in absolute value it is near zero, as “small” if it is near $d = 0.2$, as “medium” if it is near $d = 0.5$, as “large” if it is near $d = 0.8$, and as “very large” if it exceeds 1. To assess this statistically, we find the standard error of $\bar{d}_+$ as $se[\bar{d}_+] = 1/\sqrt{\sum_{j=1}^J w_j}$, and build a $1 - \alpha$ confidence interval on the true effect size. Simplest is the large-sample “Wald” interval $\bar{d}_+ \pm z_{\alpha/2} se[\bar{d}_+]$, with critical point $z_{\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2})$.

Example: Manganese Toxicity

In a study of pollution risk, Ashraf and Jaffar [17] reported on metal concentrations in the scalp hair of males exposed to industrial plant emissions in Pakistan. For purposes of comparison and control, hair concentrations were also determined from an unexposed urban population ($i = C$). Of interest was whether exposed individuals ($i = T$) exhibited increased metal concentrations in their scalp hair, and if so, how this can be quantified *via* effect size calculations. The study was conducted for a variety of pollutants; for simplicity, we consider a single outcome: manganese concentration in scalp hair. Among six ostensibly homogenous male cohorts (the “studies”), the sample sizes, observed mean concentrations, and sample variances were found as given in Table 1.

(Note that the sample sizes are all large enough with these data to validate use of the large-sample approximation for $\bar{d}_+$.) The per-cohort effect sizes, d_j , based on these data can be computed as $d_1 = 0.558$, $d_2 = 0.785$, $d_3 = 0.763$, $d_4 = 0.544$, $d_5 = 1.080$, and $d_6 = 0.625$. That is, for all cohorts the exposure effect leads to increased manganese concentrations in the T group relative to the C group (since all d_j ’s are positive), ranging from an increase of 0.558 standard deviations for cohort 1 to an increase of over one standard deviation for cohort 5. To determine a combined effect size, we calculate the inverse-variance weights as $w_1 = 8.154$, $w_2 = 7.133$, $w_3 = 10.482$, $w_4 = 10.595$, $w_5 = 7.082$, and $w_6 = 8.581$. From these values, one finds using equation (6) that

$$\begin{aligned} \bar{d}_+ &= \frac{(8.154)(0.588) + \cdots + (8.581)(0.625)}{8.154 + \cdots + 8.581} \\ &= 0.710 \end{aligned} \quad (7)$$

For a 95% confidence interval, calculate

$$se^2[\bar{d}_+] = \frac{1}{8.154 + \cdots + 8.581} = 0.019 \quad (8)$$

with which we find $\bar{d}_+ \pm z_{0.025} se[\bar{d}_+] = 0.71 \pm (1.96)\sqrt{0.019} = 0.71 \pm 0.27$. Overall, a “medium-to-large” combined effect size is indicated with these data: on average, an exposed male has manganese concentrations in scalp hair between 0.44 and 0.98, standard deviations larger than an unexposed male.

Assessing Homogeneity

One can also assess homogeneity across studies, using the statistic $Q_{\text{calcd}} = \sum_{j=1}^J w_j (d_j - \bar{d}_+)^2$ [18], where as above, $w_j = 1/\text{Var}[d_j]$. Under the

Table 1 Sample sizes, sample means, and sample variances for manganese toxicity data, where outcomes were measured as manganese concentration in scalp hair

Cohort, j	1	2	3	4	5	6
<i>i</i> = C (controls)						
N_{Cj}	18	17	22	19	12	18
$\bar{Y}_{Cj}$	3.50	3.70	4.50	4.00	5.56	4.85
s_{Cj}^2	2.43	14.06	4.41	12.25	9.99	4.54
<i>i</i> = T (exposed)						
N_{Tj}	16	14	23	26	24	18
$\bar{Y}_{Tj}$	4.63	6.66	6.46	5.82	9.95	6.30
s_{Tj}^2	5.57	12.74	8.24	9.73	18.58	5.76

null hypothesis of homogeneity across all studies, $Q_{\text{calcd}} \sim \chi^2(J-1)$. Here again, this is a large-sample approximation that operates well when the samples sizes, N_{ij} , are all roughly equal and are at least 10. For smaller sample sizes the test can lose sensitivity to detect departures from homogeneity, and caution is advised.

Reject homogeneity across studies if the P value $P[\chi^2(J-1) \geq Q_{\text{calcd}}]$ is smaller than some predetermined significance level, α . If study homogeneity is rejected, combination of the effects sizes *via* equation (6) is contraindicated since the d_j 's may no longer estimate a homogeneous quantity. Hardy and Thompson [19] give additional details on use of Q_{calcd} and other tools for assessing homogeneity in a meta-analysis; also see [4].

Applied to the manganese toxicity data in the example above, we find $Q_{\text{calcd}} = \sum_{j=1}^6 w_j (d_j - 0.71)^2 = \dots = 1.58$. The corresponding P value for testing the null hypothesis of homogeneity among cohorts is $P[\chi^2(5) \geq 1.58] = 0.90$, and we conclude that no significant heterogeneity exists for this endpoint among the different cohorts. The calculation and reporting of a combined effect size here is validated.

Informative Weighting

Taken broadly, an “effect size” in an environmental study can be any valid quantification of the effect, change, or impact under study [15], not just the difference in sample means employed above. Thus, e.g., we might use estimated coefficients from a regression analysis, potency estimators from a cancer bioassay, correlation coefficients, etc. For any such measure of effect, the approach used in equation (6) corresponds to a more general, weighted averaging strategy to produce a combined estimator. Equation (6) uses “informative” weights based on inverse variances. Generalizing this approach, suppose an effect of interest is measured by some unknown parameter ξ , with estimators $\hat{\xi}_k$ found from a series of independent, homogeneous studies, $k = 1, \dots, K$. Assume that a set of weights, w_k , can be derived such that larger values of w_k indicate greater information/value/assurance in the quality of $\hat{\xi}_k$. A combined

estimator of ξ is then

$$\bar{\xi} = \frac{\sum_{k=1}^K w_k \hat{\xi}_k}{\sum_{k=1}^K w_k} \quad (9)$$

with standard error $se[\bar{\xi}] = 1/\sqrt{\sum_{k=1}^K w_k}$. If the $\hat{\xi}_k$'s are distributed as approximately normal, then an approximate $1 - \alpha$ “Wald” interval on the common value of ξ is $\bar{\xi} \pm z_{\alpha/2} se[\bar{\xi}]$. Indeed, even if the $\hat{\xi}_k$'s are not close to normal, for large K the averaging effect in equation (9) may still imbue approximate normality to $\bar{\xi}$, and hence the Wald interval may still be approximately valid. For cases where approximate normality is difficult to achieve, use of bootstrap resampling methods can be useful in constructing confidence limits on $\bar{\xi}$ [20].

A common relationship often employed in these settings relates the “information” in a statistical quantity inversely to the variance [21]. Thus, given values for the variances, $\text{Var}[\hat{\xi}_k]$, of the individual estimators in equation (9), an “informative” choice for the weights is the reciprocal (“inverse”) variances: $w_k = 1/\text{Var}[\hat{\xi}_k]$. This corresponds to the approach we applied in equation (6). More generally, inverse-variance weighting is a popular technique for combining independent, homogeneous information into a single summary measure. It was described in early reports by Birge [22] and Cochran [18], and Gaver *et al.* [13] give a concise overview.

Combining Information across Models *versus* across studies

Another area where the effort to combine information is important in quantitative risk analysis occurs when information is combined across models for a given study, rather than across studies for a given model. That is, we wish to describe an underlying phenomenon observed in a single set of data by combining the results of competing models for the phenomenon. Here again, a type of informative weighting finds popular application. Suppose we have a set of K models, $M_1, \dots, M_K$, each of which provides information on a specific, unknown parameter θ . Given only a single available data set, we can calculate for θ a point estimator, such as

the maximum-likelihood estimator (MLE) $\hat{\theta}_k$, based on fitting the k th model. Then, by defining weights, w_k , that describe the information in or quality of model M_k 's contribution for estimating θ , we can employ the weighted estimator $\bar{\theta} = \sum_{k=1}^K w_k \hat{\theta}_k$. For the weights, Buckland *et al.* [23] suggest

$$w_k = \frac{\exp(-I_k/2)}{\sum_{i=1}^K \exp(-I_i/2)} \quad (10)$$

where I_k is an information criterion (IC) measure that gauges the amount of information each model provides for estimating θ . Most typical is the general form $I_k = -2\log(L_k) + q_k$, where L_k is the value of the statistical likelihood evaluated under model M_k at that model's MLE, and q_k is an adjustment term that accounts for differential parameterizations across models. This latter quantity is chosen prior to sampling; if we set q_k equal to twice the number of parameters in model M_k , I_k will correspond to the popular Akaike information criterion (AIC) [24]. Alternatively, if we set q_k equal to the number of parameters in model M_k times the natural log of the sample size, I_k will correspond to Schwarz' Bayesian information criterion (BIC) [25]. Other information-based choices for I_k and hence w_k are also possible [26].

Notice that the definition for w_k in equation (10) automatically forces $\sum w_k = 1$, which is not an unusual restriction. Since differences in ICs are typically meaningful, some authors replace I_k in equation (10) with the differences $I_k - \min_{k=1, \dots, K} \{I_k\}$. Of course, this produces the same set of weights once normalized to sum to 1.

Uses of this sort of weighted *model averaging* in risk analytic applications has only recently developed; examples include the use of AIC-based weights by Kang *et al.* to average risk or dose estimates across different microbiological dose-response models [27], Morales *et al.*'s model-averaged benchmark dose estimates for arsenic exposures leading to lung cancer [28], and a study by Wheeler and Bailer [29] comparing model averaging *versus* other model selection options for benchmark dose estimation with quantal data.

Acknowledgments

Preparation of this material was supported in part by grant #R01-CA76031 from the US National Cancer Institute, and grant #RD-83241901 from the US Environmental Protection Agency. Its contents are solely the responsibility of the authors and do not necessarily reflect the official views of these various agencies.

References

- [1] Glass, G.V. (1976). Primary, secondary, and meta-analysis of research, *Educational Researcher* **5**, 3–8.
- [2] Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results, *Psychological Bulletin* **86**, 638–641.
- [3] Thornton, A. & Lee, P. (2000). Publication bias in meta-analysis: its causes and consequences, *Journal of Clinical Epidemiology* **53**, 207–216.
- [4] Tweedie, R.L. (2002). Meta-analysis, in *Encyclopedia of Environmetrics*, A.H. El-Shaarawi & W.W. Piegorsch, eds, John Wiley & Sons, Chichester, Vol. 3, pp. 1245–1251.
- [5] Baker, R. & Jackson, D. (2006). Using journal impact factors to correct for the publication bias of medical studies, *Biometrics* **56**, 785–792.
- [6] Copas, J. & Jackson, D. (2004). A bound for publication bias based on the fraction of unpublished studies, *Biometrics* **60**, 146–153.
- [7] Shi, J.Q. & Copas, J.B. (2004). Meta-analysis for trend estimation, *Statistics in Medicine* **23**, 3–19.
- [8] Fisher, R.A. (1948). Combining independent tests of significance, *American Statistician* **2**, 30.
- [9] Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. & Williams Jr, R.M. (1949). *The American Soldier: Adjustment During Army Life*, Princeton University Press, Princeton, Vol. I.
- [10] Tippett, L.H.C. (1931). *The Methods of Statistics*, Williams & Norgate, London.
- [11] Wilkinson, B. (1951). A statistical consideration in psychological research, *Psychological Bulletin* **48**, 156–158.
- [12] Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*, Academic Press, Orlando.
- [13] Gaver, D.P., Draper, D., Goel, P.K., Greenhouse, J.B., Hedges, L.V., Morris, C.N. & Waternaux, C. (1992). *Combining Information: Statistical Issues and Opportunities for Research*, The National Academy Press, Washington, DC.
- [14] Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*, Academic Press, New York.
- [15] Umbach, D.M. (2002). Effect size, in *Encyclopedia of Environmetrics*, A.H. El-Shaarawi & W.W. Piegorsch, eds, John Wiley & Sons, Chichester, Vol. 2, pp. 629–631.
- [16] Piegorsch, W.W. & Bailer, A.J. (2005). *Analyzing Environmental Data*, John Wiley & Sons, Chichester.

- [17] Ashraf, W. & Jaffar, M. (1997). Concentrations of selected metals in scalp hair of an occupationally exposed population segment of Pakistan, *International Journal of Environmental Studies, Section A* **51**, 313–321.
- [18] Cochran, W.G. (1937). Problems arising in the analysis of a series of similar experiments, *Journal of the Royal Statistical Society, Supplement* **4**, 102–118.
- [19] Hardy, R.J. & Thompson, S.G. (1998). Detecting and describing heterogeneity in meta-analysis, *Statistics in Medicine* **17**, 841–856.
- [20] Wood, M. (2005). Statistical inference using bootstrap confidence intervals, *Significance* **1**, 180–182.
- [21] Fisher, R.A. (1925). Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.
- [22] Birge, R.T. (1932). The calculation of errors by the method of least squares, *Physical Review* **16**, 1–32.
- [23] Buckland, S.T., Burnham, K.P. & Augustin, N.H. (1997). Model selection: an integral part of inference, *Biometrics* **53**, 603–618.
- [24] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Proceedings of the Second International Symposium on Information Theory*, B.N. Petrov & B. Csaki, eds, Akademiai Kiado, Budapest, pp. 267–281.
- [25] Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464.
- [26] Burnham, K.P. & Anderson, D.A. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*, 2nd Edition, Springer-Verlag, New York.
- [27] Kang, S.-H., Kodell, R.L. & Chen, J.J. (2000). Incorporating model uncertainties along with data uncertainties in microbial risk assessment, *Regulatory Toxicology and Pharmacology* **32**, 68–72.
- [28] Morales, K.H., Ibrahim, J.G., Chen, C.-J. & Ryan, L.M. (2006). Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water, *Journal of the American Statistical Association* **101**, 9–17.
- [29] Wheeler, M.W. & Bailer, A.J. (2007). Properties of model-averaged BMDLs: a study of model averaging in dichotomous risk estimation, *Risk Analysis* **27**, 659–670.

Related Articles

Meta-Analysis in Nonclinical Risk Assessment

WALTER W. PIEGORSCH AND A. JOHN BAILER

Common Cause Failure Modeling

Common cause failures (CCFs) are generally defined as simultaneous failures of multiple components due to a common cause. “Simultaneity” can be perfect or a time frame such that the failed states (downtimes, unavailabilities) of multiple components overlap. Failure causes can originate from design, manufacturing or installation weaknesses that are common to many components or from environmental stresses (impact, vibration, grit, heat, etc.), common maintenance weaknesses or human errors in operation. A cause can be external to the components, or it can be a single failure that propagates to other components in a cascade (like a pipe whip or a turbine missile). Even if such events are rare compared to single failures of a component, they can dominate system unreliability or unavailability. For example, consider a system of four parallel similar trains of components with the same unavailability $p = 0.02$. Without CCF the system unavailability is calculated as $P_0 = p^4 = 1.60 \times 10^{-7}$. Assume that a small fraction $\beta = 0.03$ of component unavailabilities are due to causes that fail all four components, causing the system to fail. Then the single failure unavailability is $(1 - \beta)p$, and the system unavailability $P_c = \beta p + (1 - \beta p)[(1 - \beta)p]^4 = 6.00 \times 10^{-4}$. The result is dominated by CCF.

Before quantified event data was available on CCF a bounding method was used, assuming the system failure probability to be a geometric mean of two bounds. The lower bound is the one that would exist if there were no CCF and all probabilities are treated as independent. The upper bound equals the largest single subsystem failure probability. In the example, this yields system probability $P = \sqrt{p \cdot p^4} = 5.66E-5$. Later on, several parametric models, like the β -factor model above, were developed under the assumption and hope that such parameters would be globally generic constants that everybody could use without having to estimate them for each plant or system separately. After many more CCF events have been observed it has become evident that also the ratio parameters like β depend on component types and failure modes,

and vary from system to system even for similar components.

In probabilistic safety assessments (PSAs) CCF can have major impacts in two ways:

- in *production systems* they can cause initiating events and demand safety systems to start and operate, and
- in *safety systems* they fail multiple components, usually similar components in redundant subsystems (trains).

In normally operating systems, a CCF occurs at a random time owing to *time-related stresses*, and a proper model is based on general multiple-failure rates (GMFR) $\lambda_{a,b,\dots}$ such that $\lambda_{a,b,\dots} dt$ is the probability of an event failing the specific components $a, b, \dots$ in a small time interval dt .

In standby safety systems, failures can be caused by two kinds of stresses. *Demand-related stresses* cause components $a, b, \dots$ to fail with a certain general multiple-failure probability (GMFP) $q_{a,b,\dots}$ per system demand, at the time of an initiator or in a periodic test. Demand-related stresses cause degradation due to cold-start-up wear, crack propagation, and thermal transients or loosening etc. due to start-up shocks. Failures can also occur owing to *time-related stresses* at random times (due to corrosion, thermal aging, creep, embrittlement, externally caused vibration and fatigue, random external loads or chemical attacks, dust, dirt, or dry-out). These are properly modeled by a probability per unit time, i.e., by standby failure rates $\lambda_{a,b,\dots}$. Usually, these failures remain unrevealed until discovered by a scheduled test. Test demands of a component are carried out at certain times, usually with regular time intervals T . If a failure is discovered in a test, the component is immediately repaired. These two categories of CCF are addressed here:

- failures caused by demand (start-up) stresses and
- failures caused by time-related stresses.

Any group of standby components can be subject to both types of stresses and may have finite values for both $q_{a,b,\dots}$ and $\lambda_{a,b,\dots}$. A reason to handle these separately is that their impact on a system depends differently on test interval T , a quantity that may be varied to control risk.

Explicit Models and Multiple-Failure Parameters

Explicit modeling of multifailure states means that a CCF of components $a, b, c, \dots$ is described in a system model by a basic event $Z_{a,b,c,\dots}$, input to the OR gate of each component $a, b, c, \dots$, parallel to the individual single failure events $Z_a, Z_b, Z_c, \dots$, indicated by the subindexes. This is illustrated by Figure 1. The components being subject to a set of common causes are called a *common cause component group* (CCCG), usually n similar components in n parallel trains (subsystems) of a safety system, and n is called the size of the group.

Common cause failures that occur owing to the stresses caused by a true system demand (an initiating event) or a system test demand (simultaneous or consecutive testing of all n components of a CCCG) are usually modeled by constant probabilities $z_{a,b,c,\dots} \equiv P(Z_{a,b,c,\dots}) = q_{a,b,c,\dots}$, which is the fraction of system demands that fail $a, b, \dots$. For another interpretation, consider initiating events occurring evenly at any time during a test interval. Then, $q_{a,b,c,\dots}$ is the probability that the preceding system test has damaged or weakened components $a, b, \dots$ so that they can not perform if a true demand occurs before the next test cycle. The time window of vulnerability for this CCF is the interval T , because at the next test the damage would be discovered and repaired. The fraction of vulnerable intervals is the basic event probability $z_{a,b,c,\dots} = q_{a,b,c,\dots}$.

Often, subsets of components of a CCCG are assumed to be identically subject to common

causes so that all combinations of the same multiplicity have the same probability. This *symmetry assumption* allows definitions $q_{1/n} \equiv q_a = q_b = \dots$, $q_{2/n} \equiv q_{a,b} = q_{a,c} = q_{b,c} = \dots$, $q_{3/n} \equiv q_{a,b,c} = q_{a,c,d} \dots$, etc. With simultaneous testing, the basic event probabilities are $z_{k/n} = q_{k/n}$. The total probability of a k/n event (CCF of k components out of n) in a system demand is $Q_{k/n} = n!/(k!(n-k)!)q_{k/n}$.

An alternative scheme to test n components is *evenly staggered* testing, in which case there is a time shift T/n between successive tests of components. Assume that there is a rule to do extra testing of all other components when one is found to demand repair, so called extra-testing rule (ETR). The effect of staggered testing and ETR is that a CCF is discovered and repaired sooner than with simultaneous testing. Because the test cycle length of any specific subset of k components is T , the mean time between consecutive tests of these components is T/k . Thus, a test-caused CCF of any k components is discovered and repaired within T/k in the average. This is the average downtime for a specific CCF ($Z_{k/n}$) in fraction $q_{k/n}$ of the cycles. Summarizing for demand-caused CCF the basic event probabilities (slightly idealized) are

$$z_{k/n} = q_{k/n} \quad (1)$$

with simultaneous or consecutive testing

$$z_{k/n} = \frac{q_{k/n}}{k} \quad (2)$$

with staggered testing and ETR.

In both cases one may add the repair time contribution $q_{k/n}\tau_{k/n}/T$, where $\tau_{k/n}$ is the downtime of a

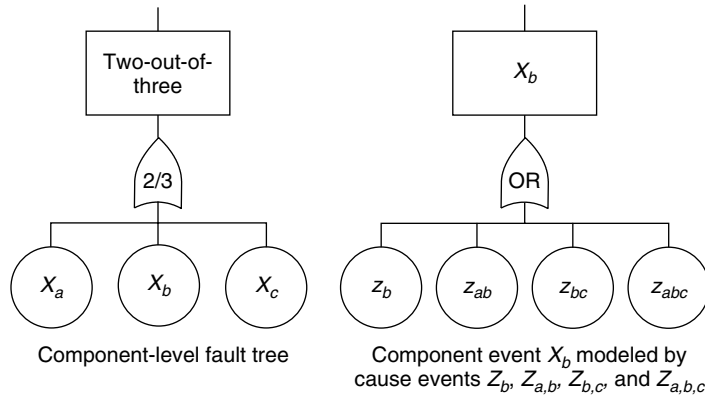


Figure 1 Developing a component-event fault tree to a cause-event fault tree

group of k components due to repair of CCF. This does not depend on the testing scheme.

On the other hand, time-related stresses cause time-dependent unavailabilities between tests. $P(Z_{a,b,\dots}) = \lambda_{a,b,\dots}(t - T_i)$ is the probability of failed states of components $a, b, \dots$ at time t due to a common cause failing exactly these components simultaneously with rate $\lambda_{a,b,\dots}$ when the last possible discovery and repair of such failure occurred at T_i . The time factors are usually small so that the linear approximation is valid. The symmetry assumption reduces the number of rates because then $\lambda_{1/n} \equiv \lambda_a = \lambda_b = \dots, \lambda_{2/n} \equiv \lambda_{a,b} = \lambda_{a,c} = \lambda_{b,c} = \dots, \lambda_{3/n} \equiv \lambda_{a,b,c} = \lambda_{a,c,d} \dots$, etc. All together k/n events occur with the total rate $\Lambda_{k/n} = n!/(k!(n-k)!) \lambda_{k/n}$.

Often, the time-average risk is analyzed rather than the risk as a function of time. The basic event probabilities $P(Z_{a,b,\dots})$ that yield the time-average risk (slightly idealized again) are [1]

$$z_{k/n} \approx \frac{1}{2} \lambda_{k/n} T \quad (3)$$

when all trains are tested simultaneously or consecutively

$$z_{k/n} \approx \frac{1}{2k} \lambda_{k/n} T \quad (4)$$

for staggered testing with ETR.

In both cases one may add the repair-time contribution $\lambda_{k/n} \tau_{k/n}$. Without ETR some $z_{k/n}$ are larger than equations (2, 4) indicate.

Other parametric models will be presented in relation to the GMFRs $\lambda_{k/n}$ and GMFPs $q_{k/n}$ because these parameters are easy to estimate and they do not depend on testing schemes or test intervals.

Basic Parametric Models

The early CCF models were developed under the assumption that a few parameters could model and predict CCFs. These models and their common features are now described.

β -Factor model

This model was first introduced for the case $n = 2$ with a ratio parameter, $\beta = (\lambda_{2/2})/(\lambda_{1/2} + \lambda_{2/2})$ [2]. If this is known, one needs to estimate only a single failure rate $\lambda_{1/2}$ for a specific plant and

component type, then obtain $\lambda_{2/2} = \beta/(1 - \beta)\lambda_{1/2}$, and use this in the basic event probability equation for $z_{2/2}$. Numerical values of β have been presented for several types of components [3, 4] in the range 0.01–0.09, with error factors about 2.5. Likewise, β can be defined for demand-caused failures as $q_{2/2}/(q_{1/2} + q_{2/2})$. The model can be generalized for larger CCCG sizes $n > 2$ as was illustrated in the introduction for $n = 4$.

Binomial Failure-Rate Model (BFRM)

The binomial model [5, 6] was also originally based on time-related failures. It makes the symmetry assumption and has four parameters, λ , μ , p , and ω . The independent failure rate of a single component is λ , not counting any faults due to common causes, and ω is the rate of “lethal” shocks that fail all components of a CCCG, independent of the size n . The third rate μ is the rate of nonlethal shocks that can fail any one of the n components with a certain probability (“coupling parameter”) p . The number failed by a nonlethal shock is binomially distributed, and

$$\lambda_{1/1} = \lambda + \mu p + \omega \quad (5)$$

$$\lambda_{1/n} = \lambda + \mu p(1 - p)^{n-1}, n > 1 \quad (6)$$

$$\lambda_{k/n} = \mu p^k (1 - p)^{n-k}, 1 < k < n \quad (7)$$

$$\lambda_{n/n} = \mu p^n + \omega, n > 1 \quad (8)$$

One can conclude that data from a system with $n = 2$ or 3 alone would not be enough to estimate four parameters, but data from systems with $n = 4$ would be just enough. If data is available from both systems with $n = 2$ and 3, then estimated five rates $\lambda_{1/2}$, $\lambda_{2/2}$, $\lambda_{1/3}$, $\lambda_{2/3}$, and $\lambda_{3/3}$ would be more than enough to solve four parameters. In general, some fitting is needed when the number of observable rates is larger than four. Numerical parameters for several types of components have been published [7].

One can define the binomial model for demand-caused failures as well. Then, λ , μ , and ω are redefined as probabilities per system demand rather than per unit time, and equations (5–8) are for the probabilities $q_{k/n}$. Another model with four parameters is a recently introduced process-oriented simulation (POS) model [8].

Stochastic Reliability Analysis Model (SRAM)

The stochastic reliability analysis model (SRAM) assumes a random probability P with a probability density $g(p)$ so that specific k out of n components fail with probability per demand $q_{k/n} = \int_0^1 g(p) p^k (1-p)^{n-k} dp$ [9]. This can also be applied to time-related failures by defining a general rate parameter Λ so that $\lambda_{k/n} = \Lambda q_{k/n}$. A model similar to binomial failure-rate model (BFRM) can be obtained by defining $g(p)$ as a set of Dirac's δ functions. A single beta distribution for $g(p)$ has been defined as a random probability shock model [10]. A more general distributed failure probability model (DFPM) with $g(p) = \sum_j C_j f_j(p)$ has also been suggested, where C_j is the fraction of time "environment j " is present. The coefficients C_j and the densities $f_j(p)$ are empirically determined β densities associated with failure multiplicities j [11]. A model in this category is also the coupling model [12].

Common Load Model

The common load model (CLM) assumes all n components of a CCCG to be subject to a common load X having a cumulative distribution $S(x)$, and each component has a strength Y with a cumulative distribution $F(y)$. Any specific k components then fail with probability $q_{k/n} = \int_{-\infty}^{\infty} F^k(x) [1 - F(x)]^{n-k} dS(x)$ for any k and n [13]. This model can be extended to time-related causes by defining a common rate Λ so that $\lambda_{k/n} = \Lambda q_{k/n}$. The distributions $S(x)$ and $F(x)$ need to be fitted by calibrating the model with relevant empirical data.

Common Features of the Parametric Models

The early parametric models were based on a few essential parameters, some built into the distributions $g(p)$, $F(x)$, and $S(x)$. An important characteristic of each model is the number of free parameters available to determine all basic event probabilities or initiating event rates. It is common to assume that the more parameters in the model, the more flexible it is and potentially more accurate. However, it has been shown that all the models defined so far are constrained by the so-called external cause mapping rule [14]

$$\lambda_{k/n} = \lambda_{k/n+1} + \lambda_{k+1/n+1} \quad (9)$$

and similarly for $q_{k/n}$. This indicates how the rates of systems with different CCCG sizes are coupled under these models. Equation (9) is a constraint because it means that the models have no more than n genuinely free rates associated with all CCCG sizes $n' \leq n$, when in fact the total number of such rates $\lambda_{k'/n'}$ with $k' \leq n' \leq n$ is equal to $n(n+1)/2$. The constraint equation (9) is built also into SRAM and CLM no matter how general and complex distributions are used for $g(p)$, $F(x)$, and $S(x)$. Furthermore, it has been shown that [14]

- for any BFRM and CLM there is an equivalent SRAM, and
- any set of failure rates $\lambda_{k/n}$ satisfying equation (9) can be presented by an SRAM.

Thus, using any of the early parametric models entails inherent adoption of equation (9). Still, there seems to be no physical reason or convincing empirical evidence to prove equation (9) or any other simple mapping rule. This is because systems with different sizes n are often designed by different vendors using different designs, separation, and layout principles. Many plant-specific details and operation and maintenance practices can also mask the impact of n and induce plant-to-plant variations in CCF parameters.

General Ratio Models

Two other parametric CCF models, the alpha-factor model (AFM) [15] and the multiple Greek letter model (MGLM) [16], are based on the ratios of multiple-failure rates or probabilities. These models are general but need at least one rate or probability from outside of the models to determine numerical values for the basic event probabilities.

α -Factor Model (AFM)

This model is based on the following ratios of the basic event rates,

$$\alpha_{k/n} \equiv \frac{\Lambda_{k/n}}{\Lambda_n} \quad (10)$$

where $\Lambda_n \equiv \Lambda_{1/n} + \Lambda_{2/n} + \dots + \Lambda_{n/n}$ is the total event rate. The definition can be made analogously with the demand-related failure probabilities, $\alpha_{k/n} \equiv Q_{k/n}/Q_n$ with $Q_n = Q_{1/n} + \dots + Q_{n/n}$. Obviously,

$\alpha_{1/n} + \alpha_{2/n} + \dots + \alpha_{n/n} = 1$ so that the α s are not mutually independent. If the α 's and at least one of the rates $\Lambda_{k/n}$ or the sum Λ_n are known, all other rates $\Lambda_{k/n}$ and then the basic event probabilities can be solved. For example, $\lambda_{k/n} = \alpha_{k/n} \Lambda_n / \binom{n}{k}$ can be used in equations (3, 4). In this way the model has the same number of free parameters as the total number of rates, and the model is not inherently constrained by equation (9). Nonetheless, an analyst who estimates a plant-specific rate and combines it with global or generic α factors makes a definite assumption about the validity of the average α 's for his target plant.

Multiple Greek Letter Model (MGLM)

In this model the parameters $\beta, \gamma, \delta, \varepsilon, \mu, \dots$ are conditional probabilities defined in terms of the ratios of the partial sums $S_{m,n} = \sum_{k=m}^n k \Lambda_{k/n}$,

$$\begin{aligned} \beta &= \frac{S_{2,n}}{S_{1,n}}, & \gamma &= \frac{S_{3,n}}{S_{2,n}}, \\ \delta &= \frac{S_{4,n}}{S_{3,n}}, & \varepsilon &= \frac{S_{5,n}}{S_{4,n}}, \text{ etc.} \end{aligned} \quad (11)$$

The definitions can be made similarly in terms of the demand failure probabilities $Q_{k/n}$ in place of $\Lambda_{k/n}$. If the Greek letters and one of the sums $S_{m,n}$ is known, all rates can be solved, e.g.,

$$\begin{aligned} \Lambda_{1/n} &= (1 - \beta) S_{1,n}, \\ \Lambda_{2/n} &= \frac{1}{2} \beta (1 - \gamma) S_{1,n}, \\ \Lambda_{3/n} &= \frac{1}{3} \beta \gamma (1 - \delta) S_{1,n}, \text{ etc.} \end{aligned} \quad (12)$$

One approach is to estimate $S_{1,n}$ based on data at a specific plant, and take values of the Greek letters from published generic values for each component type. This entails an assumption of plant-independence of the generic or mean value ratios.

Data and Application Issues

The choice of model depends mainly on answers to the following two questions:

- What kind of data is available for the CCF parameters and observed events?
- What is the intended use of the system reliability or risk model?

One hypothesis made in the past for reaching plant specificity is to assume that one can take generic or average ratio parameters, such as β or α factors, and combine them with plant- or site-specific single failure rates or total failures rates. This hybrid approach is attractive because single-failure data accumulate more rapidly at a single plant or within a family of identical systems. A model based on this approach may be used to assess orders of magnitude risks. But it may need extra care and safety margins if applied to most serious plant-specific operative decisions.

A large collection of α factors and MGL parameters for many types of components and failure modes has been published [17]. It was concluded that CCF parameters depend on component types, and even for similar components they vary among systems and failure modes. For example, β factors are not the same in different systems or for different failure modes of a specific component type. Table 1 shows a sample of mean values and variations of α 's for many types of components and failure modes [18].

If a system model is intended for detailed risk-informed decision making on allowed plant configurations, system modifications, inspections, test intervals, maintenance programs, etc., then the CCF model must have sufficient degrees of freedom and the input parameters should be as plant-specific as possible. This is possible if the plant collects failure data or is part of a failure data collection system, and

Table 1 α -Factor generic mean values and component type variations for $n = 2, 3$, and $4^{(a,b)}$

α factor	$\alpha_{1/2}$	$\alpha_{2/2}$	$\alpha_{1/3}$	$\alpha_{2/3}$	$\alpha_{3/3}$	$\alpha_{1/4}$	$\alpha_{2/4}$	$\alpha_{3/4}$	$\alpha_{4/4}$
Mean value	0.955	0.0454	0.948	0.0281	0.0239	0.944	0.0279	0.0110	0.0173
Standard deviation ^(c)	0.0427	0.0427	0.0381	0.0185	0.0278	0.0367	0.0179	0.0089	0.0209

^(a)Reproduced from [18]. © Elsevier, 2007

^(b)Over 36 component types and 66 failure modes based on estimates in NUREG/CR-5497

^(c)These do not include plant-to-plant variations

has access to a large national CCF-event data collection system (such as [19] or [20]) or the multinational project international common-cause data exchange (ICDE) [21]. Plant-specific CCF parameters may then be estimated by combining data from many plants with an empirically based Bayesian framework. This is briefly described for the parameters GMFR and GMFP.

Estimation of Parameters

Estimation of $\lambda_{k/n}(\Lambda_{k/n})$ and $q_{k/n}(Q_{k/n})$ is described because the basic event probabilities and all other parameters can be presented in terms of these GMFR and GMFP.

Estimation Based on Ideal Data

Consider first demand-caused failures at a single plant or system. The k/n events occur with a total probability $Q_{k/n} = n!/(k!(n-k)!)q_{k/n}$ per system demand ($1 \leq k \leq n$). When N_n system demands (cycles of demands of all n components of a CCG) have been made, the number of k/n events, $N_{k/n}$, has a binomial distribution with the probability parameter $Q_{k/n}$. The expected value is $E(N_{k/n}) = Q_{k/n}N_n$. Estimation of $Q_{k/n}$ based on the observed $N_{k/n}$ and N_n yields a probability density for $Q_{k/n}$, a beta distribution with the following mean value and variance:

$$E(Q_{k/n}) = \frac{N_{k/n} + 1/2}{N_n + 1} \quad (13)$$

$$\sigma^2(Q_{k/n}) = \frac{(N_{k/n} + 1/2)(N_n + 1/2 - N_{k/n})}{(N_n + 1)^2(N_n + 2)} \quad (14)$$

These results can be obtained by Bayesian formalism using noninformative priors, and also by classical-confidence bounds.

The procedure with time-caused CCF is analogous. The observable k/n events occur with a total rate $\Lambda_{k/n} = n!/(k!(n-k)!) \lambda_{k/n}$ and the number of k/n events $N_{k/n}$ in observation time T_n has a Poisson distribution with mean value and variance equal to $E(N_{k/n}) = \Lambda_{k/n}T_n$. Estimation yields the probability distribution gamma with the mean value and the

variance

$$E(\Lambda_{k/n}) = \frac{N_{k/n} + 1/2}{T_n} \quad (15)$$

$$\sigma^2(\Lambda_{k/n}) = \frac{N_{k/n} + 1/2}{T_n^2} \quad (16)$$

It matters little for these estimations how the $N_{k/n}$ events were found: in staggered or nonstaggered tests or in some other system demands. After $\lambda_{k/n} = \Lambda_{k/n}/\binom{n}{k}$ and $q_{k/n} = Q_{k/n}/\binom{n}{k}$ have been estimated, they determine the basic event probabilities (1–4) without any need for other models.

Assessment Uncertainties

Due to uncertainties in event observations, interpretations, and documentation, it is not always known with certainty how many components actually were failed at the time of the observation, and were they caused by demands or time-related stresses. Then the numbers $N_{k/n}$ are not known exactly. The effects of this uncertainty can be quantified by assessing the following conditional assessment probabilities for each plant, for each event observed, and for each event multiplicity k/n : $w_{\lambda,i}(k/n)$ [$w_{q,i}(k/n)$] = the probability, conditional on the symptoms and characteristics observed, that observation i entails a k/n -failure event and the failure is caused by time-related (demand-related) stresses; $0 \leq w_{\lambda,i}(w_{q,v}) \leq 1$.

To quantify these assessment probabilities one has to consider the evidence of a coupling mechanism (common cause), the degrees of degradation of each component, and the simultaneity of the failures. A component is considered failed if it needs repair to be able to work in the next test or true demand. A basic framework for the assessment was presented e.g., in [22], and further refinements and options with multiple k/n events and higher moments in [1, 18, 23, 24]. The assessment probabilities can be used to determine the distributions and moments of $N_{k/n}$, separately for time-related and demand-related causes. The distributions of each $\lambda_{k/n}$ and $q_{k/n}$ are weighted averages of gamma and beta distributions, respectively. The probabilities have the following mean and variance:

$$E(Q_{k/n}) = \frac{\sum_{i=1}^N w_{q,i}(k/n) + 1/2}{N + 1} \quad (17)$$

$$\begin{aligned}
& \sigma^2(Q_{k/n}) \\
&= \frac{\left[\sum_{i=1}^N w_{q,i}(k/n) + 1/2 \right] \times \left[N + 1 - \sum_{i=1}^N w_{q,i}(k/n) - 1/2 \right]}{(N+1)^2(N+2)} \\
&+ \frac{\sum_{i=1}^N w_{q,i}(k/n)[1 - w_{q,i}(k/n)]}{(N+1)(N+2)} \quad (18)
\end{aligned}$$

where N is the total number of observed system demands. Corresponding moments for the time-related rates are

$$E(\Lambda_{k/n}) = \frac{\sum_{i=1}^N w_{\lambda,i}(k/n) + 1/2}{T_n} \quad (19)$$

$$\begin{aligned}
\sigma^2(\Lambda_{k/n}) &= \frac{\sum_{i=1}^N w_{\lambda,i}(k/n) + 1/2}{T_n^2} \\
&+ \frac{\sum_{i=1}^N w_{\lambda,i}(k/n)[1 - w_{\lambda,i}(k/n)]}{T_n^2} \quad (20)
\end{aligned}$$

As an example, consider a hypothetical system of four standby diesel generators. Five observations of CCF events “failure to start” have been made in time $T_4 = 10$ years. The assessment probabilities are listed in Table 2.

Additional 22 single-failure events with $w_{\lambda,i}(1/4) = 1$ are not listed but can be used for estimating $\Lambda_{1/4}$. Equations (19, 20) yield (in units/

year) $E(\Lambda_{1/4}) = 2.25$, $E(\Lambda_{2/4}) = 0.285$, $E(\Lambda_{3/4}) = 0.260$, $E(\Lambda_{4/4}) = 0.105$, $\sigma(\Lambda_{1/4}) = 0.474$, $\sigma(\Lambda_{2/4}) = 0.195$, $\sigma(\Lambda_{3/4}) = 0.192$, $\sigma(\Lambda_{4/4}) = 0.121$. Dividing by the binomial factors this data set yields the mean values and standard deviations

$$\begin{aligned}
E(\lambda_{1/4}) &= 0.563, & \sigma(\lambda_{2/4}) &= 0.119 \\
E(\lambda_{2/4}) &= 0.0475, & \sigma(\lambda_{2/4}) &= 0.0325 \\
E(\lambda_{3/4}) &= 0.0650, & \sigma(\lambda_{3/4}) &= 0.0480 \\
E(\lambda_{4/4}) &= 0.105, & \sigma(\lambda_{4/4}) &= 0.121 \quad (21)
\end{aligned}$$

With staggered testing, ETR, and test interval $T = 4$ weeks the expected basic event probabilities can be obtained from equation (4) as $z_{1/4} = 0.02165$, $z_{2/4} = 0.00091$, $z_{3/4} = 0.00083$, and $z_{4/4} = 0.00101$. With consecutive testing equation (3) yields $z_{1/4} = 0.02165$, $z_{2/4} = 0.00183$, $z_{3/4} = 0.00250$, and $z_{4/4} = 0.00404$.

Extension of the method is available to cases with more than one k/n event (of the same multiplicity) in a single observation [23, 24].

Quantification Using Multiple Data Sources

If event data for a certain type of equipment can be assumed to be from identical systems (e.g., the same n , similar physical separation principles, similar maintenance practices, and similar operating conditions as the system of interest at the target plant), one may pool the observations, the numbers of failures of each multiplicity k/n and observation times (or numbers of system demands) and use the earlier estimation equations for all $\lambda_{k/n}$ and $q_{k/n}$. This highest degree of similarity is rarely the case.

A more realistic approach is to obtain CCF data from a family of systems whose members are not quite identical but are “similar enough” to the system of interest (the target system) in the sense that one can consider each system as a sample from the same family. For example, high-pressure safety injection systems designed by the same vendor with the same CCCG size n and the same layout and separation principles between redundant trains are quite similar, but installation, testing, and maintenance practices may vary between plants causing some variation in CCF frequencies. Assuming such plants to be samples from the same family allows the methodology of

Table 2 Assessment probabilities for diesel generator system $n = 4$ for $k = 2, 3, 4$

Observation i	$w_{\lambda,i}$ (2/4)	$w_{\lambda,i}$ (3/4)	$w_{\lambda,i}$ (4/4)
1	0.90	0.10	0.0
2	0.45	0.50	0.05
3	0.25	0.50	0.25
4	0.50	0.50	0.0
5	0.25	0.50	0.25

empirically based Bayesian methods to be used in estimating CCF rates for each plant of the family, including the plant of interest. This method uses available data to maximal extent without assuming identity, and yields proper values also for plants that have experienced no CCF. Input to this method are the sets $\{(N_{k/n}, T_n)\}$ or $\{(N_{k/n}, N_n)\}$, data pairs for each plant in the family. From these the empirical or hierarchical Bayes methods construct a prior distribution and posterior distributions for all plants, or just for a single target plant.

When assessment uncertainties are present, the numbers $N_{k/n}$ are not exactly known. One approach is to define effective pairs $\{N_{k/n}^*, T_n^*\}$ or $\{N_{k/n}^*, N_n^*\}$ so that equations (13–16) yield exactly the moments (equations 17–20). Then such virtual data input to the Bayesian process yields prior and posterior distributions for the rates and the probabilities consistent with the observations up to the first two moments. The procedure is described in [1] for time-related failures, and the essentials in [25] for demand-related failures. Figure 2 shows the overall procedure (*see Bayesian Statistics in Quantitative Risk Assessment*).

Applications of this procedure for emergency cooling system pumps and diesel generators indicate that the choice of the data base and source plants influences the results, and plant-to-plant variations can be large [1].

As long as CCF data is sparse there is a temptation to adopt event data even from plants with different CCCG sizes n . This requires assumptions about how to judge the applicability of events and modify the rates or assessment weights from larger to smaller systems and *vice versa*. There are alternatives to the mapping equations (9) [1, 24]. These somewhat controversial issues may be better to avoid until there is sufficient empirical evidence to show which rules, if any, are valid between the rates of different group sizes n . So far it seems that only the ratios of the total CCF rates $\eta_{n,n+1} = \lambda_{n+1/n+1}/\lambda_{n/n}$ for $n > 1$ are reasonably common for many types of components. This can be concluded from Table 3.

Other Dependent Failures and Methods

The models described so far apply to safety systems vulnerable to CCF occurring randomly in time or

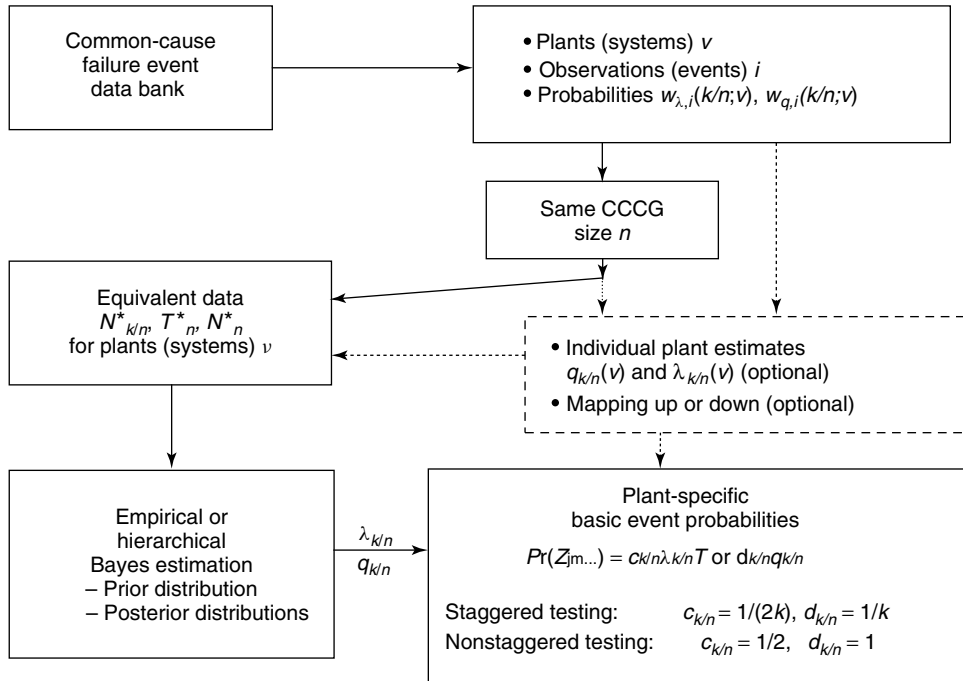


Figure 2 Common cause failure quantification procedure

Table 3 The ratios of total CCF rates $\eta_{n,n+1}$ and component type variations for $n = 1, 2, 3, 4, 5^{(a)}$

$\eta_{n,n+1} = \lambda_{n+1/n+1}/\lambda_{n/n}$	$\eta_{1,2}$	$\eta_{2,3}$	$\eta_{3,4}$	$\eta_{4,5}$	$\eta_{5,6}$
Mean value	0.098	0.73	0.95	1.00	1.05
Standard deviation ^(b)	0.063	0.16	0.044	0.050	0.21

^(a)Over 36 component types and 66 failure modes based on estimates in NUREG/CR-5497

^(b)These do not include plant-to-plant variations

due to stresses caused by tests and demands. The generic multifailure rates apply also for initiating events caused by CCF. Other types of dependent events are as follows:

- Design, manufacturing, and installation errors that make multiple components vulnerable to some initiators from the very beginning. Examples include incorrect pipe supports against seismic events, or temporary flow strainers left in emergency cooling pipelines. Explicit modeling by failure probabilities is rather straightforward, but data for parameters is rare.
- CCF initiators are initiating events that simultaneously fail components in safety systems intended to protect against the initiator. These may be modeled as special initiating events.
- Plant-specific dependences may be created if a system has design features different from earlier plants. These can induce special CCF not yet observed in the past experience or CCF data sources. System-specific walkthroughs, interviews and questionnaires can be used to identify and quantify such vulnerabilities.
- Consecutive human actions are generally dependent [26]. Repeated human errors can be treated as CCF events in risk assessments. The basic dependency models available can account for operator actions and periodic testing and maintenance actions [27].

There is also an *implicit* method for incorporating CCF in system analysis. It is based on first solving the system probability expression without CCF, and then modifying the multicomponent terms to account for CCF [28]. Quite general dependencies between components can also be modeled explicitly by defining independent virtual basic events and their probabilities [29].

A challenge to system designers is how to eliminate or minimize CCF contributions to system failures. Many defenses against CCF have been identified, with the degree of redundancy and diversity usually on top of the list [4]. Identification of defenses is also a part of CCF data collection programs [21].

References

- [1] Vaurio, J.K. (2005). Uncertainties and quantification of common cause failure rates and probabilities for system analysis, *Reliability Engineering and System Safety* **90**, 186–195.
- [2] Fleming, K.N. (1975). A reliability model for common mode failures in redundant safety systems general atomics report GA13284, *Proceedings of the Sixth Annual Pittsburgh Conference on Modeling and Simulation*, Instrument Society of America, Pittsburgh.
- [3] U.S. Nuclear Regulatory Commission. (1987). *Reactor Risk Reference Document*, Report NUREG 1150, U.S. Nuclear Regulatory Commission Washington, DC.
- [4] Andrews, J.D. & Moss, T.R. (2002). *Reliability and Risk Assessment*, Professional Engineering Publishing, London, pp. 269–285.
- [5] Vesely, W.E. (1977). Proceedings of the international conference on estimating common cause failure probabilities in reliability and risk analyses: Marshall-Olkin specializations, in *Proceedings of the international conference on Nuclear Systems Reliability Engineering and Risk Assessment*, J.B. Fussell & G.R. Burdick, eds, SIAM, Philadelphia, pp. 314–341.
- [6] Atwood, C.L. (1980). *Estimators for the Binomial Failure Rate Common Cause Model*, Report NUREG/CR-1401, U.S. Nuclear Regulatory Commission, Washington, DC.
- [7] Atwood, C.L. (1987). Distributions for binomial failure rate parameters, *Nuclear Technology* **79**, 66–81.
- [8] Berg, H.-P., Görtz, R., Schimetschka, E. & Kesten, J. (2006). The process-oriented simulation (POS) model for common cause failures: recent progress, *Kerntechnik* **71**(1–2), 54–59.
- [9] Dörre, P. (1989). Basic aspects of stochastic reliability analysis for redundant systems, *Reliability Engineering and System Safety* **24**, 351–371.
- [10] Hokstad, P. (1988). A shock model for common-cause failures, *Reliability Engineering and System Safety* **23**, 127–145.
- [11] Hughes, R.P. (1987). A new approach to common-cause failure, *Reliability Engineering* **17**, 211–236.
- [12] Kreuser, A., Peschke, J. & Stiller, J.C. (2006). Further development of the coupling model, *Kerntechnik* **71**(1–2), 50–53.
- [13] Mankamo, T. & Kosonen, M. (1992). Dependent failure modeling in highly redundant structures – application to

- BWR safety valves, *Reliability Engineering and System Safety* **35**, 235–244.
- [14] Vaurio, J.K. (1994). The theory and quantification of common cause shock events for redundant standby systems, *Reliability Engineering and System Safety* **43**, 289–305.
- [15] Mosleh, A. & Siu, N.O. (1987). A multi-parameter common-cause failure model, *Proceedings of the 9th International Conference on Structural Mechanics in Reactor Technology*, Lausanne August 17–21, 1987.
- [16] Fleming, K.N. & Kalinowski, A.M. (1983). *An Extension of the Beta Factor Method to Systems with High Levels of Redundancy*, Report PLG-0289, Pickard, Lowe and Garrick.
- [17] Marshall, F.M., Rasmuson, D.M. & Mosleh, A. (1998). *Common-Cause Failure Parameter Estimations*, Report NUREG/CR-5497, U.S. Nuclear Regulatory Commission, Washington, DC.
- [18] Vaurio, J.K. (2007). Consistent mapping of common cause failure rates and alpha factors, *Reliability Engineering and System Safety* **92**, 628–645.
- [19] Fleming, K.N. Rao, S.B. Tinsley, G.A. Mosleh, A. & Afzali, A. (1992). *A Database of Common-Cause Events for Risk and Reliability Applications*, Report EPRI TR-100382, Electric Power Research Institute, Palo Alto.
- [20] Marshall, F.M., Mosleh, A. & Rasmuson, D.M. (1998). *Common-Cause Failure Database and Analysis System*, Report NUREG/CR-6268, U.S. Nuclear Regulatory Commission, Washington, DC.
- [21] Johanson, G., Kreuser, A., Pyy, P., Rasmuson, D. & Werner, W. (2006). OECD/NEA International Common Cause Failure Data Exchange (ICDE) project – insights and lessons learned, *Kerntechnik* **71**(1–2), 13–16.
- [22] Mosleh, A., Rasmuson, D.M. & Marshall, F.M. (1998). Guidelines on *Modeling Common-Cause Failures in Probabilistic Risk Assessment*, Report NUREG/CR-5485, U.S. Nuclear Regulatory Commission, Washington, DC.
- [23] Vaurio, J.K. (2002). Extensions of the uncertainty quantification of common cause failure rates, *Reliability Engineering and System Safety* **78**, 63–69.
- [24] Vaurio, J.K. (2006). Is mapping a part of common cause failure quantification? *Kerntechnik* **71**(1–2), 41–49.
- [25] Vaurio, J.K. (2005). Estimation of failure probabilities based on uncertain binomial event data, in *Proceedings (CD) of PSA'05*, September 11–15, San Francisco, American Nuclear Society, pp. 879–885.
- [26] Swain, A.D. & Guttman, H.E. (1983). *Handbook of Human Reliability Analysis*, Report NUREG/CR-1278, U.S. Nuclear Regulatory Commission.
- [27] Vaurio, J.K. (2001). Modelling and quantification of dependent repeatable human errors in system analysis and risk assessment, *Reliability Engineering and System Safety* **71**, 179–188.
- [28] Vaurio, J.K. (1998). An implicit method for incorporating common-cause failures in system analysis, *IEEE Transactions on Reliability* **47**(2), 173–180.
- [29] Vaurio, J.K. (2002). Treatment of general dependencies in fault tree and risk analysis, *IEEE Transactions on Reliability* **51**(3), 278–287.

Related Articles

Parametric Probability Distributions in Reliability
Reliability Data
Structural Reliability
Systems Reliability

JUSSI K. VAURIO

Comonotonicity

Aggregating Nonindependent Risks

In an actuarial or financial context, one often encounters a random variable (r.v.) S of the type

$$S = \sum_{i=1}^n X_i \quad (1)$$

For an insurer the X_i may represent the claims from individual policies over a specified time horizon, while S represents the aggregate risk related to the entire insurance portfolio. In another context, the X_i may denote the risks of a particular business line, and S is then the aggregate risk across all business lines (see **Actuary; Inequalities in Risk Theory; Insurance Applications of Life Tables; Longevity Risk and Life Annuities; Risk Classification/Life**). In the context of a pension fund, r.v.'s of this type appear when determining provisions and related optimal investment strategies. Another field of application concerns personal finance problems where a decision maker faces a series of future consumptions and looks for optimal saving and investment strategies. Random variables of this type are also used to describe the payoffs of Asian and basket options. Finally, they also appear in the context of capital allocation or capital aggregation (see **From Basel II to Solvency II – Risk Management in the Insurance Sector**). All these applications amount to the evaluation of risk measures related to the cumulative distribution function (cdf) $F_S(x) = \Pr[S \leq x]$ of the r.v. S . The interested reader may refer to [1–6], for more details.

Let us denote the random vector $(X_1, X_2, \dots, X_n)$ by $\underline{X}$. To avoid technical complications, we always assume that the expectations of the X_i exist. When the r.v.'s X_i are mutually independent, the cdf of S can be computed using the technique of convolution or by recursive computation methods such as De Pril's recursion and Panjer's recursion, amongst others; see, for example [7–9].

However, in many actuarial and financial applications, the individual risks X_i in the sums S cannot be assumed to be mutually independent for instance, because all the X_i are influenced by the same economic or physical environment. As a consequence, it is not easy to determine the cdf of S .

Comonotonic Random Variables

Let us consider the situation where the individual risks X_i of the random vector $\underline{X}$ are subject to the same claim generating mechanism such that

$$\underline{X} \stackrel{d}{=} (g_1(Z), g_2(Z), \dots, g_n(Z)) \quad (2)$$

for some r.v. Z and nondecreasing functions g_i . Here, “ $\stackrel{d}{=}$ ” stands for “equality in distribution”. In this case, the random vector $\underline{X}$ is said to be “comonotonic” and the distribution of $\underline{X}$ is called the *comonotonic distribution* (see **Premium Calculation and Insurance Pricing**).

One can prove that the comonotonicity of $\underline{X}$ can also be characterized by

$$\underline{X} \stackrel{d}{=} (F_{X_1}^{-1}(U), F_{X_2}^{-1}(U), \dots, F_{X_n}^{-1}(U)) \quad (3)$$

where $F_{X_i}^{-1}$ denotes the quantile function of the r.v. X_i , and U is a uniformly distributed r.v. on the unit interval. It is easy to prove that comonotonicity of $\underline{X}$ is also equivalent to

$$F_{\underline{X}}(\underline{x}) = \min[F_{X_1}(x_1), F_{X_2}(x_2), \dots, F_{X_n}(x_n)] \quad (4)$$

Hoeffding [10] and Fréchet [11] have shown that the function $\min[F_{X_1}(x_1), F_{X_2}(x_2), \dots, F_{X_n}(x_n)]$ is the cdf of a random vector which has the same marginal distributions as the random vector $\underline{X}$.

Let us denote the sum of the components of the comonotonic random vector $(F_{X_1}^{-1}(U), F_{X_2}^{-1}(U), \dots, F_{X_n}^{-1}(U))$ by S^c :

$$S^c = \sum_{i=1}^n F_{X_i}^{-1}(U) \quad (5)$$

Several important actuarial quantities of S^c , such as quantiles and stop-loss premiums, exhibit an additivity property in the sense that they can be expressed as a sum of corresponding quantities of the marginals involved. For the quantiles, we have that

$$F_{S^c}^{-1}(p) = \sum_{i=1}^n F_{X_i}^{-1}(p), \quad 0 < p < 1 \quad (6)$$

Assuming that the marginal cdf's F_{X_i} are strictly increasing, one can prove that

$$E[S^c - d]_+ = \sum_{i=1}^n E[F_{X_i}^{-1}(U) - d_i^*]_+ \quad (7)$$

for any d such that $0 < F_{S^c}(d) < 1$, and with the d_i^* given by

$$d_i^* = F_{X_i}^{-1}(F_{S^c}(d)) \quad (8)$$

Notice that $\sum_{i=1}^n d_i^* = d$. The expressions (7) and (8) can be generalized to the case of general distribution functions (see [12] and [13]). An expression similar to equation (7) can also be found in [14] where it is proved that in the Vasicek [15] model, a European call option on a portfolio of zero coupon bonds decomposes into a portfolio of European call options on the individual zero coupon bonds in the portfolio (see **Credit Risk Models; Simulation in Risk Management**).

A Comonotonic Upper Bound Approximation

As opposed to the case of independent or comonotonic r.v.'s X_i , it is, in general, not easy to determine the cdf of S . It may therefore be helpful to find a dependency structure for the random vector $\underline{X} = (X_1, X_2, \dots, X_n)$ that leads to a “less favorable” or “more dangerous” sum of the marginal terms X_i with a cdf that is easier to determine. Decisions based on the “less favorable” distribution are prudent or conservative.

To define what we mean by “less favorable”, we have to decide how to order risks. In this respect, it is convenient to consider convex ordering: A r.v. X is convex smaller than a r.v. Y if $E[X] = E[Y]$ and $E[(X - d)_+] \leq E[(Y - d)_+]$ for all real d . In this case, we write

$$X \leq_{cx} Y \quad (9)$$

In von Neumann and Morgenstern's [16] “expected utility theory”, as well as Yaari's [17] “dual theory of choice under risk”, convex order represents the common preferences of risk averse decision makers when choosing between risks with equal expectations; see, for example [18–20] or [21] (see **Axiomatic Measures of Risk and Risk-Value Models; Optimal Risk-Sharing and Deductibles in Insurance**).

One can prove that for any random vector $(X_1, X_2, \dots, X_n)$ it holds that

$$\sum_{i=1}^n X_i \leq_{cx} \sum_{i=1}^n F_{X_i}^{-1}(U) \quad (10)$$

Hence, replacing S by S^c , and making decisions based on the cdf of S^c can be considered as a prudent strategy. Moreover, quantiles and stop-loss premiums of S^c can be easily determined from equations (6) and (7). The comonotonic upper-bound approximation F_{S^c} will be “close” to the exact cdf F_S when $(X_1, X_2, \dots, X_n)$ has a strong positive dependency structure (see **Credit Risk Models; Reinsurance; Risk Measures and Economic Capital for (Re)insurers**). An insightful geometric proof of equation (10) can be found in [22]. Earlier references to closely related results are [23–25].

Several actuarial and financial problems that were mentioned in the section titled “Aggregating Non-independent Risks” involve the evaluation of the present value or the accumulated value of a series of future cash flows, which can be expressed as a sum S as in equation (1) where the r.v.'s X_i are given by

$$X_i = \alpha_i e^{Y_i} \quad (11)$$

Here, the α_i are deterministic real numbers and $(Y_1, Y_2, \dots, Y_n)$ is a random vector. The accumulated value at time n of a series of future deterministic saving amounts α_i can be written in this form, where Y_i denotes the cumulative logreturn over the period $[i, n]$. Similarly, the present value of a series of future deterministic payments α_i can be written in this form, where now e^{Y_i} denotes the random discount factor over the period $[0, i]$. In both cases (compounding and discounting), the random vector $(X_1, X_2, \dots, X_n)$ will not be comonotonic, although neighboring components X_i and X_j will be rather strongly dependent random variables. This is because there is a natural overlapping process when compounding (or discounting) over the different time periods. A continuous version (with payments continuously spread over time) is considered in [26].

Let us now further assume that $\alpha_i > 0$ and that the Y_i are normally distributed. We find that

$$S^c = \sum_{i=1}^n \alpha_i e^{E[Y_i] + \sigma_{Y_i} \Phi^{-1}(U)} \quad (12)$$

where Φ denotes the standard normal cdf. The quantiles and the stop-loss premiums of S^c are then given

by

$$F_{S^c}^{-1}(p) = \sum_{i=1}^n \alpha_i e^{E[Y_i] + \sigma_{Y_i} \Phi^{-1}(p)}, \quad 0 < p < 1 \quad (13)$$

and

$$E[S^c - d]_+ = \sum_{i=1}^n \alpha_i e^{E[Y_i] + \frac{1}{2}\sigma_{Y_i}^2} \Phi(\sigma_{Y_i} - \Phi^{-1}(F_{S^c}(d))) - d(1 - F_{S^c}(d)), \quad 0 < d < \infty \quad (14)$$

respectively.

For a general random vector $(X_1, X_2, \dots, X_n)$ and real d and d_i ($i = 1, 2, \dots, n$), such that $\sum_{i=1}^n d_i = d$, we have the following result:

$$\left[\sum_{i=1}^n X_i - d \right]_+ \leq \sum_{i=1}^n [X_i - d_i]_+ \quad (15)$$

When the X_i represent asset prices at some future date, say t , then the r.v. $\left[\sum_{i=1}^n X_i - d \right]_+$ can be interpreted as the payoff of a European type basket call option at expiration date t , whereas each of the terms $[X_i - d_i]_+$ can be interpreted as the payoff of a European call option on the i -th asset involved at the same expiration date. The inequality equation (15) provides (*see Equity-Linked Life Insurance; Options and Guarantees in Life Insurance*) an infinite number of ways to superreplicate the payoff of the basket option in terms of the individual asset options involved. The superhedging strategy consisting of buying the n European calls with respective exercise prices d_i^* corresponds to the cheapest super-replicating hedging strategy for the basket option under consideration. Similar results hold for Asian options. For more details, we refer to [2, 6, 27–31].

Comonotonic Lower Bound Approximations

In this section, we look at less crude and hence, better approximations for F_S without losing the convenient properties of comonotonic sums. The technique of taking conditional expectations will help us to achieve this goal.

For an appropriate r.v. Λ , we propose to approximate the cdf of S by the cdf of S^l defined by

$$S^l = E[S \mid \Lambda] = \sum_{i=1}^n E[X_i \mid \Lambda] \quad (16)$$

Notice that a continuous version of this technique applied to Asian option pricing is considered in [32].

Let us now assume that all $E[X_i \mid \Lambda]$ are increasing in Λ . In this case, we find that S^l is a comonotonic sum and the quantiles and the stop-loss premiums related to S^l can be expressed as a sum of corresponding quantities for the individual terms $E[X_i \mid \Lambda]$.

As regards an appropriate choice for Λ , notice that when Λ is chosen equal to S we find that $S^l = S$. Therefore, intuitively it is clear that the “closer” Λ is to S , the better the approximation S^l will perform. However, for the Λ to be useful, we must be able to derive an explicit expression for the different $E[X_i \mid \Lambda]$.

The most prominent case which leads to closed form expressions for quantiles and stop-loss premiums of S^l is the one where $X_i = \alpha_i e^{Y_i}$, with all $\alpha_i > 0$ and $(Y_1, Y_2, \dots, Y_n)$ a multivariate normally distributed random vector. We further concentrate on this particular case.

We choose Λ to be a linear combination of the $Y_1, Y_2, \dots, Y_n$:

$$\Lambda = \sum_{i=1}^n \gamma_i Y_i \quad (17)$$

for appropriate choices of the coefficients γ_i . In the literature, several choices for the γ_i have been proposed, see [13, 33, 34].

For the general Λ , as considered in equation (17), we find that

$$S^l = \sum_{i=1}^n \alpha_i e^{E[Y_i] + \frac{1}{2}(1-r_i^2)\sigma_{Y_i}^2 + r_i\sigma_{Y_i} \frac{\Lambda - E[\Lambda]}{\sigma_\Lambda}} \quad (18)$$

where the r_i are the correlations between the Y_i and Λ .

From equation (18), we see that S^l is a comonotonic sum when all correlation coefficients r_i are nonnegative. Notice that the particular choices for the γ_i , as proposed in [13, 33, 34], lead to nonnegative r_i . The quantiles and stop-loss premiums of S^l are

then given by

$$F_{S^l}^{-1}(p) = \sum_{i=1}^n \alpha_i e^{E[Y_i] + \frac{1}{2}(1-r_i^2)\sigma_{Y_i}^2 + r_i\sigma_{Y_i}\Phi^{-1}(p)},$$

$$0 < p < 1 \quad (19)$$

and

$$E[S^l - d]_+ = \sum_{i=1}^n \alpha_i e^{E[Y_i] + \frac{1}{2}\sigma_{Y_i}^2}$$

$$\times \Phi(r_i \sigma_{Y_i} - \Phi^{-1}(F_{S^l}(d)))$$

$$- d(1 - F_{S^l}(d)), \quad 0 < d < \infty \quad (20)$$

respectively. These results can be generalized. Indeed, in [33] a particular pattern of cash flows with mixed signs of the α_i is considered, whereas in [35] the case that some of the r_i are negative is dealt with. The quality of the upper and lower bound approximations is investigated in [2, 34, 36, 37].

Dependencies in a Non-Gaussian World

In the previous sections, we considered the problem of how to determine comonotonic lower and upper bounds for sums of r.v.'s, and we illustrated the technique for sums of lognormal r.v.'s. We showed that the results can be applied to evaluate risk measures of stochastic net present or accumulated value of future cash flows in a Gaussian world. However, it is well-known that daily returns are correlated and exhibit fat tails, and cannot be adequately modeled through normal r.v.'s. On the other hand, several of the applications we encountered concern long time investment horizons (typically some decades), and hence involve monthly or yearly returns. In this case, assuming a Gaussian model for the vector of cumulative investment returns $(Y_1, Y_2, \dots, Y_n)$ appears to be appropriate, see, for instance [35, 38].

The theoretical developments concerning the comonotonic lower and upper bounds continue to hold for non-Gaussian random vectors. The comonotonic upper bound can readily be applied in the general case. For sums of log-elliptical r.v.'s, we refer to [39]. The performance of the upper bound when Lévy processes are involved is investigated in [27] and [39].

The comonotonic lower bound results are more difficult to use for general distribution functions, mainly because closed form expressions for $E[X_i | \Lambda]$ are often not available. In [3], the lower bound based on the conditioning technique is investigated for sums consisting of a combination of lognormal and normal r.v.'s. The case of sums of log-elliptical r.v.'s is considered in [40].

References

- [1] Dhaene, J., Denuit, M., Goovaerts, M.J., Kaas, R. & Vyncke, D. (2002). The concept of comonotonicity in actuarial science and finance: theory, *Insurance Mathematics & Economics* **31**(1), 3–33.
- [2] Dhaene, J., Denuit, M., Goovaerts, M.J., Kaas, R. & Vyncke, D. (2002). The concept of comonotonicity in actuarial science and finance: applications, *Insurance Mathematics & Economics* **31**(2), 133–161.
- [3] Dhaene, J., Goovaerts, M.J., Lundin, M. & Vanduffel, S. (2005). Aggregating economic capital, *Belgian Actuarial Bulletin* **5**, 14–25.
- [4] Dhaene, J., Vanduffel, S., Goovaerts, M.J., Kaas, R. & Vyncke, D. (2005). Comonotonic approximations for optimal portfolio selection problems, *Journal of Risk and Insurance* **72**(2), 253–301.
- [5] Dhaene, J., Vanduffel, S., Tang, Q., Goovaerts, M.J., Kaas, R. & Vyncke, D. (2006). Risk measures and comonotonicity: a review, *Stochastic Models* **22**, 573–606.
- [6] Simon, S., Goovaerts, M.J. & Dhaene, J. (2000). An easy computable upper bound for the price of an arithmetic Asian option, *Insurance Mathematics & Economics* **26**(2–3), 175–184.
- [7] Panjer, H.H. (1981). Recursive evaluation of a family of compound distributions, *ASTIN Bulletin* **23**, 227–258.
- [8] De Pril, N. (1989). The aggregate claims distribution in the individual model with arbitrary positive claims, *ASTIN Bulletin* **19**, 9–24.
- [9] Dhaene, J., Ribas, C. & Vernic, R. (2006). Recursions for the individual risk model, *Acta Mathematica Applicatae Sinica, English Series* **22**(4), 543–564.
- [10] Hoeffding, W. (1940). Masstabinvariante Korrelations-theorie, *Schriften des Mathematischen Instituts und des Instituts für Angewandten Mathematik der Universität Berlin* **5**, 179–233.
- [11] Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont donnés, *Annales de l'Université de Lyon, Section A, Series* **3**(14), 53–77.
- [12] Dhaene, J., Wang, S., Young, V. & Goovaerts, M.J. (2000). Comonotonicity and maximal stop-loss premiums, *Mitteilungen der Schweiz. Aktuarvereinigung* **2000**(2), 99–113.

- [13] Kaas, R., Dhaene, J. & Goovaerts, M. (2000). Upper and lower bounds for sums of random variables, *Insurance Mathematics & Economics* **27**, 151–168.
- [14] Jamshidian, F. (1989). An exact bond option formula, *Journal of Finance* **XLIV**(1), 205–209.
- [15] Vasicek, O. (1977). An equilibrium characterization of the term structure, *Journal of Financial Economics* **5**, 177–188.
- [16] von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.
- [17] Yaari, M.E. (1987). The dual theory of choice under risk, *Econometrica* **55**, 95–115.
- [18] Wang, S. & Young, V. (1998). Ordering risks: expected utility versus Yaari's dual theory of choice under risk, *Insurance Mathematics & Economics* **22**, 145–162.
- [19] Shaked, M. & Shantikumar, J.G. (1997). *Stochastic Orders and their Applications*, Academic Press, New York.
- [20] Kaas, R., Goovaerts, M.J., Dhaene, J. & Denuit, M. (2001). *Modern Actuarial Risk Theory*, Kluwer Academic Publishers, pp. 328.
- [21] Denuit M., Dhaene J., Goovaerts M. & Kaas R. (2005). *Actuarial Theory for Dependent Risks—Measures, Orders and Models*. John Wiley & Sons, pp. 437.
- [22] Kaas, R., Dhaene, D., Vyncke, D., Goovaerts, M. & Denuit, M. (2002). A simple geometric proof that comonotonic risks have the convex-largest sum, *ASTIN Bulletin* **32**(1), 71–80.
- [23] Meilijson, I. & Nadas, A. (1979). Convex majorization with an application to the length of critical paths, *Journal of Applied Probability* **16**, 671–676.
- [24] Rüschendorf, L. (1983). Solution of statistical optimization problem by rearrangement methods, *Metrika* **30**, 55–61.
- [25] Müller, A. (1997). Stop-loss order for portfolios of dependent risks, *Insurance Mathematics & Economics* **21**, 219–223.
- [26] Dufresne, D. (1990). The distribution of a perpetuity with applications to risk theory and pension funding, *Scandinavian Actuarial Journal* **9**, 39–79.
- [27] Albrecher, H., Dhaene, J., Goovaerts, M. & Schoutens, W. (2005). Static hedging of Asian options under Lévy models: the comonotonicity approach, *Journal of Derivatives* **12**(3), 63–72.
- [28] Hobson, D., Laurence, P. & Wang, T.H. (2005). Static-arbitrage upper bounds for the prices of basket options, *Quantitative Finance* **5**(4), 329–342.
- [29] Vanmaele, M., Deelstra, G., Liinev, J., Dhaene, J. & Goovaerts, M.J. (2006). Bounds for the price of discrete Asian options, *Journal of Computational and Applied Mathematics* **185**(1), 51–90.
- [30] Reynaerts, H., Vanmaele, M., Dhaene, J. & Deelstra, G. (2006). Bounds for the price of a European-style Asian option in a binary tree model, *European Journal of Operational Research* **168**(2), 322–332.
- [31] Chen, X., Deelstra, G., Dhaene, J. & Vanmaele, M. (2007). *Static super-replicating strategies for a class of exotic options*, Research Report, Department of Applied Economics, K.U.Leuven.
- [32] Rogers, L.C.G. & Shi, Z. (1995). The value of an Asian option, *Journal of Applied Probability* **32**, 1077–1088.
- [33] Vanduffel, S., Chen, X., Dhaene, J., Goovaerts, M., Henrard, L. & Kaas, R. (2005). Optimal approximations for risk measures of sums of lognormals based on conditional expectations, *Journal of Computational and Applied Mathematics* **4**(1), 17–30.
- [34] Vanduffel, S., Chen, X., Dhaene, J., Goovaerts, M., Henrard, L. & Kaas, R. (2006). *A note on optimal lower bound approximations for risk measures of sums of lognormals*. Research Report, Department of Applied Economics, K.U.Leuven.
- [35] Cesari, R. & Cremonini, D. (2003). Benchmarking, portfolio insurance and technical analysis: a Monte Carlo comparison of dynamic strategies of asset allocation, *Journal of Economic Dynamics and Control* **27**(6), 987–1011.
- [36] Huang, H., Milevsky, M. & Wang, J. (2004). Ruined moments in your life: how good are the approximations? *Insurance Mathematics & Economics* **34**(3), 421–447.
- [37] Vanduffel, S., Hoedemakers, T. & Dhaene, J. (2005). Comparing approximations for sums of non-independent lognormal random variables, *North American Actuarial Journal* **9**(4), 71–82.
- [38] Levy, H. (2004). Asset return distributions and the investment horizon, *Journal of Portfolio Management* **2004**, 47–62.
- [39] Albrecher, H. & Predota, M. (2004). On Asian option pricing for NIG Lévy processes, *Journal of Computational and Applied Mathematics* **172**(1), 153–168.
- [40] Valdez, E. & Dhaene, J. (2003). Bounds for sums of dependent log-elliptical random variables, *7th International Congress on Insurance: Mathematics & Economics*. Lyon, France, June 25–27.

Further Reading

Deelstra, G., Diallo, I. & Vanmaele, M. Bounds for Asian basket options, *Journal of Computational and Applied Mathematics*, In Press.

JAN DHAENE, STEVEN VANDUFFEL AND MARC J. GOOVAERTS

Comparative Efficacy Trials (Phase III Studies)

A comparative efficacy trial (a phase III study) is a clinical trial (*see* **Randomized Controlled Trials**) that is designed to compare the efficacy (*see* **Efficacy**) of an experimental treatment (e.g., a therapy, preventive agent, or medical device, or any other type of intervention) with that of an active control treatment (or a placebo). Clinical trials to study new treatments are often classified into three successive phases. Phase I trials typically are small studies that are designed to study the safety and toxicity of the new treatment and to estimate the maximum tolerated dose (MTD) of a drug. Once a dose has been determined, preliminary assessments of therapeutic activity and further evaluations of safety are typically carried out in relatively small phase II trials. Phase II trials can include dose ranging studies to identify additional doses and schedules below the MTD that may also be efficacious. Such trials can provide information that can lead to the decision to compare the efficacy of an experimental treatment in large, expensive phase III trials with that of a standard/control treatment [1].

Phase III trials are definitive steps in the evaluation of experimental treatments. They play a pivotal role in the determination of the effectiveness (and/or the incidence and severity of adverse events) of an experimental treatment relative to a standard/control treatment. These trials are usually the third and final step in the test of an experimental treatment, prior to submission of an investigational new drug application (INDA) for approval by relevant governing authority such as the Food and Drug Administration (FDA) of the United States or other relevant governing agencies such as the European Agency for the Evaluation of Medicines.

Phase III trials are, of necessity, often large, and costly, and can require many years to complete. The design of such trials should consider the feasibility of conducting the study in a timely manner. Ongoing monitoring of patients during these trials is required to identify any risks to patients that may be associated with the experimental treatment, as well as to minimize risks of denying patients a treatment that may be truly effective or of exposing patients to an ineffective treatment. Finally, statistical analysis that

incorporates the study design (fixed or sequential), potential patient dropout or noncompliance to treatment, and other issues, is critical to avoid various biases and errors (*see* **Compliance with Treatment Allocation**). Thus, appropriate design, conduct and monitoring, and careful statistical analysis are all important requirements for the eventual success of these important Phase III clinical trials.

This article provides a brief introduction to design and analysis of Phase III trials. The issues discussed include appropriate study designs and patient populations for study, randomization and blinding, selection of efficacy endpoints, the risks of type I and II errors, the determination of sample size, multicenter trials, interim analysis, and the assessment and comparison of treatment effects between/among different treatments.

Design of Trials

In this section, we discuss some common issues that are key to the design of appropriate phase III clinical trials. We restrict our discussion to controlled clinical trials that compare a new treatment or treatments or combinations of treatments or interventions to standard treatment(s) or placebo, or in some circumstances, usual care or observation only.

Randomization and Blinding

In phase III trials, subjects are usually randomized to the treatment arms to avoid selection bias. Selection bias may affect the validity of the comparative study. The randomization may be simple or blocked, and may or may not be carried out in strata. The classic example of a large randomized comparative efficacy trial (phase III) is the Salk vaccine trial of 1954. Over 1 million children participated in the trial. This trial was conducted to assess the effectiveness of the Salk vaccine to provide protection against paralysis or death from poliomyelitis as described in detail by Meier [2]. In one part of this study, 401 973 children were randomly assigned either to injection with the Salk vaccine or with a placebo salt solution. The trial was double blind, that is, neither the children nor the diagnosing physicians were aware of who had received the Salk vaccine or the placebo. Patients and/or their physicians who participate in a study are often blinded to the treatment assignment to reduce

the likelihood of treatment-related bias because of the knowledge of treatment assignment (*see Blinding*). The Salk vaccine trial was conducted in a relatively short period of time, and the endpoints were observed within this relatively short time period. In trials with long accrual times and/or long observation times, to ensure that the effect of an experimental treatment is measured against a comparison treatment administered over the same time period and under similar conditions, a blocked randomization procedure is often used.

Selection of Endpoints

Definitive or surrogate endpoints can be used to assess the effect of the treatment or intervention. If a surrogate endpoint or marker is used, this marker should be correlated with the clinical efficacy endpoint (e.g., survival) and should capture the effect of the treatment or intervention on the definitive clinical endpoint so that it becomes a reliable replacement [3]. For example, in HIV studies, CD4+ counts, which were correlated with the disease, were used as surrogate endpoints for mortality in early AIDS trials; however, when long term follow-up became available, CD4+ counts were not predictive of mortality. Viral load, causally related to outcome, was identified as a better potential surrogate endpoint [4]. In recent studies to examine the efficacy of vaccines for human papilloma virus (HPV) for the prevention of cervical cancer, prevention of infection from HPV was accepted by regulatory authorities as a surrogate for cervical cancer [5]. Thus, the trials were conducted relatively rapidly rather than requiring observation of vaccinated subjects for the development of cervical cancer, a long term endpoint. The association between cervical cancer and HPV has been found to be unique and this link enabled the development of a prophylactic vaccine for cervical cancer [6].

Superiority and Equivalence Trials

Comparative efficacy trials (phase III studies) may be further classified by their primary objectives. These include phase III superiority trials to demonstrate superiority of a new experimental treatment in terms of improved efficacy compared to the efficacy of an active control treatment. Note that, in general, all trials designed to compare a new intervention to a placebo, would be superiority trials. Noninferiority

trials (or equivalence trials), on the other hand, are designed with the objective of showing that the experimental treatment is clinically not much worse than the active control treatment by a specified margin (*see Inferiority and Superiority Trials*). Investigators may decide to conduct a noninferiority trial when they believe that the efficacy of the active control cannot be surpassed, but that the experimental treatment has similar efficacy and may offer important safety and/or cost advantages.

It is common to specify a threshold for the margin of indifference when comparing the treatment effects between the experimental treatment and the control treatment in a noninferiority trial. For example, when comparing the proportion of success in the experimental treatment (P_E) with that of the control treatment (P_C), the noninferiority margin (or the margin of indifference), δ , is often defined as the maximum difference amount by which the experimental treatment is not considered clinically much worse than the control treatment if $P_E > P_C - \delta$. The noninferiority margin δ , should be prespecified and determined on the basis of clinical judgment [7–9].

Let μ_E and μ_C represent the mean value of the efficacy variable for the experimental and control treatment, respectively. Further, suppose that positive values of the parameter of interest (or values greater than one, in case of a ratio) indicate superiority of the standard treatment. The standard null and alternative hypotheses for testing noninferiority are,

$$\begin{aligned} H_0(-\delta) : \mu_C - \mu_E &\geq \delta \quad \text{versus} \\ H_1(-\delta) : \mu_C - \mu_E &< \delta \end{aligned} \quad (1)$$

where $\delta > 0$ is the noninferiority margin, that is, the amount by which μ_C can exceed μ_E with the experimental treatment still being considered noninferior to the active control. The null hypothesis states that the active control μ_C exceeds the experimental treatment μ_E by at least δ ; if this cannot be rejected, then the active control is considered superior to the experimental treatment with respect to efficacy. The alternative hypothesis states that the active control may indeed have better efficacy than the experimental treatment, but by no more than δ . In such a case, we say the experimental treatment is not inferior to or no worse than the active control. Thus, rejection of the null hypothesis

concludes noninferiority. For superiority trials, the hypotheses can be formulated as above by letting $\delta = 0$ in equation (1).

Power and Sample-Size Analysis

In the design of phase III clinical trials, in addition to selecting the appropriate randomization procedures and/or blinding methods, one must also guard against the risk of committing any of the two common types of errors in typical hypothesis-testing-based trial designs. A type I error is made if the study leads to a false rejection of the null hypothesis H_0 , and a type II error is made if the study fails to establish the alternative hypothesis H_1 when it is indeed true. The probabilities of type I and type II errors are often denoted by α and β , respectively. The statistical power of a test is defined as the probability $1 - \beta$.

In the design of a phase III trial, the investigators often have some idea of the order of magnitude of the efficacy difference that they consider clinically meaningful. A difference, Δ , between the projected efficacy for the experimental treatment and that for the control treatment, that is worth detecting is often specified [10]. The primary efficacy endpoints often have normal distributions. The sample size needed to compare the means of two normal distributions with equal and known variances σ^2 , using a one-sided z -test, can be obtained as

$$n_C = \frac{\sigma^2(z_\alpha + z_\beta)^2[1 + 1/k]}{\Delta^2}, \quad n_E = kn_C \quad (2)$$

where one sample (n_E) is k times as large as the other sample (n_C), σ^2 is the common variance and Δ is the minimum meaningful difference between the means (say, μ_C and μ_E) to be detected, and z_α and z_β are the upper percentiles of the standard normal distribution corresponding to α and β . Additional details on sample-size formulas for various efficacy endpoints can be found in many sources [1, 9, 11, 12]. Similarly, the sample-size estimation procedure for the design of noninferiority trials can be found in many textbooks and classical papers [1, 9, 13, 14].

Multicenter Trials

Phase III trials, in general, are large, and, of necessity, are usually conducted at multiple sites. Multicenter trials can yield results that could be applicable to

a wider range of treatment settings or patient groups than a single center trial [15]. Such trials should allow the investigators to estimate an overall treatment effect in the patient populations under study at the various sites. The multicenter trial, while generally expedient, incorporates some particular difficulties in the conduct and implementation that require attention in both the design and analysis stages. In particular, critical for the successful completion of a multicenter trial, are the development of the protocol that specifies accrual requirements from each center, details of patient management, criteria for patient entry and randomization, details of treatment regimens, criteria for patient evaluation and for alterations in planned patient management, procedures to ensure consistency across centers, and definitions of response and endpoints along with documentation requirements.

Analysis of Trial Outcomes

The approach to the comparison of different treatment effects depends on the primary outcome measurement. The comparison in the case of an event, such as death or recurrence of disease, is based on the event rate for different treatment groups. In case of a continuous outcome, such as weight or blood pressure, the change from baseline to some defined point after enrollment is determined for each subject and then summarized using means or medians. The treatment effect is estimated as the difference between the change in mean or median from the baseline for the control and experimental groups.

Frequently, a formal hypothesis test is conducted to compare the efficacy of the competing treatments based on data from the trial. In practice, a two-sided test is often recommended, even though the investigators are often only interested in a one-sided alternative hypothesis [10]. The efficacy parameter can be estimated and its confidence interval obtained for each treatment arm. The efficacy differences among treatment groups can be evaluated using appropriate measures such as, the difference in response rates, hazard rates, means or median survival times, etc., along with corresponding confidence intervals [1, 10, 12].

Statistical analysis that appropriately incorporates the planned trial design (fixed or sequential designs), data monitoring and interim analysis results, patient dropout and noncompliance to treatments, misclassification errors [10, 16], and other issues is essential

to the final success of a clinical trial. The subsequent discussion focuses only on a few of the most common issues that arise in the analysis of clinical trials.

Sequential Designs and Interim Analysis

Trials are either designed with a fixed or sequential/group sequential sample size. Most phase III trials use fixed sample size, i.e., the sample size is specified at the start of the study. In sequential trials, enrollment and observation continues until a stopping boundary is crossed. Inference about the treatment effect in sequential designs has to consider the fact that the stopping time of the trial is random, in order to avoid bias. Methods for the design and analysis of phase III trials using group sequential procedures with interim analysis are described in Proschan et al [17]. The two most common approaches are (a) O'Brien–Fleming boundaries that permit early stopping for success with very stringent α -level requirements and maintain almost all of the α level for the final planned look [18], and (b) Pocock boundaries that divide the α level equally among the planned looks, making it easier to stop the trial early [19]. These approaches fall under the class of Lan–DeMets spending function approach with different parameterizations [20].

Repeated Measures Analyses

Repeated measures data arise when multiple measurements of response variables and/or covariates are obtained from each experimental unit over the duration of the clinical trial. The repeated measures design (e.g., the crossover design) provides opportunities for investigators to examine the pattern of response variable under different measurement occasions, to understand the relationship between repeated outcomes and treatments as well as covariates, and to discern the dependence pattern among the repeated measurements. A special feature of repeated measures data is that the measurements from the same unit are usually positively correlated; thus, valid and effective analysis of repeated measures must incorporate this within-subject correlation (*see Repeated Measures Analyses*).

Analysis of Multicenter Trials

In multicenter trials, randomization is usually carried out within centers (strata) to ensure balance among

treatment groups within centers and to take into account the subtle differences among centers with respect to interpretation and implementation of the protocol, equipment, supportive care, and patient populations. As a consequence, the final analysis will generally incorporate the center effect as a fixed effect (weighted by center or unweighted) or as a random effect [15]. The discussion of the implication of unequal center sizes on the treatment-by-center interaction term in the regression modeling framework can be found in Gallo [21]. In this sense, the multicenter trial analysis can be considered to be a meta-analysis that combines results from different centers [22].

Missing Data and Noncompliance

Patient dropout and/or noncompliance to assigned treatments are common in large clinical trials with differing efficacy, and/or side effects for the treatments under study, as well as in trials with long durations. For example, survival time data arise frequently in many clinical trials, where incomplete or censored observations are common due to multiple factors. Statistical methods and models for survival analysis such as Cox's semiparametric proportional hazards models are widely used. Per-protocol based analysis may be compared to per-randomization based analysis to avoid biased analysis and interpretation of the data [1, 12].

This article provides a brief discussion on the design and analysis of phase III clinical trials. An in depth discussion of various issues is beyond the scope of this short review. Fortunately, there are a large number of monographs and textbooks about phase III trials. Interested readers can consult these useful reference books.

References

- [1] Friedman, L.M., Furberg, C.D. & DeMets, D.L. (1998). *Fundamental of Clinical Trials*, 3rd Edition, Springer, New York.
- [2] Meier, P. (1989). The biggest public health experiment ever: the 1954 field trial of the Salk poliomyelitis vaccine, *Statistics: A Guide to the Unknown*, 3rd Edition, J.M. Tanur, F. Mosteller, W.H. Kruskal, E.L. Lehmann, R.F. Link, R.S. Pieters & G.R. Rising, eds, Duxbury, California.
- [3] Prentice, R.L. (1994). Surrogate endpoints in clinical trials: definition and operational criteria, *Statistics in Medicine* **8**, 431–440.

- [4] Fleming, T.R. & DeMets, D.K. (1996). Surrogate end points in clinical trials: are we being misled? *Annals of Internal Medicine* **7**, 605–613.
- [5] FDA Vaccine and Related Biological Products Advisory Committee (2001). November 28–29.
- [6] Harper, D.M., Franco, E.L., Wheeler, C.M., Moscicki, A.-B., Romanowski, B., Rotel-Martins, C.M., Jenkins, D., Schuind, A., Clemmens, S.A.C. & Dubin, G. (2006). Sustained efficacy up to 4.5 years of a bivalent L1 virus-like particle vaccine against human papilloma virus types 16 and 18: follow-up from a randomized control trial, *Lancet* **367**, 1247–1255.
- [7] International Conference on Harmonization (1998). E9: guidance on statistical principles for clinical trials, *Federal Register* **63**(179), 49583–49598.
- [8] European Agency for the Evaluation of Medicinal Products, Committee for Proprietary Medicinal Products (2005). *Guideline on the Choice of the Non-Inferiority Margin*, European Medicines Agency (EMA), at <http://www.emea.eu.int/pdfs/human/ewp/215899en.pdf>.
- [9] Chow, S.-C. & Liu, J.-P. (2003). *Design and Analysis of Clinical Trials: Concepts and Methodologies*, 2nd Edition, John Wiley & Sons.
- [10] Fleiss, J.L., Levin, B. & Paik, M.C. (2003). *Statistical Methods for Rates and Proportions*, 3rd Edition, John Wiley & Sons.
- [11] Chow, S.-C., Shao, J. & Wang, H. (2003). *Sample Size Calculation in Clinical Research*, Marcel Dekker, New York.
- [12] Piantadosi, S. (2005). *Clinical Trials: A Methodologic Perspective*, 2nd Edition, John Wiley & Sons.
- [13] Blackwelder, W.C. (1982). “Proving the null hypothesis” in clinical trials, *Controlled Clinical Trials* **3**, 345–353.
- [14] Farrington, C.P. & Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk, *Statistics in Medicine* **9**, 1447–1454.
- [15] Goldberg, J.D. & Koury, K.J. (1989). Design and analysis of multicenter trials, in *Statistical Methodology in the Pharmaceutical Sciences*, D.A. Berry, ed, Marcel Dekker, New York, Chapter 7, pp. 201–237.
- [16] Goldberg, J.D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table, *Journal of American Statistical Association* **70**, 561–567.
- [17] Proschan, M.A., Lan, G.K.K. & Wittes, J.T. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*, Springer, New York.
- [18] O’Brien, P.C. & Fleming, T.R. (1979). A multiple testing procedure for clinical trials, *Biometrics* **35**, 549–556.
- [19] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*, John Wiley & Sons, New York.
- [20] DeMets, D.L. & Lan, K.K.G. (1994). Interim analysis: the alpha spending function approach, *Statistics in Medicine* **13**, 1341–1352.
- [21] Gallo, P. (1998). Practical issues in linear models analyses in multi-center trials, *Biopharmaceutical Report* **6**, 1–9.
- [22] Friedman, H. & Goldberg, J.D. (1996). Meta-analysis: an introduction and point of view, *Hematology* **23**, 917–928.

Related Articles

Blinding

Intent-to-Treat Principle

JUDITH D. GOLDBERG AND YONGZHAO SHAO

Comparative Risk Assessment

In its most general sense, comparative risk assessment (CRA) is an analytical process of (a) defining a type of undesired outcome, e.g., human mortality, (b) selecting a location, e.g., worldwide, (c) gathering information on factors relevant to that outcome, e.g., infectious disease and environmental toxins, and (d) ranking these factors in terms of their relative importance or contributions to the undesired outcome. Importance can be ranked either in terms of the expected harm or the potential for reduction in harm from controlling the different sources being compared. The term *comparative risk assessment* is more frequently used in the context of environmental risk (see **Environmental Health Risk**), but the principles and methodology are applicable in other contexts.

CRA is both a scientific and a political process. Scientific methods are used to gather data, classify the type and severity of risks, and identify and verify relationships between outcomes and risk factors. CRA is not intended as an academic or purely scientific exercise. It is a guide to decision making and, therefore, also involves the political functions of defining the scope of the assessment, determining the stakeholders and involving them in the process, and putting the results into a context that can be effectively communicated and used to develop consensus concerning appropriate action (see **Role of Risk Communication in a Comprehensive Risk Management Approach; Stakeholder Participation in Risk Management Decision Making**).

This article will review the history of CRA for environmental risk, discuss methods of risk ranking used in environmental risk, review methods of communicating risk, look at a few examples of CRA in nonenvironmental areas, and touch on United States Environmental Protection Agency (USEPA) and other regulations, as they relate to CRA.

Comparative Risk Assessment in Environmental Risk

History of Comparative Risk Assessment in the United States

Formalized CRA is a relatively young field, with documented uses by public agencies beginning in

the 1980s. A brief timeline on CRA milestones is as follows.

In 1981, the USEPA used comparison of the rate of occupational injury in retail trades, i.e., a safe occupation, to help set the allowable level of occupational exposure to radiation [1].

In 1983, the US National Academy of Sciences published the study “Risk Assessment in the Federal Government: Managing the Process” [2]. The focus of this work was to breach the gap between scientific information and public policy as they relate to the potential health effects of toxic substances, with primary focus on increased risk of cancer due to chemicals in the environment. The report recommended that a clear distinction between risk assessment and risk-management policy be maintained and that explicit guidelines for the process of risk assessment be published.

In 1985, the US Occupational Safety and Health Administration (OSHA) used the comparison of rates of job-related illness for firefighters, miners, and other high-risk occupations and rates for low risk occupations such as service employment to help determine if the estimated risk associated with benzene exposure (see **Benzene**) was sufficiently high to be considered “significant” [3].

The US Environmental Protection Agency Task Force was established in 1986. This task force published “Unfinished Business: A Comparative Assessment of Environmental Problems – Overview Report” [4–8]. The concept of comparative ranking of risk to human health was clearly established in this work. A team of 31 EPA scientists and managers ranked 31 environmental problems in terms of: (a) human cancer risk, (b) human noncancer health risk, (c) ecological risk, and (d) welfare risk. This report noted that public perception of risk does not always match risk ratings of scientists, and that public policy is frequently driven by public perception of risk. The US Science Advisory Board responded to “Unfinished Business” with “Reducing Risk: Setting Priorities and Strategies for Environmental Protection” in which they recommended broad public participation in the ranking of risks, emphasizing the uncertainty in many scientific risk assessments.

In 1993, the Carnegie Commission on Science Technology and Government published “Risk and the Environment: Improving Regulatory Decision Making” [9], recommending placing problems in broad risk categories, developing multidimensional

risk rankings, and finding methods to integrate societal values into relative risk analyses. The use of relative or comparative risk rankings was also recommended in the 1995 Harvard Center for Risk Analysis report, "Reform of Risk Regulation: Achieving More Protection at Less Cost" [10].

By 1997, the Presidential/Congressional Commission on Risk Assessment and Risk Management was explicitly recommending that "(US government) agencies should use a comparative risk assessment approach for risk management on an experimental or demonstration basis to test the effectiveness of seeking consensus on setting priorities for environmental, health, and safety hazards" [11].

Examples of Comparative Risk Analysis Projects

CRA projects range from very general to very specific. The examples listed here are by no means exhaustive. We start with further detail on the groundbreaking "Unfinished Business", followed by an example so general as to tackle worldwide health risk, and end with examples specific to a type of risk (cancer) and to a small number of risk factors. Further narrowing of the scope is illustrated by examples specific to a product or specific to a location.

A widely cited example of environmental comparative risk analysis is the study by the USEPA of 31 environmental risk factors documented in the 1987 report "Unfinished Business: A Comparative Assessment of Environmental Problems – Overview Report". Working groups were established to evaluate and rank 31 specified environmental problems in each of the four risk categories: (a) human cancer risk, (b) human noncancer risk (c) ecological risk, and (d) welfare risk. The concept of a risk ranking method in which the concerned public and other stakeholders as well as scientists were engaged in the ranking process was established here as well.

The cancer risk group ranked worker exposure to chemicals (formaldehyde, tetrachloroethylene, asbestos, and methyl chloride) and indoor radon (see **Radon**) as the highest environmental problems, followed by pesticide residues on foods (herbicide, fungicide, insecticide, and growth regulator), indoor air pollutants other than radon (tobacco smoke, benzene, *p*-dichlorobenzene, chloroform, carbon tetrachloride, tetrachloroethylene, and trichloroethylene), and consumer exposure to chemicals (formaldehyde, methyl chloride, *p*-dichlorobenzene, and asbestos).

The ecological risk group gave *stratospheric ozone depletion* (see **Stratospheric Ozone Depletion**) and CO₂ the highest rank followed by physical alteration of aquatic habitats, mining, gas and oil extraction, processing, and wastes.

The World Health Organization's 2002 World Health Report is an example of a comparative risk analysis at the most general level. Risks evaluated include: malnutrition and underweight, unsafe sex, high blood pressure, tobacco, alcohol, unsafe water and sanitation, high cholesterol, indoor smoke from solid fuels, iron deficiency, overweight, zinc deficiency, low fruit and vegetable intake, vitamin A deficiency, physical inactivity, and many others. Risks were ranked according to estimated disability-adjusted life years (DALYs) lost attributable to the risk factors. The global exposure to these risks and the relationship of these exposures to poverty and socioeconomic status was evaluated. Rankings were produced for developing and developed countries, and show that a relatively small number of risk factors "cause a huge number of premature deaths and account for a very large share of the global burden of disease" [12].

An analysis of the causes of cancer in the world is an example of a CRA of nine behavioral and environmental risk factors [13], where the scope of the analysis was narrowed with a detailed follow-up of risk factors identified and ranked in a larger, more global analysis. Mortality from 12 types of cancer was evaluated with a focus on potentially modifiable risk factors. The authors judged that 35% of worldwide cancer deaths could be attributed to the nine factors: smoking, alcohol use, low fruit and vegetable consumption, overweight and obesity, physical inactivity, sexual transmission of human papilloma virus, urban air pollution, indoor smoke from household use of solid fuels, and contaminated injections. These risks were ranked as the nine most important cancer risks that could be mitigated or removed.

An example of using CRA to evaluate a very specific health risk is the US Food and Drug Administration's estimate of the risk of *Escherichia coli* O157:H7 per serving of tenderized beef as compared to nontenderized beef [14] under various cooking conditions. A probability model for estimating the transmission of *E. coli* from cattle to consumer was used to find essentially the same risk of illness per serving of tenderized beef as for nontenderized beef.

An example of comparative risk analysis with a very specific location is “Chelsea Creek Community Based Comparative Risk Assessment” [15] in which five community organizations and the USEPA used CRA to provide information to improve public health for residents of the area. Community surveys were used to identify 15 top issues ranging from air pollution, trash, and water quality to housing and health care insurance.

Comparative Risk Assessment in Practice

The following sections describe details of the CRA method. While developed in relation to environmental risks, this methodology is general enough to apply to any type of risk assessment. The steps of the methodology as well as methods of communication of the results are described.

Components of Risk

Any CRA must start with a specification of the types of risks to be studied. Historically, environmental risk assessments focused upon human cancer risk associated with toxic chemicals. Other risks considered are human noncancer risk, ecological risk, and quality of life risk (or welfare effects). Both human cancer risk and human noncancer risk can be measured in terms of number of deaths per year. This can be refined as DALYs lost, giving higher value to a life lost at an early age. Ecological risk is measured by changes in basic characteristics of the ecosystem as a whole, and this assessment will depend upon the ecosystem being considered. Welfare or quality of life risk is measured as decline in economic value or value of other human activity other than health-related decline.

Risk Ranking

Once all relevant risks have been identified, the risks are ranked from the most important to the least important. The ranking may be done by discussion or other group dynamic processes leading to consensus; this could include taking of straw polls and additional discussion until the group decision becomes stable. For larger groups or groups that do not meet face to face, a simple majority rule vote could be used to rank the risk. For example, each person is given an equal number of votes to be allocated among the

various risks. The risk receiving the largest number of votes is given the highest rank and so forth. For risk assessments for which there is sufficient information concerning both the rate of death, illness, or injury per unit of exposure, as well as data on the distribution of the amount of exposure, a formula can be used to rank the risk factors. Examples of formulae used in risk ranking include estimated annual number of cancer deaths, estimated annual number of serious or fatal injuries, and estimated DALY lost.

Whichever method is used for the risk ranking, the uncertainty in the estimates of risk per unit of exposure and in the estimate of amount of exposure should be considered in the assignment of ranks. This can be done informally in the consensus or voting process or formally by calculating confidence intervals around the risk estimate calculated using formulae.

There were four EPA working groups, one for each component of risk considered in “Unfinished Business”. The ranking methods used by these four groups illustrate approaches to the ranking process.

The cancer risk group at first used available data to partition the problems to be considered into five broad categories of importance. This was followed by ordinal ranking within each category. Ordinal rankings were refined by an iterative consensus process.

The ecological risk working group noted that a quantitative model-based approach to ranking works for well-defined problems, but is less useful for broadly defined problems. This working group started by defining four general categories of ecosystems and evaluating how the 31 environmental problems identified by the USEPA are related to the four ecosystem categories; nine of the 31 environmental problems were dropped from further study in the first screening. The group then formed a preliminary rating of the remaining 22 environmental problems within each of the ecosystem categories using a three-point (high, medium, and low) scale. A background paper was prepared for each of the 22 problems; each workgroup member used this information to rank the problems. Consensus was reached by a final classification of the problems as high, medium, or low without further ranking within these levels of importance.

One primary value of consensus ranking is that the fact of consensus in the ranking of risks may generate the political will to go to the next step and begin to

take action. However, as in any group decision process, the group consensus need not generate optimal or even good decisions. Although risk ranking may consider the likelihood of actions for risk reduction during the ranking process, it may not be possible to adequately capture the importance of budget constraints in the ranking process. This is particularly true when some of the possible risk reduction actions are mutually exclusive or when possible expenditures of funds are not continuous. For example, consider the situation in which action A is expected to save 10 lives per \$ million expended, but at most 30 lives could be saved (\$3 million total cost), while action B could save only 8 lives per \$ million spent but could save 40 lives (\$5 million total cost). If a total of \$5 million is available, and it is not possible to spend the first \$3 million on action A and the remaining \$2 million on a partial implementation of action B, then action B, which has the higher cost per life saved, may be optimal.

Project Teams

CRA is distinct from the risk-management process; it is generally intended to be a source of information for the subsequent risk-management decisions. It is often recommended that the project team include representatives of all of the stakeholders. Team members may include scientists, members of regulatory agencies, public health officials, members of planning commissions, representatives of business or citizens groups, and other interested individuals. This process may take longer than simply producing scientific estimates, but it is likely to be more efficient and to actually provide the type of information that will lead to decisions to take action to mitigate risk. Even in those situations in which formulae are to be used to create the rankings, the group process is used to determine the formulae that are used for the ranking.

Communication through Comparison to Known Risks

Every product and every human activity carries with it some level of risk. The concepts of CRA can be used to communicate the importance of a risk of interest by comparing the level of risk for similar products or activities. Some of the earliest uses of CRA were of this type. As previously noted, in 1981 the USEPA used comparison to familiar risks

to set levels for allowable occupational exposure to radiation, and in 1985 the US Occupational Safety and Health Administration used comparison of injury levels in high-risk jobs to conclude that benzene exposure was a significant risk.

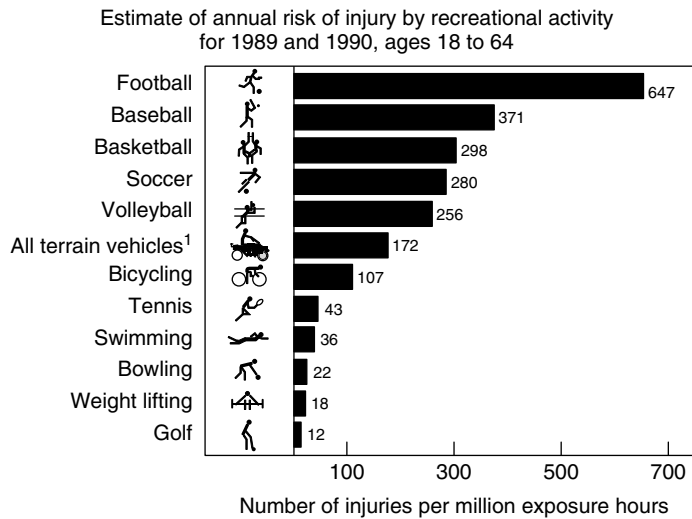
This type of comparison is most useful if the risks to be compared can be put on the same scale, e.g., risk of injury per million exposure hours. The chart in Figure 1 shows the risk of injury for various recreational activities in the United States.

The US Consumer Product Safety Commission uses this type of approach. For example, the risk of injury for skateboards was put into context by comparing it to the injury risk associated with roller skates, ice skates, and similar types of products [16]; the injury risk associated with chain saws was compared to injury risk associated with walk-behind power mowers [17]; the risk of drowning in a home swimming pool for children under age five was compared to the risk of death for children under age five in spas, motor vehicle crashes, and all home accidents [18].

The comparative risk method can also be used to evaluate the relative safety of competing products and to address questions of unreasonable product risk. For example, the year 1993–1995 Honda Civic Coupe model was alleged to have an unreasonably high risk of injury for occupants involved in side collisions [19]. Data from police reported side collision crashes was used to make a CRA of the risk of injury in side collisions for occupants of the Honda Civic Coupe and occupants of other passenger vehicles of similar age and size. This information was useful to the jury in deciding that the Honda Civic was not defective (see Figure 2).

Comparative Risk Assessment in Nonenvironmental Areas

While the use of CRA, particularly the group or team approach, has seen wide use in environmental issues, the basic concept of comparing different risks for the purpose of making risk-management decisions is applicable in any situation in which relevant data can be collected and there is interest in making informed choices. The value of comparative risk is perhaps easier to appreciate when the area for risk assessment is narrowly defined and the appropriate risk-management decision is straightforward. A few examples on nonenvironmental risk assessment follow.



¹Exposure hours computed for operators from "ATV owning households" only.

Figure 1 Comparative risk assessment of injury for recreational activities. *Notes:* Risk is the number of injuries per million exposure hours. The ATV risk estimate is based on data from 1989. All other activities have risk estimates based on from 1990. All injury estimates reflect the US Government's revision of the hospital emergency room sampling frame in 1990

CRA of Hazmat and Non-Hazmat Truck Shipments [20] used the following CRA method: (a) arrange Hazmat classes into categories, (b) estimate incident likelihood for each category, (c) estimate per incident economic impact, (d) calculate annual risk, and (e) estimate annual risk for non-Hazmat categories and compare. They found the cost of Hazmat transport accidents to range from \$0.05 to \$0.42 per mile. In comparison, non-Hazmat transport accident cost was estimated at \$0.25 per mile.

The US Department of Transportation Federal Railroad Administration used CRA to evaluate the risk of conventional switching operations with remote control switching [21]. A human reliability assessment of potential operator errors was used to compare the risks associated with the two switching methods.

The FDA/Center for Food Safety and Applied Nutrition in conjunction with the USDA/Food Safety and Inspection Service, and the US Centers for Disease Control and Prevention conducted a CRA of 23 ready-to-eat food categories as they relate to the foodborne illness caused by the bacterium *Listeria monocytogenes* [22]. This bacterium causes serious illness and fatality in high-risk groups of people such

as the elderly and persons with compromised immune systems. Two dimensions of risk were considered in the ranking process: (a) risk of illness per serving and (b) annual number of cases in the United States. A cluster analysis algorithm was used to classify each of the 23 food groups into five risk categories. The very high-risk category (e.g., deli meats) is high on both dimensions of risk, has high rates of contamination, and supports rapid growth of the bacterium under refrigeration. The high-risk category (e.g., pate and meat spreads) is high on either risk per serving or number of cases per annum and supports growth of the bacterium during extended refrigerated storage. The moderate risk category (e.g., fruits and vegetables) contains foods that are of medium risk both per serving and number of cases per year. For these products, the risk is primarily associated with recontamination. The low risk category (e.g., preserved fish) contains foods that have moderate contamination rate, but include conditions such as short shelf life that are expected to reduce the risk of bacterial growth. The very low risk category (e.g., hard cheese) includes products that have low contamination rates and characteristics that either inactivates or prevents the growth of the bacteria.

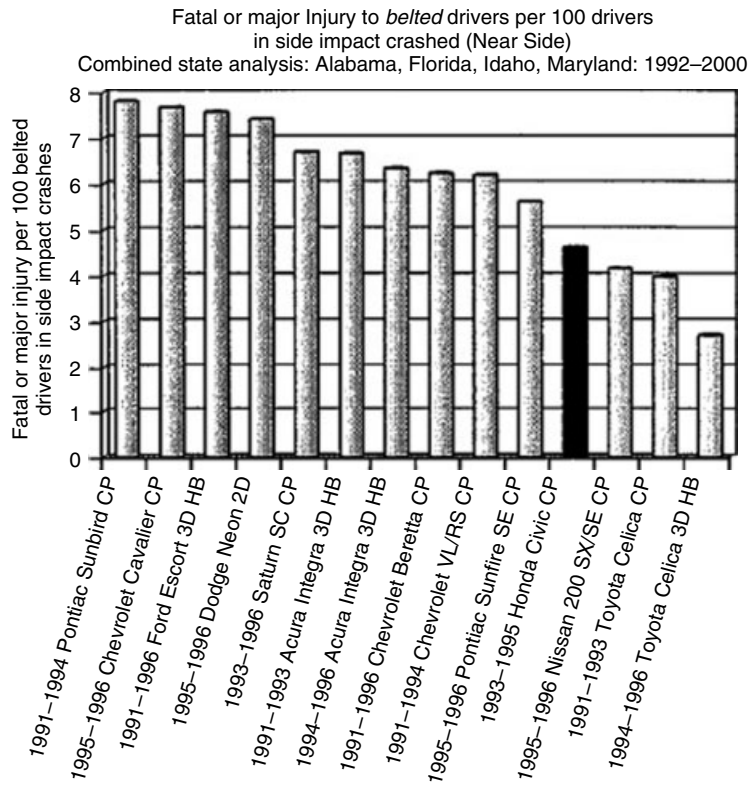


Figure 2 Comparative risk assessment of crashworthiness. *Source:* Combined state analysis: Alabama 1992–2000, Florida 1992–2000, Maryland 1992–2000; passenger car model, years 1991–1996

This is an excellent example of the use of a statistical method (cluster analysis) to handle multiple dimensions of risk to produce a meaningful risk ranking algorithm.

Comparative Risk Assessment Issues

Criticisms of the Comparative Risk Assessment Process

Several specific criticisms of the CRA process are discussed here.

Most risks are by nature multidimensional, and direct comparison is not meaningful, i.e., apples cannot be compared with oranges [23, 24]. Finkel [25] recommends explicit recognition of the multiple value systems that can be used to rank risks and delineation of the common aspects of the risks during the ranking process. For example, apples and oranges can be compared on price per pound, calories per

serving, and many other factors. Humans regularly make comparisons between disparate situations, e.g., “Should we go to a movie or spend the evening at a fine restaurant?” Consequently, the multivariate nature of many risks is not necessarily a bar to meaningful comparisons. Kadvany [24] recommends the formalization of this comparison through the use of multiattribute utility theory, a method of aggregating multiple dimensions into a single dimension by assigning weights to each dimension and taking a weighted average, or other function that combines the multiple dimensions. The difficulty is, of course, finding an appropriate weighting scheme so that the linear function is a reasonable representation of the utility of reducing the risk, or in cases in which no linear combination is appropriate identifying a function that both combines information from all of the dimensions and is useful for comparing the overall magnitude of the risk. (See the previous example for a discussion of the use of cluster analysis as a

means of combining the different dimensions of risk of foodborne illness [24].) The fact that CRA projects frequently rank the risks through voting or group consensus methods is an informal way of integrating the multiple components of risk.

Information is Frequently Incomplete. Some level of uncertainty is inherent in any risk analysis and in most decision making processes. Kavadhan [24] recommends distinguishing between uncertainty, i.e., incomplete knowledge, and variability, i.e., the risk factor has differential effects in different circumstances. Rather than reporting only point estimates of risk (e.g., five deaths per million exposed persons per year) report confidence bounds, (e.g., with 95% confidence the risk is between one and nine deaths per million). In this way, ranking of risk can be based upon both the estimated level of risk and the uncertainty associated with the risk.

The complexity of scientific methods for estimating the magnitude of various risks makes it difficult for the concerned public to evaluate the validity of the estimates [23]. The method of risk ranking based upon voting or group dynamic processes mitigates this criticism. The scientifically based numerical estimates of the level of risk provide part of the raw material for risk ranking. If representatives of all stakeholders are included in the process, the end result will not be dominated by the mathematical estimates of risk.

Risk Ranking does not go Far Enough. Risks should be ranked not according to the highest level of risks, but according to the highest level of preventable risks or according to the most cost-effective reduction of risks [24]. There is nothing in the concept of ranking risk that prevents consideration of the cost of reducing the risk as one of the attributes in the ranking process. As noted by Gutenson [26], comparative risk processes can lose their “distinct identity” and become simply a state or locally driven risk-management process.

Comparative Risk in the United States and Europe

The comparative risk process involving ranking of risks by a team of scientists, community members, business interests, and other stakeholders has been

recommended by the USEPA since the 1987 “Unfinished Business” report outlined the method.

In 2006, the US Office of Management and Budget (OMB) [27], in consultation with the Office of Science and Technology, proposed to issue new guidance on risk assessments in the form of a Risk Assessment Bulletin. The purpose of the bulletin is to “enhance the technical quality of and objectivity of risk assessments prepared by federal agencies by establishing uniform minimum standards”.

The proposed bulletin recommends making the purpose of the assessment clear by distinguishing between screening level risk assessments and comprehensive risk assessments. Five goals are enumerated: (a) the project problem should be formulated through iterative dialogue with the decision makers who will use the risk assessment. (b) The scope of the assessment should balance scientific completeness with the information needs of the decision makers; the costs and benefits of acquiring further information should be considered. (c) The level of effort should be commensurate with the importance of the risk assessment. (d) Agencies should consider the importance of the risk assessment in allocating resources. (e) Appropriate procedures for peer review and for public participation should be used. At the request of the OMB, the National Research Council reviewed the proposed bulletin and recommended that the current draft bulletin be withdrawn, and replaced with a bulletin that outlines general goals and principals of risk assessment and directs federal agencies to develop their own technical guidelines [28].

In contrast to the United States, CRA is not so widely used by European governments. Examples of CRA in Europe include the following:

The European Commission has funded a study of food safety in which CRA will be used to improve current risk analysis practices for food produced by different breeding approaches and production practices [29].

In 2003, the International Risk Governance Council (IRGC) recommended development of databases and methodology for CRA in its report of IRGC activities for the part-year June 2003–December 2003 [30].

The World Health Organization has adopted CRA in its assessments of selected major risk factors and global and regional burden of disease [12].

References

- [1] US Environmental Protection Agency (1981/notices). Federal radiation protection guidance for occupational exposures; proposed recommendations, request for written comments, and public hearings, *US Federal Register* **46**(15), 7837.
- [2] National Research Council, Committee on Institutional Means for Assessment of Risk for Public Health (1983). *Risk Assessment in the Federal Government: Managing the Process*, National Academy Press.
- [3] US Occupational Safety and Health Administration, Department of Labor, Occupational Exposure to Benzene, *US Federal Register* (1985). Proposed Rules, **50**, 50539.
- [4] U.S. Environmental Protection Agency (1987). *Unfinished Business: A Comparative Assessment of Environmental Problems, Volume I: Overview*, US EPA, reproduced by U.S. Department of Commerce, Technical Information Service, Springfield VA 22161.
- [5] Cancer Risk Work Group (1987). *Unfinished Business: A Comparative Assessment of Environmental Problems*, Appendix I, Springfield.
- [6] Non-Cancer Risk Work Group (1987). *Unfinished Business: A Comparative Assessment of Environmental Problems*, Appendix II, Springfield.
- [7] Ecological Risk Work Group (1987). *Unfinished Business: A Comparative Assessment of Environmental Problems*, Appendix III, United States Environmental Protection Agency, Springfield.
- [8] Welfare Risk Work Group (1987). *Unfinished Business: A Comparative Assessment of Environmental Problems*, Appendix IV, United States Environmental Protection Agency, Springfield.
- [9] Carnegie Commission on Science Regulation and Government (1993). *Risk and the Environment: Improving Regulatory Decision Making*, Carnegie Commission on Science Regulation and Government, New York.
- [10] Graham, J. D. (1995). Harvard group on risk management reform of risk regulation: achieving more protection at less cost, *Human and Ecological Risk Assessment* **1****3**, 183–206.
- [11] The Presidential/Congressional Commission on Risk Assessment and Risk Management (1997). *Risk Assessment and Risk Management in Regulatory Decision Making*, Final Report Volume 2, Presidential/Congressional Commission on Risk Assessment and Risk Management; www.riskworld.com, p. 46.
- [12] World Health Report (2002). *Quantifying Selected Major Risks to Health*, Chapter 4, World Health Organization, p. 7.
- [13] Danaei, G., Vander Hoorn, S., Lopez, A.D., Murray, C.J.L. & Ezziati, M. (2005). Causes of cancer in the world: comparative risk assessment of nine behavioral and environmental risk factors, *The Lancet* **366**, 1784–1793.
- [14] Risk Assessment Division, Office of Public Health Science, Food Safety and Inspection Service & U.S. Department of Agriculture (2002). *Comparative Risk Assessment for Intact (Non-Tenderized) and Non-Intact (Tenderized) Beef: Executive Summary*.
- [15] Mystic Watershed Collaborative (2003). *Chelsea Creek Community Based Comparative Risk Assessment*, <http://epa.gov/region1/eco/uep/boston/ccra.pdf>.
- [16] US Consumer Product Safety Commission (1986). *All-Terrain Vehicle Task Force Briefing*, US Consumer Product Safety Commission, p. 121.
- [17] Newman, R. (1981). *Overview of Chain Saw Related Injuries*, US CPSC, p. 7.
- [18] Baxter, L. Brown, V., Present, P., Rauchschalbe, R. & Young, C. (1984). *Report: Infant Drowning in Swimming Pools/Spas*, US CPSC, p. 4.
- [19] *Greenwood v Honda*, Superior Court of Arizona, County of Maricopa, CV99-022663.
- [20] Abkowitz, M., DeLorenzo, J., Duych, R., Greenberg, A. & McSweeney, T. (2001). Comparative risk assessment of hazmat and non-hazmat truck shipments, *Transportation Research Record*, TRB 2001 Annual Meeting, pp 1–27.
- [21] Raslear, T. & Conklin, J.A. (2006). *Comparative Risk Assessment of Remote Control Locomotive Operations versus Conventional Yard Switching Operations*, US DOT Federal Railroad Administration, Research Results, RR06-01.
- [22] Whiting, R., Clark, C., Hicks, J., Dennis, S., Buchanan, R., Brandt, M., Hitchins, A., Raybourne, R., Ross, M., Bowers, J., Dessai, U., Ebel, E., Edelson-Mammel, S., Gallagher, D., Hansen, E., Kaase, J., Levine, P., Long, W., Orloski, K., Schlosser, W. & Spease, C. (2003). *Quantitative Assessment of Relative Risk to Public Health from Foodborne Listeria Monocytogenes among Selected Categories of Ready to Eat Foods*, FDA/Center for Food Safety and Applied Nutrition, USDA/Food Safety and Inspection Service, Centers for Disease Control and Prevention, <http://www.foodsafety.gov/~dms/lmr2-toc.html>.
- [23] Schultz, H., Wiedemann, P., Hennings, W., Mertens, J. & Clauberg, M. (2006). *Comparative Risk Assessment*, John Wiley & Sons.
- [24] Kadvany, J. From Comparative Risk to Decision Analysis: Ranking Solutions to Multiple-Value Environmental Problems, *Risk: Health, Safety and Environment* (1995). **6**, 333–358, <http://www.piercelaw.edu/risk/vol6/fall/kadvany.htm>.
- [25] Finkel, A. Risk: Health, Safety and Environment, *Comparing Risks Thoughtfully* (1996). **7**, 325–359, <http://www.piercelaw.edu/risk/vol7/fall/finkel.htm>.
- [26] Gutenson, D. *Comparative Risk: What Makes a Successful Project?* (1997). <http://www.law.duke.edu/journals/delpf/articles/delpf8p69.htm>.
- [27] US Office of Management and Budget (2006). Proposed Risk Assessment Bulletin, *Proposed Risk Assessment Bulletin 3*, 1–26, http://www.whitehouse.gov/omb/inforeg/proposed_risk_assessment_bulletin_010906.pdf.

- [28] Committee to Review the OMB Risk Assessment Bulletin, Scientific Review of the Proposed Risk Assessment Bulletin from the Office of Management and Budget, National Research Council (2007). *Scientific Review of the Proposed Risk Assessment Bulletin from the Office of Management and Budget*, National Academies Press, ISBN-10: 0-309-10477-7.
- [29] CSIR (2004). South African Biotechnologists participate in major new European Food Safety Research Project, African Centre for Gene Technology, http://www.acgt.co.za/news/articles/082004_foodsafety.html (accessed Aug 2004).
- [30] Gago, J. (2003). *Annual Report of the International Risks Governance Council 2003*, IRGC, Geneva, pp. 1–9.

Related Articles

Ecological Risk Assessment

Scientific Uncertainty in Social Debates Around Risk

Statistics for Environmental Mutagenesis

ROSE M. RAY

Compensation for Loss of Life and Limb

Each year, approximately one in six Americans suffers economic losses caused by nonfatal injuries. The direct and work-loss costs associated with these injuries total close to 4% of the US gross national product (GNP) [1]. In addition, approximately 1 in 2500 Americans dies each year from an unintentional injury [2].

This article discusses the direct economic costs, indirect economic costs, and noneconomic costs associated with injuries and fatalities. Different methods for putting a value on noneconomic costs, often referred to as *pain and suffering*, are compared, and the types of compensation available for these different costs, including private insurance, public insurance, and the tort system, are examined. Next, a detailed discussion of how the tort system treats compensation for loss of life and limb is given. Finally, this article considers how life and limb are valued in other contexts, such as setting government regulations and assessing health care priorities. The focus is on the United States.

Injury Statistics

Approximately 50 million Americans suffer economic losses each year caused by nonfatal injuries. Of these persons, approximately 75% experience an economic loss from an injury that occurred the same year, whereas 25% experience an economic loss from an injury that occurred in a previous year. The breakdown of the source of injuries among these persons is given in Figure 1.

The work-loss associated with these injuries is significant. Overall, more than 6% of employed adults lose time from work each year owing to injuries [1].

In addition, there are 170 000 fatal injuries in the United States each year, of which 70% result from accidents. The remaining 30% are violence related [2].

Components of Loss

Accident victims suffer various types of loss associated with their injuries. Economic costs include

medical expenses and other direct costs, such as special equipment and home help. These direct costs account for more than half of the economic losses associated with nonfatal injuries. In addition, there may be losses because of the individual's inability to continue with some or all work activities, as well as lost wages of family members who curtail work to care for the injured [1].

Further, there are noneconomic losses, including inability to perform nonmarket activities (e.g., attending school, caring for children, participating in hobbies), loss of enjoyment of life, and the fact that money has less utility if the injured person cannot use it for the things he/she has used it for in the past. For example, the money that used to buy a person the enjoyment of a concert can no longer bring that pleasure if the person becomes deaf.

Economic losses are much easier to value than noneconomic losses. In particular, placing a value on loss of enjoyment of life, referred to as *pain and suffering* in the legal lexicon, has proved to be particularly challenging.

Valuing Life and Limb

Valuing Life

Historically, valuations of life were based on human capital theory. A life was valued by future production potential, generally calculated as the present discounted value of expected future earnings [3]. This

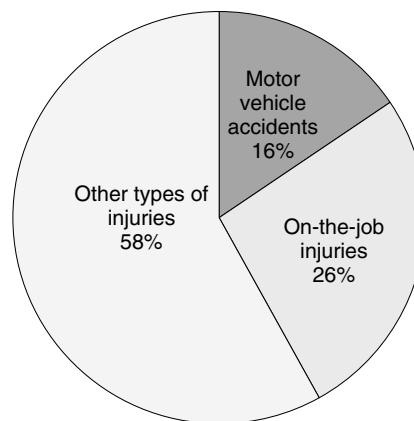


Figure 1 Nonfatal injuries resulting in economic loss in the United States

captured the economic component of loss, but did not address the noneconomic component.

The “hedonic” value of a life refers to the value of life itself, for the pleasure of living, independent of earnings potential. This type of value is much harder to quantify. Beginning in 1968, researchers have gone on to measure the implicit value of life by examining human behavior.

Economists infer value from revealed preferences – i.e., by observing how people actually behave. Some may assert that life has an infinite value, but if that were true, people would avoid dangerous activities whenever possible, eat only healthful foods, and so on. Thus, although there may not be a clear trade-off between life and monetary amounts, it is clear that people do make trade-offs between various pleasurable activities and the risks to life and limb that go along with those activities.

One way that economists measure how much people value life is by studying the wage premium for risky professions: how much extra compensation is required to convince people to take on riskier jobs? In various studies, economists have estimated the risk of death in dangerous jobs and compared the wages to those from safer jobs that are otherwise similar. The difference in wages between those jobs, divided by the extra risk of death in the more dangerous job, gives a measure for the value of a statistical life (VSL). The VSL estimates derived in this way are generally in the \$2 million to \$12 million range (US \$2000) [4–6].

Criticisms of this method are that dangerous jobs and safer jobs are never actually equivalent, with the only difference being the risk level (for example, there may be tension associated with higher risk jobs). Also, the people who are willing to take on more dangerous jobs may be those who place a lower value on their lives or who are relatively risk tolerant or risk seeking [7]. Hence, their revealed VSLs are not necessarily applicable to other people.

Other ways to measure the value of life involve evaluation of the choices that people make, which affect their risk of death or injury. Studies have examined the price-risk trade-off for automobile safety, seat belt use, bicycle helmets, cigarette smoking, smoke detectors, and property values in polluted areas. Such studies yield VSL estimates that range from approximately \$1 million to \$10 million (2000 US \$). These estimates tend to be a little lower,

although on the same order of magnitude, as those obtained from labor market studies [4, 8].

Although these measures are imperfect, they may provide a useful starting point for VSL calculations. But such measures should not be taken literally, since a body of evidence suggests that people do not have accurate perceptions of risk and so may not make fully informed choices [9]. Also, VSL estimates based on willingness to pay (WTP) naturally vary with income, since those with lower income cannot afford to pay as much for increased safety. Moreover, VSLs estimated from these studies reflect the preferences of the particular population in the study and so should not be considered a universal constant.

Valuing “Limb”

Value of life estimates obtained from revealed preferences may be coupled with assessments of the relative quality of life for different health states to get a measure of the “value” associated with different types of injuries.

In economics, utility reflects the satisfaction gained from different alternatives. Measuring health utilities involves defining a set of health states and having people provide judgments on the relative desirability of time spent in these different states. Methods of soliciting such judgments include the standard gamble, time trade-off, rating scales, and WTP [10].

The standard gamble (or “hypothetical lottery”) procedure is a classic method of measuring preferences (or, more accurately, von Neumann–Morgenstern utilities) in economics. It makes use of a hypothetical lottery to measure people’s utilities, which reflect preferences for outcomes and attitude toward risk. The lottery typically involves a choice between two alternatives: (a) a certain outcome or (b) a gamble giving some probability of obtaining a better outcome and a complementary probability of ending up with a worse outcome. The stakes and/or odds in the lottery are shifted until the person is indifferent between the hypothetical lottery and the certain outcome.

The time trade-off method was developed as a simpler alternative to the standard gamble. It trades off the quality of life against the length of time alive.

Another simple technique is a rating scale. This can take on various forms. An example is a visual analog scale consisting of a single line with the best

possible health state (perfect health) with a score of 1 and the worst possible health state (generally, death) with a score of 0. This is shown in Figure 2.

Respondents are asked to place various health states on the line at the positions the respondent thinks are appropriate. Health state utilities derived from the rating scale are consistently lower than those obtained by either standard gamble or time trade-off [10].

In Europe, the EuroQol EQ-5D is a questionnaire that is used to assess people's utility for different health states [11].

It should be noted that there are significant caveats to all of these methods. Health utilities are extremely subjective measures and may vary widely based on the measurement method (e.g., wording of questions asked and allowed responses in a survey), as well as the population being surveyed. For example, different studies have estimated the health utility for depression between 0.3 and 0.99, depending on how depression is defined and what population is being studied [12].

A catalog of health state preference scores obtained in different studies is maintained by the Tufts Cost-Effectiveness Analysis (CEA) Registry [12]. Some examples of preference scores for different health states, as listed in this registry, are given in Table 1.

To calculate the hedonic value of an injury in monetary terms, one could start by finding the loss in quality of life associated with that injury. For profound deafness, this could be calculated as 1

(state of perfect health) minus 0.6 (utility associated with profound deafness), giving 0.4. If this state is assumed to last for the rest of the person's life, and if the length of life is assumed to be unchanged from what it would have been without the injury, then the "value" of the injury could be taken as the loss in quality of life multiplied by the value of a life. For example, if a life is taken to be worth \$5 million, then the "value" of profound deafness might be estimated, very roughly, as $0.4 \times \$5 \text{ million} = \2 million . Of course, adjustments should be made for age, since a young person who is permanently injured must live with the affliction for many more years than an older person. Also, the calculation might include a discount rate to convert to present value.

Compensation Systems

Sources of Compensation

Accident victims may seek compensation for their losses from a variety of sources. The starting point is often the person's own private insurance – for example, health, automobile, accident, or disability insurance. If the person has been injured on the job, the person may seek workers' compensation. There may also be employer benefits such as paid sick leave. In addition, there are multiple public insurance programs that assist victims, including the National Vaccine Injury Compensation Program, unemployment benefits, and supplemental security income, among others. These types of compensation are generally "no fault" – they compensate the victim regardless of who was at fault in causing the injury.

If the victim believes that another party is responsible for the injury, he/she might seek liability compensation. The percentage of persons seeking liability compensation for different types of accidents is given Table 2 [1].

Although only 1 in 10 injured Americans with economic loss attempts to collect liability payments from the injurer or the injurer's insurance, this rate varies widely with the type of accident. The claiming rate for persons injured in motor vehicle accidents is more than 6 times higher than that for work accidents and almost 15 times higher than that for other nonwork accidents. A contributing factor to the lower claiming rate for work accidents is that in most cases, employees receiving workers' compensation are barred from seeking liability compensation.

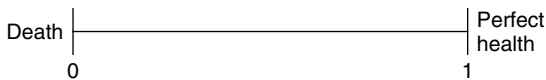


Figure 2 Visual analog rating scale

Table 1 Selected health state preference scores from the Tufts CEA Registry

Health state	Preference score (0–1)
Blindness	0.4–0.7
Trouble hearing	0.8
Profound deafness	0.6
Below knee amputation	0.6–0.8
Manual wheelchair	0.7
Electric wheelchair	0.8
Disability in bed	0.4

Table 2 Percentage of persons with economic loss arising from nonfatal injuries who make liability claims in the United States^(a)

Type of accident	Percentage of persons injured in a given type of accident who make liability claims (%)
All accidents	10
Motor vehicle	44
On the job	7
Other	3

^(a)Reproduced from [1], with permission from the RAND Institute for Civil Justice, © 1991

Recovery Rates

The average percentage of economic losses recovered by persons with nonfatal injuries is given in Table 3. Overall, persons with nonfatal injuries can expect to recover approximately 62% of their economic losses. The recovery rate is higher for direct costs (75%) than for earnings losses (34%) [1].

Efficiency of Compensation Systems

The efficiency of the different compensation systems varies widely, as shown in Table 4. In the tort system, only 46 cents of each dollar spent on the system goes toward compensating victims [13]. In contrast, in public insurance programs, a much higher fraction of each dollar goes toward compensating victims.

The following section gives a more detailed discussion of the compensation provided by some representative insurance programs.

Examples of Insurance Compensation Programs

Private Accident Insurance

Accidental death and dismemberment policies provided by the private insurance market vary widely. An example of a schedule of benefits provided by such policies is given in Table 5. These policies make only crude assessments of the relative values of different types of injuries.

Workers' Compensation

Workers' compensation is an insurance plan provided by employers as per US law. It is a no-fault system, which pays employees for job-related injuries, disabilities, or death. Workers' compensation was intended to reduce the need for litigation, as well as the need to prove fault. In the majority of states, workers' compensation is provided solely by private insurance companies, but some states have a state fund or state-owned monopolies.

Workers' compensation provides compensation for both medical expenses and earnings losses. Benefits for disability vary across states, but are typically two-thirds of the worker's wages, with set minimum and maximum amounts per week. Some states also limit the duration of benefits. Replacement rates below 100% are meant to encourage employees to return to work, as well as to reflect the fact that these benefits, unlike earnings, are nontaxable.

Death benefits also vary from state to state. Benefits to dependants are typically two-thirds of the worker's wages, with set minimum and maximum amounts per week.

Table 3 Recovery rate for different types of losses associated with injuries in the United States

Category of loss	Specific type of loss	Recovery rate (average percentage compensated from all sources combined) (%)
Economic loss	Total direct and work loss	62
Direct costs	Total direct	75
	Medical treatment	84
	Other direct costs	10
Work loss	Total work loss	34
	Short term	66
	Long term	20

Table 4 Efficiency of different methods of compensation in the United States [14]

Method of compensation	Percentage of dollars spent on system that goes toward compensating victims (remainder goes to attorney fees and other costs) (%)
Tort system	46
Workers' compensation	80
Vaccine Injury Compensation Program	85

Table 5 Example of schedule of benefits provided by private accidental death and dismemberment policies. Actual benefits vary by policy

Loss	Percentage of principal benefit (%)
Death	100
Loss of use of one limb	50
Loss of use of two limbs	75
Loss of sight of both eyes	100
Loss of hearing of both ears	50
Loss of speech	50

National Vaccine Injury Compensation Program

The National Vaccine Injury Compensation Program is an example of a public insurance program. Created by the National Vaccine Injury Act of 1986, the program was intended to “ensure an adequate supply of vaccines, stabilize vaccine costs, and establish and maintain an accessible and efficient forum for individuals found to be injured by certain vaccines.” The program is a no-fault alternative to the tort system. Claims are decided by the US Court of Federal Claims.

The program provides compensation for medical costs, custodial care, rehabilitation costs, and other direct expenses associated with vaccine-related injuries. In addition, reimbursement is given for lost earnings, and up to \$250 000 may be awarded for pain and suffering. Over the history of the program, persons judged to have vaccine-related injuries have received an average compensation of \$824 000 per

person. For deceased victims, up to \$250 000 may be given as a death benefit to the estate [15].

Insurance versus Liability Compensation

Advantages of compensation from public insurance or a person's private insurance are speed and ease of reimbursement, as well as low transaction costs (see **Pricing of Life Insurance Liabilities**). However, such programs typically do not compensate fully for economic losses (in particular, earnings losses tend to be less than fully compensated) and often do not provide compensation for noneconomic losses. Thus, if a person feels that his/her injury is the resultant of another party's fault, he/she might make a liability claim and possibly turn to the tort system in an attempt to receive more comprehensive compensation. The tort system is discussed in the next section.

The Tort System

Overview of the Tort System

A tort is a civil wrong (as opposed to a criminal wrong) that results in personal or economic harm to another person and is recognized by law as grounds for a lawsuit. The aims of the tort system are generally agreed to be compensation (making the victim “whole”), deterrence (creating disincentives for putting others at risk), and justice (compelling the injurer to pay for the losses he/she has caused).

There are three general categories of torts: intentional torts (e.g., intentional battery), negligent torts (e.g., causing a motor vehicle accident by disobeying traffic laws), and strict liability torts (e.g., making or selling a defective product). US tort law is based primarily on common law, in which rules are developed on a case-by-case basis by judges, as opposed to statutory law, which is created by legislatures [16].

Statistics about the Tort System

The US tort system is largely a state system. Approximately 95% of tort lawsuits are filed in state courts, with the remaining 5% filed in federal courts. Most tort lawsuits are settled or dropped before a verdict is reached. Study of such cases is difficult since the details of settlements are usually private [14].

Of tort cases filed in state courts, approximately 97% are terminated before a trial verdict is given. Of tort cases (both personal injury and other torts) that reached a verdict in state courts in the United States' largest 75 counties in 2001, plaintiffs won 52% of the time, with a median award of \$27 000 to plaintiff winners [17].

Of the tort cases filed in federal courts in fiscal 2002–2003, approximately 98% were terminated before a trial verdict was reached. Of the personal injury cases that reached a verdict, plaintiffs won 48% of the time, with a median award of \$181 000 to plaintiff winners [18].

For wrongful death cases in general, the median award to plaintiff winners is approximately \$900 000, according to Jury Verdict Research [19].

Although trial verdicts represent only a small minority of cases filed, they are important in setting a precedent and thereby affecting the incentive to settle for other litigants.

Overall, approximately \$205 billion in direct costs, or about 2% of the US gross domestic product (GDP), was spent on the US tort system in 2001; this works out to approximately \$720 per citizen. Of this, approximately \$95 billion went toward compensating victims, whereas the remaining \$110 billion was spent on transaction costs. Of the component that went toward compensating victims, approximately half of it was for economic loss and the remaining for pain and suffering [13].

How Compensation is Determined

In the US courts, the types of damages that may be awarded include economic damages, pain and suffering, and punitive damages.

Economic Damages. According to the Restatement (Second) of Torts, the objective of the economic component of compensation is to restore the victim to “a position substantially equivalent in a pecuniary way to that which he would have occupied had no tort been committed” [20]. To determine economic damages, the jury considers testimony from expert witnesses and other parties involved in the case and makes an assessment of the appropriate level of damages to award.

Pain and Suffering Damages. The Restatement (Second) of Torts states that noneconomic damages

are “intended to give to the injured person some pecuniary return for what he has suffered or is likely to suffer. There is no scale by which the detriment caused by suffering can be measured and hence there can be only a rough correspondence between the amount awarded as damages and the extent of the suffering” [20]. All states allow compensation for pain and suffering for injury cases, while only a few states allow such compensation for wrongful death cases [21].

In deciding pain and suffering awards, juries have wide discretion and little guidance from the courts. Since there is no general agreement on what losses should be measured, nor how they should be measured, the standard procedure is to leave it to the jury [20]. There is no institutional “memory” in the courts of damages that have been awarded in the past for different types of injuries (it is not permitted to instruct juries on patterns of awards for similar cases). Therefore, there tends to be significant variation in awards made, even for similar injuries and accidents.

Plaintiff attorneys have attempted various ways to advise juries of appropriate levels of pain and suffering damages. It is generally *not* permitted for plaintiff attorneys to ask juries to consider how much compensation they would require before the fact to accept the certainty of the injury that the victim suffered, nor can attorneys ask juries to consider the compensation they would require after the fact to accept the victim's pain and suffering. Some jurisdictions do allow the “per diem” method, whereby jurors consider the compensation appropriate for a small unit of time and then multiply by the plaintiff's life expectancy. Also, “day in the life” videos may sometimes be used [22].

After the damages have been decided by the jury, the court may adjust the award if it is found to be excessive or if it exceeds statutory caps.

It is interesting to note that in England, personal injury cases are almost always tried by a judge because it is considered inappropriate for juries to be asked to assess compensatory damages [22].

Among law and economics scholars, there is a debate about “optimal” damages for pain and suffering. From a deterrence perspective, it is argued that defendants should bear the full cost of the injuries they have caused, however this “full cost” is determined. On the other hand, from a pure compensation perspective, the argument has been made that pain and suffering damages should be equivalent to the insurance coverage a rational and

informed individual would have purchased on the open market, if such coverage existed. Even if one agrees with this “insurance” argument, it is very difficult to ascertain how much coverage individuals would demand and pay for in such a hypothetical scenario [22].

Punitive Damages. In addition to compensatory damages, punitive damages may be awarded to punish and deter egregious behavior. The Restatement (Second) of Torts states that punitive damages are designed to “punish (the defendant) for his outrageous conduct and to deter him and others like him from similar conduct in the future” [23].

Because the purpose of punitive damages is to punish defendants rather than compensate victims, there is a controversy over who should receive punitive damage payments. Typically, plaintiffs receive any punitive damages awarded, although eight states (Alaska, Georgia, Illinois, Indiana, Iowa, Missouri, Oregon, and Utah) mandate split-recovery, whereby the plaintiff and the state split punitive damage awards between themselves [23].

Studies of Jury Awards

Various studies of jury award data have been carried out to assess what factors are correlated with higher awards and whether such awards are as arbitrary as critics contend. These studies have found that as much as half of the observed variation in awards can be explained by regression models, indicating that compensation of personal injuries is not as haphazard as sometimes perceived [21, 24].

In a study of personal injury cases from Florida and Kansas City between 1973 and 1987, cases were categorized by degree of injury severity on a nine-point scale. It was found that injury severity was the best single predictor of the amount of awards, explaining approximately 40% of the award variance. However, within each severity category, there was an enormous range of award values. For the most severe category, awards ranged from approximately \$150 000 to \$18 000 000. This calls into question the horizontal equity of the system, since injuries of similar severities may result in vastly different damage awards [25].

A study of nonfatal drunk driving injury cases between 1980 and 1990 found that juries placed a

value of approximately \$2.3 million on total impairment, a value of the same order of magnitude as VSL estimates from revealed preferences studies. This research also found that more weight was given to disabilities that were obvious to observers. In line with human capital theory, juries appeared to take earnings potential into account, providing the greatest awards to people between the ages of 30 and 45. In addition, awards were found to be higher when the defendant had deep pockets. In contrast, awards were found to be lower for plaintiffs perceived to be irresponsible (for example, plaintiffs who were impaired at the time of the accident) [21].

A study of permanent injuries in birth-related and emergency room cases found that compensation tended to fall short of the full economic cost of injuries. In particular, large losses were under-compensated relative to smaller ones – a percentage point increase in loss was found to raise compensation by far less than a percentage point [26]. The next section discusses how life and limb are valued in other contexts.

Valuing Life and Limb in Other Contexts

Value of Life for Setting Government Regulations

There is general consensus that, in principle, the proper way to value a statistical life when setting government regulations that affect health and safety is by WTP [9]. (In practice, however, WTP estimates based on survey data may reflect political attitudes, cues in the framing of questions or of response options, perceptions about legitimacy and fairness, and other factors that go beyond purely economic evaluation.) For more than 20 years, US federal agencies have used labor market estimates of the value of life to do cost–benefit analysis of health, safety, and environmental regulations [27].

As shown in Table 6, the VSL used by government agencies in setting regulations varies from approximately \$1 million to \$10 million, with average values clustering in the \$3 million to \$6 million range. This matches well with VSL estimates derived from economic studies.

Value of Life for Health Care Spending

In the area of health care, it is becoming increasingly common to evaluate the cost-effectiveness of

Table 6 Value of a statistical life (VSL) used by US federal agencies [4, 19, 28]

US federal agency	VSL (\$millions)	Year
Environmental Protection Agency (EPA)	6.3 (range from 1 to 10)	2004
Food and Drug Administration (FDA)	5.5	1996
Consumer Product Safety Commission (CPSC)	5.0	2000
Department of Transportation (DOT)	3.0	2002
Federal Aviation Administration (FAA)	3.0	1996
Food Safety Inspection Service, US Department of Agriculture (USDA)	1.9	1996
Average payout of the September 11th Victim Compensation Fund	1.8 (range from 0.25 to 7)	2003

medical interventions. In particular, countries with nationalized health care take the cost-effectiveness of different health care interventions into consideration when determining how to allocate limited resources.

To compare different medical interventions that may have disparate impacts on quality and/or length of life, a common unit is required. A unit that is often used is the quality-adjusted life year (QALY), which represents one year of perfect health. A year of life in a disabled or diseased state has a value between zero and one QALY, depending on the quality of life in that state. Each medical intervention can be assigned a cost per QALY, and the cost-effectiveness of different interventions can thereby be compared. Other units used for this type of comparison include the disability-adjusted life year (DALY) and health-adjusted life expectancy (HALE).

In the United Kingdom, the National Institute for Clinical Excellence (NICE) is an independent organization that provides national guidance on the adoption of new medical interventions. NICE evaluates the cost-effectiveness of pharmaceuticals and other medical treatments, using an implicit threshold of approximately £30 000 per QALY when deciding whether or not to recommend an intervention. (In general, NICE tends to recommend interventions that cost less than £30 000 per QALY, whereas it does not typically recommend interventions that cost more than this [29, 30].)

In the United States, there is no accepted cost per QALY threshold. Indeed, Medicare explicitly rejects cost-effectiveness as a criterion for approving new medical technologies. However, interventions are generally considered to be good value if they cost less than \$50 000 to \$100 000 per QALY. The \$50 000 per QALY benchmark arises from Medicare's decision in the 1970s to cover dialysis for chronic renal failure

patients, a treatment with a cost-effectiveness ratio of approximately \$50 000 per QALY [31]. Studies using WTP and revealed preferences approaches suggest that the public values a QALY even higher than this – over \$200 000 [32].

To translate, at least very roughly, between QALYs and VSL, one could make the assumption that the average value of a person's remaining years of life is about 40 QALYs (approximately half the US life expectancy). In this case, a cost per QALY between \$50 000 and \$200 000 translates to a VSL of roughly \$1 million to \$8 million, depending on the discount rate used to calculate present value. This is in line with the range of VSLs used by government agencies in setting regulations.

Conclusion

Overall, there are trade-offs to be made in any type of compensation system. No-fault systems, such as private and public insurance, tend to be efficient, with low transaction costs. However, such systems do not accomplish other objectives such as deterrence and justice, which the tort system attempts to achieve.

Criticisms of the US tort system revolve around its high transaction costs, as well as the unpredictability and inconsistency of awards. Various proposals have been put forth to reform the tort system. Some argue that pain and suffering awards should be capped, but such caps would only come into play for the most severe injuries, while having little effect on overvaluation of minor injuries. Others propose that damages should be calculated *via* schedules, matrices, scenarios, or multipliers. However, such systems do not address the fundamental question of how damages should be valued in the first place. On that question, there is significant ongoing debate, both

on the value of a life overall and on the value of the loss of enjoyment of life associated with different types of injuries.

References

- [1] Hensler, D.R. Marquis, M.S. Abrahamse, A.F. Berry, S.H. Ebener, P.A. Lewis, E.G. Lind, E.A. MacCoun, R.J. Manning, W.G. Rogowski, J.A. & Vaiana, M.E. (1991). *Compensation for Accidental Injuries in the United States*, The Institute for Civil Justice, RAND, R-3999-HHS/ICJ.
- [2] Centers for Disease Control (CDC) (2004). *National Center for Injury Prevention and Control, Web-based Injury Statistics Query and Reporting System (WISQARS)*, <http://www.cdc.gov/ncipc/wisqars/>.
- [3] Landefeld, J.S. & Seskin, E.P. (1982). The economic value of life: linking theory to practice, *American Journal of Public Health* **72**, 555–566.
- [4] Viscusi, W.K. & Aldy, J.E. (2003). The value of a statistical life: a critical review of market estimates throughout the world, *The Journal of Risk and Uncertainty* **27**, 5–76.
- [5] Viscusi, W.K. (1993). The value of risks to life and health, *Journal of Economic Literature* **31**, 1912–1946.
- [6] Mrozek, J.R. & Taylor, L.O. (2002). What determines the value of a life? A meta-analysis, *Journal of Policy Analysis and Management* **21**, 253–270.
- [7] Friedman, D. (1982). What is fair compensation for death or injury? *International Review of Law and Economics* **2**, 81–93.
- [8] Viscusi, W.K. (2005). *The Value of Life*, John M. Olin Center for Law, Economics, and Business, Harvard, Discussion Paper No. 517.
- [9] Becker, W.E. & Stout, R.A. (1992). The utility of death and wrongful death compensation, *Journal of Forensic Economics* **5**, 197–208.
- [10] Petrou, S. (2001). What are health utilities? *Hayward Medical Communications* **1**, 1–6.
- [11] <http://www.euroqol.org/> (accessed 2007).
- [12] <http://www.tufts-nemc.org/cearegistry/> (accessed 2007).
- [13] Sutter, R. (2002). *U.S. Tort Costs: 2002 Update*, Tillin-ghast-Towers Perrin.
- [14] Congressional Budget Office (2003). *The Economics of U.S. Tort Liability: A Primer*.
- [15] <http://www.hrsa.gov/vaccinecompensation/> (accessed 2007).
- [16] Cornell Law School, Legal Information Institute, Wex (2006). <http://www.law.cornell.edu/wex/index.php/Tort> (accessed 2007).
- [17] Cohen, T.H. (2004). Tort trials and verdicts in large counties, 2001, *Bureau of Justice Statistics Bulletin*, <http://www.ojp.usdoj.gov/bjs/pub/pdf/ttvlc01.pdf>.
- [18] Cohen, T.H. (2005). Federal tort trials and verdicts 2002–03, *Bureau of Justice Statistics Bulletin*, <http://www.ojp.usdoj.gov/bjs/pub/pdf/fttv03.pdf>.
- [19] Torpy, B. (2004). Life – hard to know what price is right, *The Atlanta Journal-Constitution*, http://ase.tufts.edu/gdae/about_us/Frank/ackerman_ajc_3-04.html.
- [20] Viscusi, W.K. (1988). Pain and suffering in product liability cases: systematic compensation or capricious awards? *International Review of Law and Economics* **8**, 203–220.
- [21] Smith, S.V. (2000). Jury verdicts and the dollar value of human life, *Journal of Forensic Economics* **13**, 169–188.
- [22] Avraham, R. (2006). Putting a price on pain-and-suffering damages: a critique of the current approaches and a preliminary proposal for change, *Northwestern University Law Review* **100**, 87–119.
- [23] Sud, N. (2005). Punitive damages: achieving fairness and consistency after *State Farm v. Campbell*: despite its best efforts, the U.S. supreme court's opinion leaves many blank spaces and holes with which appellate courts have had to cope, *Defense Counsel Journal* **72**, 67.
- [24] Sloan, F.A. & Hsieh, C.R. (1990). Variability in medical malpractice payments: is the compensation fair? *Law & Society Review* **24**, 997–1039.
- [25] Bovbjerg, R.R., Sloan, F.A. & Blumstein, J.F. (1989). Valuing life and limb in tort: scheduling Pain and suffering, *Northwestern University Law Review* **83**, 908–976.
- [26] Sloan, F.A. & Van Wert, S.S. (1991). Cost and compensation of injuries in medical malpractice, *Law and Contemporary Problems* **131**, 131–168.
- [27] Viscusi, W.K. (2003). *The Value of Life: Estimates with Risks by Occupation and Industry*, John M. Olin Center for Law, Economics, and Business, Harvard, Discussion Paper No. 422.
- [28] Brannon, I. (2004–2005). What is a life worth? *Regulation* **27**(4), 60–63.
- [29] Devlin, N. & Parkin, D. (2004). Does NICE have a cost effectiveness threshold and what other factors influence its decisions? A binary choice analysis, *Health Economics* **13**, 437–452.
- [30] Raftery, J. (2001). NICE: faster access to modern treatments? Analysis of guidance on health technologies, *British Medical Journal* **323**, 1300–1303.
- [31] Neumann, P.J. (2005). *Using Cost-Effectiveness Analysis to Improve Health Care: Opportunities and Barriers*, Oxford University Press, Oxford, pp. 157–158.
- [32] Hirth, R.A., Chernew, M.E. Miller, E. Fendrick, A.M. & Weissert, W.G. (2000). Willingness to pay for a quality-adjusted life year: in search of a standard, *Medical Decision Making* **20**, 332–342.

Related Articles

Role of Risk Communication in a Comprehensive Risk Management Approach

Stakeholder Participation in Risk Management Decision Making

HELENE L. GROSSMAN

Competing Risks

General Description of Competing Risk

Competing risks are generally evaluated by comparing the predicted time to failure or time to event for the risks being considered. Failure could represent breakage of a specified part of an automobile, death from a specified cause such as lung cancer, or failure of a specified component of a mechanical or electrical system, or the event could be transition from unemployment to employment or time to place an order on a stock exchange. We consider the situation in which there could be several potential causes of death, but only one actual cause per person, or mechanical or electrical systems without repair processes in which the failure of one of several key components could cause the system to fail. In this case, it is of interest to estimate what would happen to the remaining risk if one or more causes of failure could be eliminated. More areas of interest include study of the relationships between type of failure and covariates, study of systems in which multiple failures can be expressed, and study of the relationships between different failure types. It should be noted that the common situation of right censoring, “i.e., for some members of the sample group it is only known that the failure or event has not occurred prior to the observation time, is itself a form of competing risk.”

Mathematical Formulation

Here we describe the competing-risk problem in the standard formulation of random variables. This allows a precise description of the problem. Let $R_1, R_2, \dots, R_n$ represent the risk of failure of n different components, or n distinct causes of death. Let T_i be a random variable representing the time to failure for risk R_i . Further, let F_i be the cumulative probability distribution function of the time to failure, i.e., $F_{T_i}(t) = \text{probability}\{T_i \leq t\}$. In this formulation, the n different risks compete with one another in the sense that the failure is attributed to the risk for which the failure occurs first. In other words, the time to failure, T , of the entire process is $\min\{T_1, T_2, \dots, T_n\}$.

In the case of electrical processes, or other systems in which it is possible to build a system of independent parallel and/or serial circuits, the individual risks

are likely to be statistically independent; it may even be possible to gather data for each of the risks and estimate each of the F_i . In this case, the probability distribution of time for failure for the entire system is easily derived from probability calculus as

$$F_T(t) = \text{Probability}\{T \leq t\} = \prod_{j=1}^n F_{T_j}(t) \quad (1)$$

For the case of independent identically distributed risks, $F_T(t) = \{F_{T_1}(t)\}^n$.

If the times to failure for the individual risks are not mutually independent, then to determine the probability distribution of time to failure for the entire system, it is necessary to know the joint probability distribution of the n risks, i.e.,

$$F_{T_1, T_2, \dots, T_n}(t_1, t_2, \dots, t_n) = \text{Probability}\{T_1 \leq t_1, T_2 \leq t_2, \dots, T_n \leq t_n\} \quad (2)$$

In many practical problems, all available data consist of observations for which all of the competing risks apply. And furthermore, only the pair (T, J) is observable where T is the failure time of the system, i.e., $\min\{T_1, T_2, \dots, T_n\}$ and J is the type of failure. Without assumptions about the form of the joint probability distribution of times to failure, it may not be possible to estimate even the marginal probability distribution function, much less the complete joint probability distribution. One method for dealing with the dependence problem is to assume a known form for the dependence relationship between the failure types. The dependence relationship can be specified by means of the copula function, C (see **Copulas and Other Measures of Dependency**); for example, the joint probability distribution can be expressed as

$$F_{T_1, T_2, \dots, T_n}(t_1, t_2, \dots, t_n) = C[F_{T_1}(t_1), F_{T_2}(t_2), \dots, F_{T_n}(t_n)] \quad (3)$$

where $F_{T_1}(t_1), F_{T_2}(t_2), \dots, F_{T_n}(t_n)$ are the marginal probability distributions.

Zheng and Klein [1] developed an estimator for the marginal distribution based upon using a known (or assumed) form for the copula. Their method reduces to the Kaplan–Meier estimator if the copula is the independence copula, $C(y_1, y_2, \dots, y_n) = y_1^* y_2^* \dots y_n^*$. The use of a copula to specify the range of possible dependence can be used to estimate upper

and lower bounds for the marginal probability distributions, F_{T_i} . Zheng and Klein [1] showed that if a good estimate of dependence between T_i as measured by Kendall's τ is available, then the estimate of the marginal probability distribution functions is not sensitive to the exact form of the copula as long as it is consistent with the "known" value of τ . Of course, τ must be known *a priori*; it cannot be estimated from the data when only the pairs (T, J) are observed.

It is frequently useful to describe the failure time process through the hazard function (see **Hazard and Hazard Ratio**), h_T , rather than through the probability distribution, F_T . The hazard function describes the current risk for the population that has survived up to the current time and often has a simpler mathematical form. The hazard function at time t is the conditional probability of failure, conditional upon survival up to time t .

Specifically,

$$h_T(t) = \lim_{\Delta \rightarrow 0} \frac{1}{\Delta} [F_T(t + \Delta) - F_T(t)] / F_T(t) \quad (4)$$

In the case of competing risks and continuous-time measurement, it is typically assumed that at any given instant, t , only one failure can occur. In this case, the hazard function for the combined risk is the sum of the hazard functions for each of the failure types, that is,

$$h_T(t) = \sum_{j=1}^n h_{T_j}(t) \quad (5)$$

where $h_{T_j}(t)$ is the hazard function for failure type j .

Proportional Risk Models

The proportional risk model introduced by Cox [2] has been used to generalize the competing-risk framework beyond independent identically distributed observations.

In this formulation, the observed quantities are the triples (T, J, X) , where T is the time to failure, J is the failure type, and X is a vector of covariates, i.e., a set of variables that can be measured for each individual and are thought to be related to the risk of failure.

In the proportional risk model the hazard function for failure type j is expressed as

$$h_{T_j}(t) = \lambda_0(t) \exp(\gamma_j + X\beta_j) \quad (6)$$

where γ_j and β_j are the parameters to be estimated from the data.

The method of partial likelihood, which conditions upon the observed failure times, can then be used to estimate the parameters γ_j and β_j , while $\lambda_0(t)$ is an unspecified function of time.

Notice that, in this formulation, the case-specific hazard functions must be proportional to one another, and that conditional upon the value of the covariate, X , the time to failure, T , and the failure type, J , are statistically independent.

Frailty Models

Another generalization beyond independent identically distributed risk is the frailty model, frequently used in economic and financial applications as well as in epidemiological applications.

Variation between individuals in the population is modeled through an unobservable frailty variable, W , which represents the effect of unobserved or unknown covariate or other natural variation in the population. Conditional upon the value of the frailty, the time to failure (or event) is assumed to follow a proportional hazards framework. The hazard function for event type j , in the presence of other events, for the i th observation is

$$h_{T_{ji}}(t) = W_{ji} \lambda_0(t) \exp(X_i \beta_j) \quad (7)$$

where, T_{ji} is the time to event type j for observation i , W_{ji} is the frailty for event type j for observation i , $\lambda_0(t)$ is the baseline hazard function, X_i is the observed covariate for observation i , and β_j is the parameter to be estimated from the data. Typically the frailty variable, W , is considered to be an unobserved random variable in the population under study. In order for this model to be identifiable, it is necessary to use a parametric framework for the baseline hazard and for the frailty. Typical models include Weibull, Gompertz, exponential, and piecewise-linear. Dependence between the event times for the competing risk can be handled by assuming a dependence structure for the multivariate frailty variable $(W_{1i}, W_{2i}, \dots, W_{ni})$.

Statistical Methods for Estimating Competing Risk

Independent Competing Risks and Censoring

Statistical estimation methods for time to event data include completely nonparametric methods such as lifetable analysis and Kaplan–Meier estimates, semi-parametric methods such as the Cox proportional hazard models, and completely parametric methods. All of these methodologies are suitable for the case of independent competing risks. The statistical estimation methods for each of these methods include provision for handling censored data. An observation is right censored if it is only known that the time to event is at least as long as the observed time. This situation occurs naturally in most situations. Examples include the following: (a) patients are followed until a date when the data collection part of the study ends; survival time is censored for those patients still alive at the close of the study; (b) patients are lost to follow-up; the observed time is the time of last contact. As another example, a product is stress tested until failure or for 48 h; products that do not fail within the 48 h period are right censored.

All of the statistical methods for handling censored data rely upon the assumption of random censoring, that is, stochastic independence between time to event and censor time. This condition is automatically satisfied when the only censoring is end of the study. In this case, the time to censoring is fixed and consequently independent of the time to failure.

It is clear that it would be impossible to learn anything about the distribution of time to failure if the censoring mechanism systematically censored individuals from the study if they appeared to be at high (or low) risk of failure. In the case of loss to follow-up, the independence condition is not necessarily met; some additional testing or tracing of individuals lost to follow-up is needed to verify the independence assumption.

Mathematically, censoring is equivalent to a competing risk. The observed data consist of a triple (T, J, X) , where T is the observed time, J is an indicator of whether censored or not, and X is a vector of covariates. If no covariates are observed, X can be considered as a constant.

Generalizing this to the competing-risk situation, we see that by sequentially treating each failure type of the relevant failure mechanism and all other failure

types as censoring mechanisms, all of the standard statistical methods for independent censoring can be used with competing risks as long as the assumption of independence between hazards is satisfied.

Lunn and McNeil [3] noticed that with competing risks, each failure time represents an uncensored failure time for the type of failure that was observed and a censored failure time for all other types of failure. They developed an analysis method that treats failure type as a covariate in the Cox model and is sometimes more powerful than the more common method of treating each risk separately and classifying other types of failures as censored observations.

Dependent Competing Risks

The situation is substantially complicated if the competing risks (or censoring mechanisms) are not independent. As can be seen in the previous discussion, the distribution of time to failure for dependent competing risk is not identifiable in the absence of some assumption about the nature of the dependence.

Wong [4] introduced the concept of susceptibility to model-dependent competing risks in a lifetable analysis. Let q_{ji} be the probability of failure from cause j in interval i in the presence of all competing risks of failure, and let q_{jik} be the probability of failure from cause j in interval i for units that (formerly) failed of cause k , if cause k was eliminated as a type of failure. In the Wong model $q^{ijk} = S_{jki}^* q_{ji}$, where S_{jki} is the relative susceptibility of failure from cause j for units that failed from cause k . If S_{jki} is greater than 1, then if cause k could be eliminated, units that (formerly) would have failed from cause k would be more likely to fail from cause j than the rest of the population; if $S_{jki} = 1$, then causes j and k are independent; if S_{jki} is less than 1, then if cause k could be eliminated, units that (formerly) would have failed from cause k would be less likely to fail from cause j than the rest of the population. The susceptibility is not directly estimable from lifetable data that only records actual cause of failure. Wong suggests using auxiliary information, such as contributing cause of death, to produce external estimates of the susceptibilities. Wong demonstrates a lifetable calculus for estimating the probability of failure from cause j , when cause k has been eliminated from the population.

A two-stage nonparametric method for estimating the cumulative distribution of cause-specific time to failure is described in Kabfleisch and Prentice [5] and illustrated by Satagopan *et al.* [6]. The Kaplan–Meier method is used to estimate the overall failure rate for all types of failure combined. In the second stage, the failure rate is partitioned into failure classes by computing the cause-specific failure rate in each interval given survival up to that interval. Let t_{ji} be the i th failure from cause j . As in the Kaplan–Meier method, the probability of cause-specific failure at time t_{ji} is the product of the all-cause probability of survival up to time t_{ji} and the cause-specific probability of failure in the interval t_{ji-1} to t_{ji} . The causes of failure need not be independent in this method.

The bivariate frailty model is one form of dependence of competing risks that has found application in the analysis of data from twin studies and other studies of genetic markers. Not only is it necessary to assume that the dependence can be expressed through the multiplication of the unobserved frailty variable upon the underlying hazard function, and that lifespan (or time to event) is conditionally independent given the frailty, but also a parametric form for the frailty variable must be assumed to allow estimation of the distribution of time to event.

The hazard function for the j th risk and the i th observation is $h_{T_{ji}}(t) = W_{ji}\lambda_0(t)\exp(X_i\beta_j)$ in a frailty model. For some combinations of distributions for the baseline hazard $\lambda_0(t)$ and the frailty variable W_{ji} , the frailty term can be integrated from the likelihood function and maximum-likelihood estimates can be calculated. This is the case when $\lambda_0(t) = A\exp(Bt)$ (Gompertz) and the W_{ji} are gamma functions [7]. In other cases, e.g., when the W_{ji} follow a lognormal distribution, explicit representation of the likelihood function is not possible. Estimation methods include penalized partial likelihood [8], a modification of the EM algorithm [9, 10], and Bayesian analysis with Gibbs sampling [7].

The multitude of ways in which competing risk can depend upon one another means that there is no simple general way to model a dependent competing-risk process. The frailty models are a useful way of modeling dependent competing risks by choosing a well-behaved probability distribution for the underlying hazard functions and for the distribution of individual susceptibility or frailty.

Applications of Competing Risk

Engineering

Reliability. Consider a system that contains n components, and failure of any one of which will cause the system to fail (sequential system). Assume that components will fail independently of one another. What is the reliability (probability that the system has not failed) at time τ ?

Let T_j represent the random time to failure of component j , and F_{T_j} be the probability distribution function of T_j . Then $1 - F_{T_j}(\tau)$ is the probability that component j does not fail prior to time τ . The reliability of the system at time τ is the probability that no components have failed by time τ ; and this is given by

$$1 - F_T(\tau) = \text{Probability}\{T > \tau\} = \prod_{j=1}^n [1 - F_{T_j}(\tau)] \quad (8)$$

A second system consists of n components. If any one component continues to work the system will work (parallel system). How many parallel components are needed for a specified reliability at time τ ? Assuming identical components, let F_{T_0} be the probability distribution function of the time to failure for each component and $R(\tau)$ be the required reliability to time τ . Then the number of components required for $R(\tau)$ is the solution to the following equation:

$$R(\tau) = \text{Probability}\{T > \tau\} = 1 - F_{T_0}(\tau)^n \quad (9)$$

$$\text{Or } n = \frac{\log[1 - R(\tau)]}{\log[F_{T_0}(\tau)]} \quad (10)$$

Weibull Models. In some situations multiple causes of failure, i.e., competing risks, may exist, but the cause of failure is not observable or has not been investigated for the failed units and only the time to failure, or censoring, is known. The Weibull distribution with probability distribution function $F(t) = 1 - \exp[-(t/\alpha)^\beta]$ is often used to model failure time in this context.

A plot of the log of the empirical cumulative distribution function of failures, $F(t)$, against time is used to judge whether the observed failure process has only one failure type or multiple failure

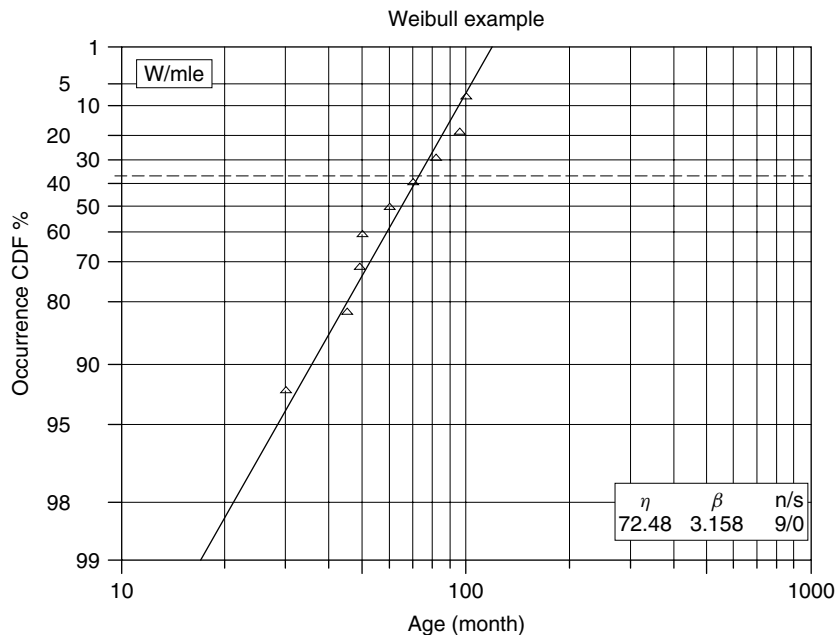


Figure 1 Plot for Weibull failure time with single failure type “W/mle: with maximum likelihood estimates”; “n/s: n is the total number of units”; “s is the number of censored units”

types. Figure 1 shows a plot that indicates that a single failure type Weibull distribution fits the data. Figure 2 shows the corresponding plot when there are two failure types. The sharp change in the slope of plot of $\log F(t)$ versus time is the indication of two failure types in this example. The WeibullSmith program for Weibull analysis contains algorithms to fit competing-risk models with an unspecified number of causes of failure, each of which follows a Weibull distribution. Note that, for the failure types to be distinguishable, they must have substantial differences in the parameters of the Weibull distributions. Two or more failure types that each have the same distribution of time to failure could not be distinguished without the observation of the failure type.

Medicine

Use of the mathematics of competing risk in medicine go back to Daniel Bernoulli (1766) and his analysis of mortality due to smallpox and normal mortality [11]. Here, a two state model was used. Individuals could be in a normal state with normal force of mortality or in a smallpox state in which individuals would

either die very quickly or recover and no longer be susceptible to smallpox.

The Cox proportional risk model and partial likelihood method as well as the Kaplan–Meier method [12] are now widely used in epidemiology and medicine. A few recent articles that give the flavor of the wide use of competing-risk ideas in medicine are [13–15].

The Kaplan–Meier method is considered suboptimal by some for use in short-term situations such as waiting lists for transplant or death during hospitalization. Some researchers [16] recommend logistic regression or other methods unrelated to survival time, while others [4, 17] recommend nonparametric lifetable type methods or Poisson regression methods.

Economics and Finance

Competing-risk models in general and the Cox proportional hazards model in particular are now widely used in economic and financial applications. The publication of Han and Hausman [18] is cited in at least 50 publications. An [19] credits Suiyoshi and McCall for introducing competing-risk methods to economic applications.

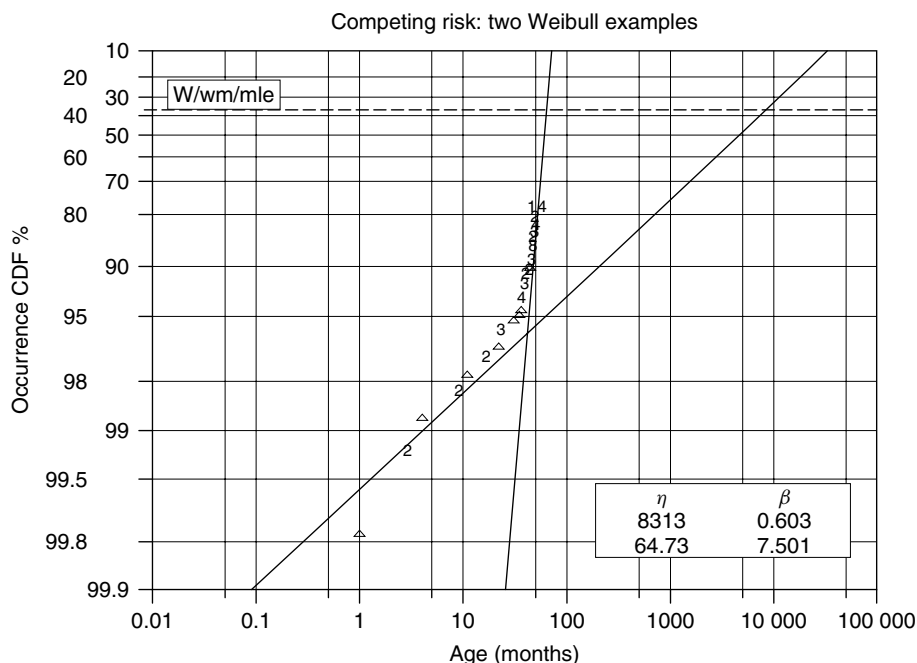


Figure 2 Plot for Weibull failure times with two failure modes

Applications of competing risk include analysis of length of unemployment, time to graduation or drop out from college, length of time to execution or cancellations of orders in financial markets, time to default or prepayment of mortgages and acquisitions of banking institutions – in short, any situation in which there are multiple reasons for transitioning from one state to another.

References

- [1] Zheng, M. & Klein, J. (1995). Estimates of marginal survival for dependent competing risks based upon an assumed copula, *Biometrika* **82**, 127–138.
- [2] Cox, D. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society* **26**, 103–110.
- [3] Lunn, M. & McNeil, D. (1995). Applying Cox regression to competing risk, *Biometrics* **51**, 524–532.
- [4] Wong, O. (1977). A competing-risk model based on the life table procedure. Epidemiological studies, *International Journal of Epidemiology* **6**, 153–159.
- [5] Kalbfleisch, J. & Prentice, R. (1980). *The Statistical Analysis of Failure Time Data*, John Wiley & Sons.
- [6] Satagopan, J., Ben-Porat, L., Berwick, M., Robson, M., Kutler, D. & Auerbach, A. (2004). A note on competing risks in survival data analysis, *British Journal of Cancer* **91**, 1229–1235.
- [7] Wienke, A., Arbeev, K., Locatelli, I. & Yashin, A. (2003). *A Simulation Study of Different Correlated Frailty Models and Estimation Strategies*, Max Planck Institute for Demographic Research, Germany.
- [8] Ripitti, S. & Palgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood, *Biometrics* **56**, 1016–1022.
- [9] Arveeb, K. & Yashin, A. (2003). *Bivariate Lognormal Frailty Models: Estimation Methods, Simulation Studies and Application to Danish Twins Data*, MPIDR, working paper.
- [10] Ripitti, S. & Palgren, J. (2002). Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm, *Lifetime Data Analysis* **8**, 349–360.
- [11] Seal, H. (1977). Studies in the history of probability and statistics. XXXV, multiple decrements or competing risks, *Biometrika* **64**, 429–439.
- [12] Kaplan, E. & Meier, P. (1956). On the identifiability crisis in competing risk analysis, *Journal of the American Statistical Association* **53**, 457–481.
- [13] Albertson, P., Hanley, J., Gleason, D. & Barry, M. (1998). Competing risk analysis of men aged 55 to 74 years at diagnosis managed conservatively for clinically localized prostate cancer, *Journal of the American Medical Association* **280**, 975–980.

- [14] Yan, Y. (2000). Competing risk adjustment reduces overestimation of opportunistic infection rates in AIDS, *Journal of Clinical Epidemiology* **53**, 817–822.
- [15] Hays, J.C., Pieper, C.F. & Purser, J.L. (2003). Competing risk of household expansion or institutionalization in late life, *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* **58**, S11–S20.
- [16] Schoenfeld, D. (2006). Survival methods, including those using competing risk analysis, are not appropriate for intensive care unit outcome studies, *Critical Care* **10**, 103.
- [17] Putter, H. (2004). *Survival Analysis for Complicated Data*, Leiden University Medical Center, <http://www.lumc.nl/3020/research/MS/Survival.html#Healthcare>.
- [18] Han, A. & Hausman, J. (1990). Flexible parametric estimation of duration and competing risk models, *Journal of Applied Econometrics* **5**, 325–353.
- [19] An, M. (2004). *Likelihood-Based Estimation of a Proportional-Hazard, Competing-Risk Model with Grouped Duration Data*, Urban/Regional 0407013, Economics WPA.

Further Reading

- Holt, J. (1978). Competing risk analyses with special reference to matched pair experiments, *Biometrika* **65**, 159–165.
- Kauermann, G. & Khomski, P. (2006). *Full Time or Part Time Reemployment: A Competing Risk Model with Frailties and Smooth Effects using a Penalty based Approach*, Department of Economics and Business Administration, University of Bielefeld.

ROSE M. RAY

Competing Risks in Reliability

Competing risks models (*see* **Repair, Inspection, and Replacement Models; Absolute Risk Reduction**) are a class of models in reliability that can be used to analyze operational field data. They assume that different failure risks and maintenance events are in competition with one another to be the cause of the system being taken out of service and that these risks censor each other, that is, only the first occurring cause is observed. Particularly significant for many applications is the fact that maintenance masks the failure event that would have happened in the future. Competing risk models implicitly assume the renewal property (*see* **Mathematics of Risk and Reliability: A Select History**) to hold. In many situations, the cost of critical failures in a system is so large that it is necessary, when components are repaired or maintained, to have a policy that ensures that they are restored to a state that is as good as new. Thus, the event times can be modeled as a renewal process with interevent times $Y_1, Y_2, \dots$, which are modeled as independent and identically distributed random variables. Raw field data for reliability is frequently found in maintenance and repair logs. Generic reliability databases make implicit use of competing risk models to convert this raw field data into information about system failure rates. Examples of these include the Centre for Chemical Process Safety (CCPS) [1] and the European Industry Reliability Database (EIREDA) [2]. A discussion of how such databases are built and their use in competing risks problems can be found in [3]. It is because of the censoring (*see* **Detection Limits; Recurrent Event Data; Lifetime Models and Risk Assessment; Hazard and Hazard Ratio**) feature of competing risk models that problems arise when trying to identify marginal distributions for the time to failure. From a purely statistical point of view it is not possible, without making nontestable assumptions, to identify the underlying marginal distributions for times to failure (mode) or time to maintenance. This non-identifiability problem can only be surmounted by making assumptions about the underlying dependencies between the competing risks variables, requiring a deeper understanding of the maintenance policies in place, which leads to more qualitative forms of model

validation. For a more general review of competing risks, which includes more general applications to survival analysis, see Crowder's recent book [4].

Main Theoretical Ideas

For simplicity, let us consider one failure cause and one preventive maintenance action. Let X denote the time to the next removal from service due to failure and let Z denote the time to the next preventative maintenance. For simplicity, we shall assume that X and Z cannot occur at the same time. In practice we can only observe $\min(X, Z)$, the smallest of X and Z , and we can also observe which variable it is $I_{\{X < Z\}}$ where

$$I_{\{X < Z\}} = \begin{cases} 1 & \text{if } X < Z \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

It would be very useful to know the distribution function of X to ascertain how changes in the maintenance policy will impact on the reliability of the system. However, we can only observe the subdistribution function

$$F_X^*(t) = P\{X \leq t, X < Z\} \quad (2)$$

This function increases to $P\{X < Z\}$ as t tends to infinity. An equivalent function that is often of interest is the subsurvival function

$$S_X^*(t) = P\{X > t, X < Z\} = P\{X < Z\} - F_X^*(t) \quad (3)$$

The normalized subsurvival function $S_X^*(t)/S_X^*(0)$ is equal to 1 at $t = 0$. Similar functions are also defined for Z as well. Another important quantity is

$$\Phi(t) = P\{Z < X | Z > t, X > t\} \quad (4)$$

the probability that, at time t since the component went into service, the next event will be Z . The functions $F_X^*, F_Z^*, S_X^*, S_Z^*$, and Φ can be directly estimated from observable data. These functions can be useful when selecting models.

Nonidentifiability Problem

Historically, assumptions about independence of X and Z were made in competing risks to be able

to identify marginal distributions from the data. However, Tsiatis [5] showed that this assumption was not testable using the data by showing that many joint distributions (including an independent model) could share the same subdistribution functions. This feature of competing risk models is frequently referred to as the nonidentifiability problem. Accepting the problem of nonidentifiability, Peterson [6] was able to show that the marginal distribution functions of X and Z are bounded pointwise in the following manner

$$F_X^*(t) \leq F_X(t) \leq P\{\min(X, Z) \leq t\} \quad (5)$$

Crowder [7] was able to demonstrate functional bounds which showed that the marginal distribution function $F_X(t)$ minus Peterson lower bound is a nondecreasing function in t . Bedford and Meilijson [8] were able to improve these results while demonstrating a simple geometric construction. The Peterson bounds are illustrated in Figure 1. This shows the upper Peterson bound, which is simply the distribution function for $Y = \min(X, Z)$, the lower Peterson bound (which is simply the subdistribution function for X), and two possible distribution functions for X (that is, marginal distribution functions for X to a joint distribution for X and Z that has the given subdistribution functions). It also shows an exponential distribution function that does not satisfy

the Peterson bounds, and therefore can be excluded as a possible marginal for X . In this particular case, no exponential distribution will fit into the Peterson bounds.

Although, in general, we cannot identify marginal distributions from competing risk data, it may be reasonable to make assumptions that restrict attention to a subclass within which we can identify marginal distributions.

Independent Model

If X and Z are independent, nonatomic, and share essential suprema, then it has been shown [9–11] that their marginal distributions can be uniquely identified. If we are given a pair of subdistribution functions, then we can always assume independence and calculate the unique marginal distributions. However, it is possible that one of the random variables will have an atom at infinity [12]. This means that we can always identify marginal distributions if we assume independence, but that we need to justify this assumption outside of the competing risks data. It was shown [13] that any other model involving dependence between X and Z would give a higher estimate of the failure rate than the independent model. Indeed, considering the class of all models in which the failure rate is constant, the interval containing these

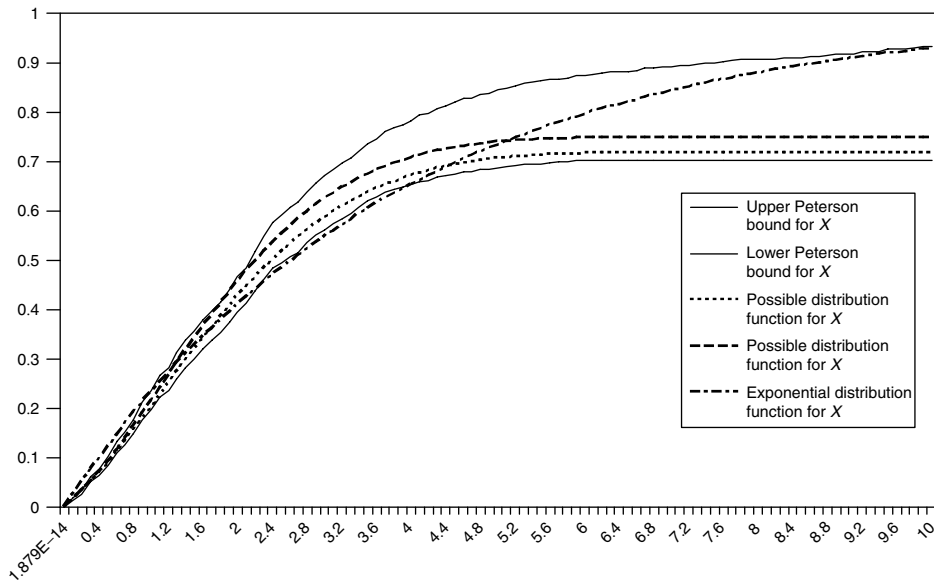


Figure 1 Illustration of the Peterson bounds

failure rates has a minimum value equal to the failure rate associated with the independent model.

Other Competing Risk Models

As discussed above, without making some assumptions, which cannot be validated statistically, it is not possible to identify underlying marginal distributions. However, a number of models have been proposed, which try to circumvent this problem usually by hypothesizing an engineering scenario within which the data has been generated. This has given rise to a number of different models.

Dependent Copula Model

The first model is a simple one that allows us to consider the effect of dependence on the estimate of the marginal. Copulas (*see Credit Risk Models; Default Correlation; Extreme Value Theory in Finance; Copulas and Other Measures of Dependency*) can be used to model the dependency between X and Z , and it was shown in [14] that a generalization of the Kaplan–Meier (K–M) (*see Detection Limits; Individual Risk Models*) estimator could be defined that gives a consistent estimator on the basis of an

assumption about the underlying copula of (X, Z) . By selecting a family of copulas, parameterized by the (rank) correlation of X and Z , we can perform a sensitivity analysis to see how the dependence parameter affects the results. For example, [15] used a family of copulas to show how sensitive optimum maintenance costs are to the choice of an underlying model for interpreting the data. Figure 2 shows upper and lower Peterson bounds for the marginal distribution of X , the actual analytical marginal used, and the K–M estimate for the marginal of X . The latter estimator, of course, makes the assumption that X and Z are independent when, in fact, in this simulation, they had a positive rank correlation. It has been observed that the assumption of independence gives a more optimistic view of the marginal distribution for X than an assumption of positive dependence, and this can be seen from the figure, where the K–M estimator lies to the right of the true distribution. All quantiles for the K–M estimator are higher than the corresponding true quantile.

Delay Time Model

The delay time model (*see Repair, Inspection, and Replacement Models*) has been used a lot

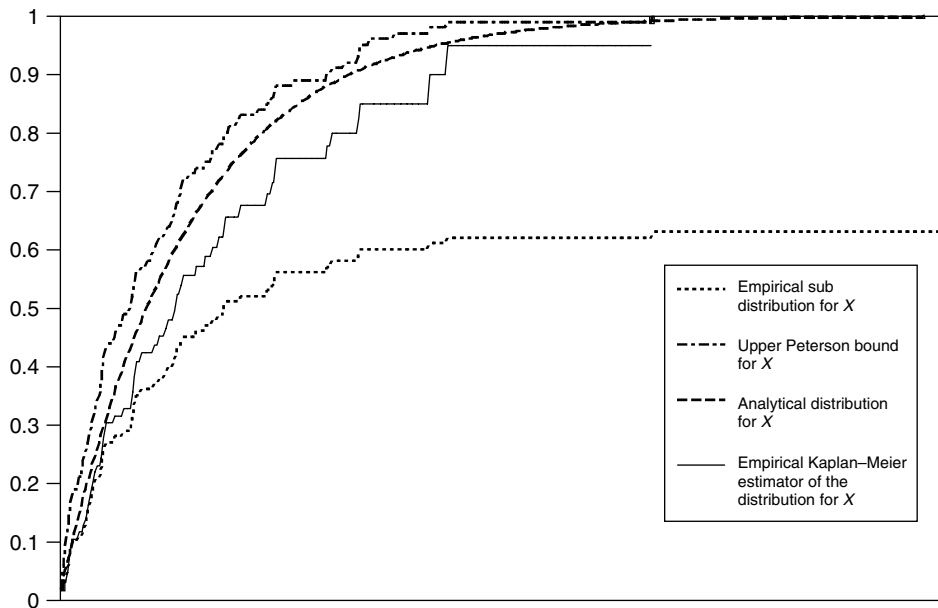


Figure 2 Illustration of the K–M estimator

in maintenance modeling [16]. It assumes that a degradation signal is given by a component, say at time U , after which there is another period, say V , until failure. The failure time is thus, $X = U + V$. To make a competing risk variant of this model, we can assume that a maintainer is walking round the plant, and if he happens to observe the equipment in its degraded state, then he will maintain it. Hence, the time to preventive maintenance is $Z = U + W$ where W is exponentially distributed. It is assumed that U , V , and W are mutually independent. A special case of this model discussed in [17] shows that if we assume that these variables are exponentially distributed, then $\Phi(t)$ is constant and a simple formula exists for estimating the model parameters.

Alert Delay Model

This model, presented in [18], assumes that

$$Z = p + \varepsilon \quad (6)$$

$$X = pX + (1 - p)X \quad (7)$$

where X and ε are exponentially distributed and independent of each other, $p \in [0, 1]$. This looks similar to the delay time model except that pX and $(1 - p)X$ are not independent. pX represents the time in which the system issues a warning signal, and ε represents the delay needed to carry out the preventative maintenance.

Random Signs

This model, discussed in [19], assumes that the preventative maintenance is related to the failure time by

$$Z = X + \delta\varepsilon \quad (8)$$

where $\delta = \pm 1$ is independent of X , and ε is a strictly positive random quantity. The random signs model represents a policy of preventative maintenance where in order to maximize its effectiveness, such maintenance should only be carried out as close to the time of failure as possible. Whether or not such maintenance succeeds in repairing the system before failure is represented by δ .

LBL Model

This model is a variation of the random signs model presented in [20]. The model assumes that

$$P\{Y \leq y | Y < X, X = x\} = \frac{H(y)}{H(x)} \quad 0 \leq y \leq x \quad (9)$$

$$H(x) = \int_0^x r_X(u) du \quad (10)$$

This means that the likelihood of an early intervention of preventative maintenance on a system is proportional to the unconditional failure intensity of the system.

Repair Alert Model

This model is presented in [21] and is a special case of the random signs model, which imposes additional structure by specifying a general function $G(t)$ in place of $H(t)$. $G(t)$ reflects the alertness of the maintenance team. This model generalizes the LBL model, which takes $G(t)$ to be the failure rate function of X . One of the properties of this model is that $\Phi(t)$ is increasing.

Random Clipping

We assume that X is exponentially distributed and that there is a variable $Z = X - W$ where W is a strictly positive random variable, independent of X . Z represents the time till a warning signal is issued by the system in some way. It was shown in [22] that regardless of the distribution for W , $X - W$ has the same distribution as X . This means that these warnings can be used to estimate important properties such as the mean time before failure of the system.

Mixed Exponential Model

This is a special case of independent competing risk models, which assumes that X is modeled by a pair of exponentially distributed variables, while Z is also exponentially distributed and independent of X . A useful property described in [23] is that $\Phi(t)$ is continually increasing from the origin. The underlying interpretation of this model is that the data consists of two distinct subpopulations, but that the maintenance regime is the same for both.

Summary

The identifiability competing risk problem is one that arises naturally in analyzing reliability field data. Taking the lead from work survival analysis, it is common to assume that the censoring is independent. However, consideration of the engineering context suggests that this is unlikely to be a reasonable assumption in many cases. The Peterson bounds give best possible upper and lower bounds for the marginals, and these are generally very wide, and therefore, not useful in practice. A number of models have been constructed however, that are built on an explicit engineering scenario. These models can be validated on engineering grounds, and are identifiable, making them useful for the practical interpretation of reliability field data.

References

- [1] <http://www.aiche.org/CCPS/ActiveProjects/PERD/index.aspx> (2007).
- [2] Procaccia, H., Aufort, P. & Arsenis, S. (1998). *The European Industry Reliability Data Bank EIReDA*, 3rd Edition, Crete University Press.
- [3] Cooke, R. & Bedford, T. (2002). Reliability databases in perspective, *IEEE Transactions on Reliability* **51**, 294–310.
- [4] Crowder, M. (2001). *Classical Competing Risks*, Chapman & Hall/CRC.
- [5] Tsiatis, A. (1975). A nonidentifiability aspect in the problem of competing risks, *Proceedings of the National Academy of Sciences of the United States of America* **72**, 20–22.
- [6] Peterson, A. (1976). Bounds for a joint distribution function with fixed subdistribution functions: application to competing risks, *Proceedings of the National Academy of Sciences of the United States of America* **73**, 11–13.
- [7] Crowder, M. (1991). On the identifiability crisis in competing risks analysis, *Scandinavian Journal of Statistics* **18**, 223–233.
- [8] Bedford, T. & Meilijson, I. (1997). A characterisation of marginal distributions of (possibly dependent) lifetime variables which right censor each other, *Annals of Statistics* **25**, 1622–1645.
- [9] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal American Statistical Association* **53**, 457–481.
- [10] Nádas, A. (1970). On estimating the distribution of a random vector when only the smallest coordinate is observable, *Technometrics* **12**, 923–924.
- [11] Miller, D.R. (1977). A note on independence of multivariate lifetimes in competing risk models, *Annals of Statistics* **5**, 576–579.
- [12] Van der Weide, J.A.M. & Bedford, T. (1998). Competing risks and eternal life, safety and reliability In *Proceedings of ESREL'98*, S. Lydersen, G.K. Hansen & H.A. Sandtorv, eds, Balkema, Rotterdam, Vol. 2, pp. 1359–1364.
- [13] Bedford, T. & Meilijson, I. (1995). In *The Marginal Distributions of Lifetime Variables which Right Censor each other*, *IMS Lecture Notes Monograph Series* 27, H. Koul & J. Deshpande, eds, Institute of Mathematical Statistics.
- [14] Zheng, M. & Klein, J.P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula, *Biometrika* **82**, 127–138.
- [15] Bunea, C. & Bedford, T. (2002). The effect of model uncertainty on maintenance optimization, *IEEE Transactions in Reliability* **51**, 486–493.
- [16] Christer, A. (2002). A review of delay time analysis for modelling plant maintenance, *Stochastic Models in Reliability and Maintenance*, Springer.
- [17] Hokstadt, P. & Jensen, R. (1998). Predicting the failure rate for components that go through a degradation state, in *Safety and Reliability*, S. Lydersen, G.K. Hansen & H.A. Sandtorv, eds, Balkema, Rotterdam, pp. 389–396.
- [18] Dijoux, Y. & Gaudion, O. (2006). A dependent competing risks model for maintenance analysis, the alert-delay model, *Proceedings of Degradation, Damage, Fatigue and Accelerated Life Models in Reliability Testing*, ALT'2006, Angers.
- [19] Cooke, R. (1993). The total time on test statistic and age-dependent censoring, *Statistics and Probability Letters* **18**, 307–312.
- [20] Langseth, H. & Lindqvist, B. (2003). *A Maintenance Model for Components Exposed to Several Failure Mechanisms and Imperfect Repair*, *Mathematical and Statistical Methods in Reliability*, World Scientific Publishing, pp. 415–430.
- [21] Lindqvist, B., Støve, B. & Langseth, H. (2006). Modelling of dependence between critical failure and preventive maintenance: the repair alert model, *Journal of Statistical Planning and Inference* **136**, 1701–1717.
- [22] Cooke, R. (1996). The design of reliability databases, Part I and II, *Reliability Engineering and System Safety* **51**(2), 137–146 and 209–223.
- [23] Bunea, C., Cooke, R. & Lindqvist, B. (2002). Competing risk perspective over reliability databases, in *Proceedings of Mathematical Methods in Reliability*, H. Langseth & B. Lindqvist, eds, NTNU Press.

TIM BEDFORD, BABAKALLI ALKALI
AND RICHARD BURNHAM

Compliance with Treatment Allocation

Consider a male in his early 70s. From his previous experiences, he thinks that if he catches a cold or influenza, it is almost certainly going to lead to a potentially serious or, at least, very unpleasant bout of bronchitis. Would it be worth his while to get an influenza vaccination (a “flu jab”) each autumn? Would a flu jab reduce his risk of bronchitis or similar flu-related complications? He decides to search the Internet for information and finds the results of a randomized trial designed to answer this question. But how does he evaluate the results? The investigators randomly allocated half of the trial participants (people aged 65 or over who are registered with their family physician or general practitioner (GP)) to receive a letter from their GP encouraging them to attend their local health clinic to get a flu jab. The other half do not get the letter, but continue to receive any health checkups, advice and treatment as usual (TAU). Those allocated to receive the letter also, of course, get access to TAU; so the trial is evaluating the difference in outcomes for patients who receive TAU+ letter and those who get access to TAU alone. The letters were sent out at the beginning of October, and the primary outcome measure was whether the participant suffered from serious flu-related complications over the following 6 months.

The authors report an intention-to-treat (ITT) effect (see **Intent-to-Treat Principle**) indicating that sending out the letter reduces the proportion who get complications from 0.10 to 0.08 (i.e., a drop of 2%). The trial was large, so the difference was statistically significant, but to the man, the difference looks trivial. It may be an important difference from the point of view of public health (or the cost of care to the health services), but is this small reduction of any consequence to him personally? Looking at the trial results more closely, he now notices that only 30% of those participants who received the letter took up the offer and actually received a flu jab. And 10% of the participants in the other arm of the trial had a flu jab, even though they had not received the letter inviting them to get one. The ITT effect is the effect of sending out the letter. It does not estimate the effect of receiving the flu vaccine; or, if it is treated as such,

it is an estimate that can be quite severely biased (attenuated) by both the failure of those offered the vaccination to take up the offer and by the receipt of the flu jab in many of those not invited to attend the clinic for the vaccination. What is the scale of this attenuation? How can we adjust for it to get a better idea of the reduction in risk of flu-related complications if we go out and get vaccinated?

A naïve approach to the problem would simply compare the outcomes in those people who were vaccinated with those who were not (the “as-treated” estimator – see below). But here we would not necessarily be comparing the outcomes in people with similar prognoses. There would be confounding, leading to so-called selection biases: we would not be comparing like with like. Those people who turn down the offer of vaccination are likely to be fitter on average than those who comply with the offer. Those who turn down the offer of vaccination will, on average, have a better treatment-free prognosis. The reverse will be true for those who are vaccinated irrespective of the letter offering them vaccination. Those are the people who because of prior medical conditions, for example, are thought to have a poor treatment-free prognosis. If we were able to take measurements of baseline (prerandomization) covariates that were associated with the receipt of vaccination (i.e., compliance) and also had an influence on the outcome (i.e., confounders), then we could make appropriate adjustments to the as-treated estimate. In practice, however, there will always be unmeasured or hidden confounders. So, where do we go from here? The aim of the present article is to illustrate how one might allow for selection effects (hidden, confounding) to obtain valid estimate of the effects of treatment receipt in those who comply with their randomized allocation.

Randomized Encouragement Trials and Randomized Controlled Trials (RCTs) with Noncompliance with Allocated Treatments

The hypothetical influenza vaccine trial described above is an example of what has been termed a *randomized encouragement trial* (for a real example, see Hirano *et al.* [1]). Participants are encouraged to change their behavior in a particular way (or not, if they are allocated to the control condition), but there is little expectation on the part of the trial

investigators that participants will fully comply with the request, or that there will not be a substantial minority in the control condition who change their behavior in the required way without actually being asked (by the trial investigators, at least) to do so. Another example involves randomizing expectant mothers (who are also cigarette smokers) to receive advice to cut down on or completely stop smoking during pregnancy [2]. Many of the mothers may fail to reduce their smoking even though they have received encouragement to do so. On the other hand, many may cut down on their smoking whether or not they have received the advice suggesting that they should do so. Randomized encouragement to take part in cancer screening programs is another familiar example [3–5]. Again, adherence to the offer of screening will be far from complete but, on the other hand, there are always people who will ask for the examination in any case. Similar, if not identical, in design and purpose to the randomized encouragement trial is Zelen's *randomized consent* trial [6, 7]. Zelen's design might be applied to both screening trials and to conventional clinical (treatment) trials. Randomized consent designs were introduced by Zelen to improve recruitment to randomized trials. In a conventional trial, the investigators seek informed consent from the potential participants prior to randomization. In the randomized consent design the order is reversed:

“After the patient's eligibility is established, the patient is randomized to one of two groups. Patients randomized to A are approached for patient consent. They are asked if they are willing to receive therapy A for their disease. All potential risks, benefits, and treatment options are discussed. If the patient agrees, treatment A is given. If not, the patient receives treatment B or some other alternative treatment that the physician is willing to recommend. Similarly, those patients assigned to B are approached for consent with a comparable discussion.” [7, p. 646]

If, say, treatment A is an established form of care, and B is a novel untested alternative, a variant of the above design (the single consent design) is to seek consent only from those patients who have been randomized to treatment B. Patients allocated to A are receiving exactly the same care as they would have received if they were not part of the trial, so their agreement to participate is not sought. In this situation, patients allocated to receive A do not have access to B (they may not even be aware

of its existence). Such designs may be objected to on ethical grounds, and they are potentially quite controversial. The point of describing them here, however, is not to promote their use, but to illustrate how treatment switching might be a direct result of the choice of the trial design.

Although it is not always adequately reported, it is increasingly obvious that most conventional randomized clinical trials (*see Randomized Controlled Trials*) (i.e., ones in which informed consent is sought prior to randomization) are subject to noncompliance. Adherence to the offered treatment or management package is rarely complete. People fail to turn up for therapy (Bloom's “no shows” [8]) or they fail to take sufficient amounts (or even any) of their prescribed medication. Sometimes, they are switched to a treatment that is not mentioned in any of the arms of the trial (rescue medication or emergency surgery, for example). It is important, however, that this noncompliance to the allocated treatment protocol does not necessarily indicate “delinquent” behavior on the part of the patient. It is often quite legitimate and in the interests of the patient's health or even survival. The patient's physician is always free to vary or switch the patient's medication if he or she feels that it is in the interest of the patient. The patient may be withdrawn from the trial altogether or offered a treatment unrelated to those being compared in the trial. For this reason, nonjudgmental terms such as “nonadherence” or even the completely neutral “departures from randomization” [9] are increasingly used instead of the apparently more disapproving “noncompliance”. Here, we used the terms “noncompliance” and “nonadherence” interchangeably.

Matthews [10, pp. 116–117] illustrates the problem of treatment switches in a trial to compare surgical and medical treatments for angina (European Coronary Surgery Study Group [11]). Data on treatment actually received, together with 2-year mortality, are shown in Table 1. Here the switches are likely to have resulted from clinical decisions, rather than noncompliance on the part of the patients. For example, some of the patients allocated to receive surgery may have been too ill for the procedure to take place (Table 1 shows that this subgroup has the highest 2-year mortality). Creed *et al.* [12] describe what we will refer to as the *day care trial* to compare the effects of day and inpatient treatment of acute psychiatric illness. Randomization produced 93 inpatients and 94 day

Table 1 A surgical intervention trial with both noncompliance and contamination

	Allocated to surgical treatment ($z = 1$)		Allocated to medical treatment ($z = 0$)	
	Actually received surgical treatment ($t = 1$)	Actually received medical treatment ($t = 0$)	Actually received surgical treatment ($t = 1$)	Actually received medical treatment ($t = 0$)
Number of patients (n)	369	26	48	323
Number who died (s)	15 (4.1%)	6 (23.1%)	2 (4.0%)	27 (8.4%)

patients. We then have the following sequence of events:

“Eight were excluded because of diagnosis or early discharge, leaving 89 inpatients and 90 day patients. Five randomized inpatients were transferred to the day hospital because of lack of beds, and 11 randomized day patients were transferred to the inpatient unit because they were too ill for the day hospital.” [12, p. 1382]

The apparently safe method of analysis for all of these trials is based on ITT. It is the analysis that is less open to abuse and is, quite rightly, the one expected by various regulatory and other official bodies. But, as was pointed out above, it is not necessarily an estimate of the effect we are particularly interested in. What is the efficacy of the treatment? What is the effect of getting the treatment as opposed to being allocated to get it? As an estimate of treatment efficacy, the ITT estimate will be attenuated in the presence of noncompliance and/or treatment switches. But one advantage of using the attenuated ITT estimate is that it is conservative (less likely to lead to unwanted type 1 errors). Perhaps the attenuation is not such a problem – it is a way of implicitly preventing us being too enthusiastic about the trial’s results (this is presumably what the regulator wishes to see). This may be true in a trial designed to test superiority of one treatment over another, but in a trial to evaluate noninferiority or equivalence, however, the last thing we need is attenuation (bias) toward the null effect. In this case the apparent equivalence may have arisen from the noncompliance (if no one takes their tablets, it does not matter what the tablets actually contain!). The ITT estimate is not always the safe, conservative one.

Finally, loss to follow-up (failure to obtain outcome data) is also a frequent problem for the analysis

of clinical trials. And, as might be expected, loss to follow-up is often associated with noncompliance with the allocated treatment. If a patient fails to turn up for the allocated therapy, he or she is less likely to turn up for an assessment of outcome. JOBS II was a randomized prevention trial to test the efficacy of a job training intervention in prevention of deterioration in mental health as a result of job loss [13]. The control group did not have access to the training package and 46% of those offered it did not actually take up the offer. The Outcomes of Depression International Network (ODIN) trial was undertaken to evaluate the effect of psychological treatments on depression in primary care [14, 15]. A characteristic of both JOBS II and ODIN is that there was considerable loss to follow-up and that probability of having missing outcome data was dependent on the previous compliance with therapy. In ODIN, for example, about 73% of the control group provided outcome data. Of those allocated to therapy, there was 92% follow-up in the compliers, but only 55% in the noncompliers (the figures depend on the exact definition of compliance used, but they illustrate the potentially dramatic effect of compliance behavior on dropout). So, we should not ignore adherence patterns, even if we are primarily interested in estimating ITT effects. Failure to do so is likely to result in biases and invalid significance tests [16].

We now turn to the challenge of estimation in the presence of noncompliance. We first introduce the required notation and definitions of types of participant. We then introduce the complier-average causal effect (CACE) and its estimation. Following this, we review and criticize other estimation methods (per-protocol and as-treated estimates). Finally, we have another look at our causal model assumptions and warn readers against uncritically accepting the results of any of these analyses!

Notation

The notation to be used in the rest of this article is summarized in Table 1. For simplicity, we are mainly concerned with a two-arm clinical trial to compare an active treatment and a control condition, or perhaps two potentially active treatments, A and B. We use the random variable z to indicate the outcome of randomization ($z = 1$ for the active treatment, say, and $z = 0$ for the controls). We assume that receipt of treatment is all-or-none and is restricted to one of the two options made available by the randomization (life is considerably more complicated if the participant receives some of treatment A and then, for example, moves to B, or even to an unforeseen option, C). Treatment receipt is represented by the random variable T ($t = 1$ for receipt of active treatment, for example, and $t = 0$ for the control condition). R indicates availability of follow-up (outcome) data with $r = 1$ when not missing and $r = 0$ when missing. Finally, a binary outcome variable, Y , indicates treatment failure ($y = 1$) or otherwise ($y = 0$).

The results of a trial are summarized as follows: s_{zt} is the number of participants *observed* to fail (outcome $y = 1$) if randomized to arm z and receiving treatment t . The count n_{zt} is the number of participants randomized to arm z and receiving treatment t with an observed outcome, and m_{zt} is the corresponding number with a missing outcome. The total number of participants randomized to arm z and receiving treatment t is given by $n_{zt} + m_{zt}$. Finally, the total number of participants allocated to arm z is $N_z = n_{z+} + m_{z+}$.

Types of Participants

We postulate that for any given trial, prior to randomization, we have the following three types of participants [17]. Always Takers (A) are participants who will always receive treatment, irrespective of their randomized allocation. These are the people we subsequently observe to be treated when they have been allocated to the control arm of the trial. We assume that (following from randomization) on average there will be an equal number of Always Takers in the control arm, but in this case, we have no way of knowing who they might be. These are people who if, contrary to fact, they had been allocated to the controls, would have received treatment regardless. Similarly Never Takers (N) are participants who

will never receive treatment, irrespective of randomized allocation. Again we can subsequently see who they are in the treatment arm, and we assume that there will, on average, be the same number allocated to the control arm (but, again, we cannot see who they are). Finally, we postulate that there are potential compliers – participants who will receive treatment if and only if allocated to the treatment arm ($z = 1$).

It is also logically possible that there might be defiers – those who will always do the opposite of their allocation, but here, we assume that there are none.

Note that, if we are interested in measuring outcome on a simple additive (linear) scale, the ITT effect is a weighted average of the ITT effects amongst the three types of patients. Furthermore, it should be reasonable to assume that for both the Always Takers and the Never Takers, the ITT effect is, in fact, zero. Randomization makes no difference to the treatment they actually receive and we conclude therefore that it has no effect on outcome. Note that this is a pretty strong assumption, and it is always open to challenge. However, if we are prepared to accept that it is true, then the following relationship holds:

$$ITT_{All} = \pi_c \times ITT_c \quad (1)$$

Here ITT_{All} and ITT_c are the ITT effects for everyone in the trial and for the compliers, respectively, and π_c is the proportion of potential compliers in the trial. ITT_c is known as the *complier-average causal effect* [17]. It is straightforward to show that the attenuation (π_c) is estimated by the difference between the proportion of participants receiving treatment in the treatment arm and the proportion receiving treatment in the control arm (equivalent to the ITT effect on receipt of treatment). Assuming that we have a valid estimate of ITT_{All} , then the CACE estimate is simply the estimate of ITT_{All} divided by the estimate of π_c . In the hypothetical flu jab trial introduced above, ITT_{All} was estimated to be 0.02. The proportion of compliers is estimated as 0.30–0.10 (= 0.20). The effect adjusted for attenuation (CACE) is estimated as 0.10. We should bear in mind, however, that we do not know (from what we have seen so far) the proportion with a poor outcome among the compliers. We need to examine the data in slightly greater detail (but unfortunately the data to let us do this is not always reported in published trial reports).

Estimation of Complier-Average Causal Effects

No Loss to Follow-up

Table 2(a) summarizes the characteristics of the three types of trial participants, indicating their prevalence (the π 's), expected outcomes (the β 's) and ITT effects on three different measurement scales (i.e., as a risk difference, relative risk, and odds ratio (OR), *see Relative Risk; Odds and Odds Ratio*). The key assumption is that randomization has no effect on outcome in the Always Takers and in the Never Takers. Table 2(b) gives the pattern of outcome data expected according to the model parameters described and introduced in Table 2(a). In the right-hand column of Table 2(b) are given the relative frequencies (f 's) observed from the trial. We proceed by simply equating observed frequencies with their

expectations and then solving the set of simultaneous equations to get the parameter estimates. That is,

$$\pi_c \beta_{c0} = f_{00} - f_{10} \quad (2)$$

$$\pi_c \beta_{c1} = f_{11} - f_{01} \quad (3)$$

$$\pi_c = p_1 - p_0 \quad (4)$$

Therefore,

$$\widehat{\beta}_{c0} = \frac{f_{00} - f_{10}}{p_1 - p_0} \quad (5)$$

and

$$\widehat{\beta}_{c1} = \frac{f_{11} - f_{01}}{p_1 - p_0} \quad (6)$$

From these estimates of $\widehat{\beta}_{c1}$ and $\widehat{\beta}_{c0}$, we can obtain the various treatment-effect estimates given in

Table 2 A simple causal model

(a) Types of trial participant and their average treatment effects

Type	Proportion	Expected outcomes ($\Pr(Y = 1)$)		Average effects of randomization (ITT effects)		
		$Z = 0$	$Z = 1$	Risk difference	Relative risk	Odds ratio
Always Takers (A)	π_a	β_a	β_a	0	1	1
Never Takers (N)	π_n	β_n	β_n	0	1	1
Compliers (C)	$\pi_c = 1 - \pi_a - \pi_n$	β_{c0}	β_{c1}	$\beta_{c1} - \beta_{c0}$	$\frac{\beta_{c1}}{\beta_{c0}}$	$\frac{\beta_{c1}(1 - \beta_{c0})}{\beta_{c0}(1 - \beta_{c1})}$

(b) The data expected according to the model

Randomized group	Treatment received	Participant combinations			Outcome	
		Types	Pr (combination)	Observed proportion ^(a)	$\Pr(Y = 1)$	Observed fraction
0	0	N or C	$\pi_n + \pi_c$	$1 - p_0$	$\pi_n \beta_n + \pi_c \beta_{c0}$	$f_{00} = s_{00}/n_{0+}$
0	1	A	π_a	$p_0 = n_{01}/n_{0+}$	$\pi_a \beta_a$	$f_{01} = s_{01}/n_{0+}$
1	0	N	π_n	$1 - p_1 = n_{10}/n_{1+}$	$\pi_n \beta_n$	$f_{10} = s_{10}/n_{1+}$
1	1	A or C	$\pi_a + \pi_c$	p_1	$\pi_a \beta_a + \pi_c \beta_{c1}$	$f_{11} = s_{11}/n_{1+}$

^(a) p_0 and p_1 are the observed fractions of subjects receiving treatment when randomized to the control and treatment conditions, respectively

$p_0 = 48/(48 + 323) = 0.129$	$p_1 = 369/(369 + 26) = 0.934$	$\widehat{\pi}_c = p_1 - p_0 = 0.805$
$f_0 = 29/(48 + 323) = 0.078$	$f_{00} = 27/(48 + 323)$	$f_{01} = 2/(48 + 323)$
$f_1 = 21/(369 + 26) = 0.053$	$f_{10} = 6/(369 + 26)$	$f_{11} = 15/(369 + 26)$

Table 2(a). It can also be shown that these method-of-moments estimators are also maximum likelihood [18]. We can obtain standard errors (SEs) and corresponding confidence intervals (CIs) for the treatment effects *via* maximum likelihood and the delta method [18], or *via* the bootstrap [19].

Here, we work through an example (see table above), based on the surgical trial data given in Table 1.

$$\begin{aligned}\text{Risk difference (ITT)} &= f_1 - f_0 = 0.053 - 0.078 \\ &= -0.025\end{aligned}$$

$$\begin{aligned}\widehat{\beta}_{c0} &= \frac{f_{00} - f_{10}}{p_1 - p_0} \\ &= \frac{27/(323 + 48) - 6/(369 + 26)}{369/(369 + 26) - 48/(48 + 323)} \\ &= 0.0716 \\ \widehat{\beta}_{c1} &= \frac{f_{11} - f_{01}}{p_1 - p_0} \\ &= \frac{15/(369 + 26) - 2/(48 + 323)}{369/(369 + 26) - 48/(48 + 323)} \\ &= 0.0405\end{aligned}$$

$$\begin{aligned}\text{Risk difference (CACE)} &= \widehat{\beta}_{c1} - \widehat{\beta}_{c0} \\ &= 0.0405 \\ &\quad - 0.0716 \\ &= -0.031 \\ &= (f_1 - f_0)/\widehat{\pi}_c \\ &= -\frac{0.025}{0.805} \\ &= -0.031\end{aligned}$$

$$\text{Relative risk (CACE)} = \frac{0.0405}{0.0716} = 0.566$$

$$\begin{aligned}\text{Odds ratio (CACE)} &= 0.0405 \times \frac{(1-0.0716)}{(1-0.0405)} \times \\ 0.0716 &= 0.547 \text{ (i.e., } \log(\text{OR}) = -0.603\text{)}.\end{aligned}$$

Coping with Nonignorable Missing Data

Now, let us assume that we have missing outcome data and that the missing data mechanism is what has been called *latently ignorable (LI)* [16]. The essence of this missing data model is that the probability of providing outcome data depends on the type of participant (which is latent or hidden), and also that randomization may have an effect on the proportion with follow-up data in the compliers, but has no effect in either the Always Takers or the Never Takers. Note, again, that these strong assumptions are always open to challenge. The revised model (now a joint model for both outcome and nonmissing data) is shown in Tables 3 and 4. Again, we can equate observed frequencies with their expectations, and solve the resulting simultaneous equations, as follows:

$$\pi_c \phi_{c0} \beta_{c0} = f_{00}^* - f_{10}^* \quad (7)$$

$$\pi_c \phi_{c1} \beta_{c1} = f_{11}^* - f_{01}^* \quad (8)$$

$$\pi_c \phi_{c0} = q_{00} - q_{10} \quad (9)$$

$$\pi_c \phi_{c1} = q_{11} - q_{01} \quad (10)$$

Therefore,

$$\widehat{\beta}_{c0} = \frac{f_{00}^* - f_{10}^*}{q_{00} - q_{10}} \quad (11)$$

and

$$\widehat{\beta}_{c1} = \frac{f_{11}^* - f_{01}^*}{q_{11} - q_{01}} \quad (12)$$

Again, we can provide SEs or CIs *via* maximum likelihood and the delta method [18], or *via* the bootstrap. We illustrate the calculations by reference

Table 3 Allowing for missing follow-up – types of trial participant and their average treatment effects

Type	Proportion	Probability of follow-up (Pr($R = 1$))		Expected outcomes (Pr($Y = 1$))	
		$Z = 0$	$Z = 1$	$Z = 0$	$Z = 1$
Always Takers (A)	π_a	ϕ_a	ϕ_a	β_a	β_a
Never Takers (N)	π_n	ϕ_n	ϕ_n	β_n	β_n
Compliers (C)	$\pi_c = 1 - \pi_a - \pi_n$	ϕ_{c0}	ϕ_{c1}	β_{c0}	β_{c1}

Table 4 Patterns of outcomes and missing data (assuming latent ignorability)

(a) The nonmissing data expected according to the missing data model

Z	T	Types	Pr (types)	Observed proportion	Nonmissing outcome	
					Pr($R = 1$)	Observed fraction
0	0	N or C	$\pi_n + \pi_c$	$1 - p_0$	$\pi_n \phi_n + \pi_c \phi_{c0}$	$q_{00} = n_{00}/(n_{0+} + m_{0+})$
0	1	A	π_a	$p_0 = (n_{01} + m_{01})/(n_{0+} + m_{0+})$	$\pi_a \phi_a$	$q_{01} = n_{01}/(n_{0+} + m_{0+})$
1	0	N	π_n	$1 - p_1$	$\pi_n \phi_n$	$q_{10} = n_{10}/(n_{1+} + m_{1+})$
1	1	A or C	$\pi_a + \pi_c$	$p_1 = (n_{11} + m_{11})/(n_{1+} + m_{1+})$	$\pi_a \phi_a + \pi_c \phi_{c1}$	$q_{11} = n_{11}/(n_{1+} + m_{1+})$

(b) The outcome data expected according to the combined missing data and outcome models

Z	T	Types	Pr (types)	Outcome	
				Pr($Y = 1$)	Observed fraction
0	0	N or C	$\pi_n + \pi_c$	$\pi_n \phi_n \beta_n + \pi_c \phi_{c0} \beta_{c0}$	$f_{00}^* = s_{00}/(n_{0+} + m_{0+})$
0	1	A	π_a	$\pi_a \phi_a \beta_a$	$f_{01}^* = s_{01}/(n_{0+} + m_{0+})$
1	0	N	π_n	$\pi_n \phi_n \beta_n$	$f_{10}^* = s_{10}/(n_{1+} + m_{1+})$
1	1	A or C	$\pi_a + \pi_c$	$\pi_a \phi_a \beta_a + \pi_c \phi_{c1} \beta_{c1}$	$f_{11}^* = s_{11}/(n_{1+} + m_{1+})$

to a hypothetical community care (or day care) trial for patients suffering from schizophrenia [20]. Table 5(a) shows the underlying structure of the data (but not that which is observed). The table

presents the observed number of treatment failures in each category, the total number with an observed (nonmissing) outcome and, finally, the total number randomized at the beginning of the trial. Table 5(b)

Table 5 Results of a hypothetical trial of community care (counts)

(a) The true data structure (counts)

		Complier	Never inpatient	Always inpatient
Randomized to day care ($z = 1$)	Failures	200	10	180
	Observed	400	50	200
	Total	500	200	300
Randomized to inpatient care ($z = 0$)	Failures	100	10	180
	Observed	300	50	200
	Total	500	200	300
True CACE		$= 200/400 - 100/300$ $= 6/12 - 4/12$ $= 1/6$		

(b) What we observe (counts)

	Randomized to day care ($z = 1$)		Randomized to inpatient care ($z = 0$)	
	Received day patient care ($t = 1$)	Received inpatient care ($t = 0$)	Received day patient care ($t = 1$)	Received inpatient care ($t = 0$)
Total	700	300	200	800
Nonmissing (n)	450	200	50	500
Missing (m)	250	100	150	300
Poor outcome (s)	210	180	10	280

shows the data that would actually be observed in such a trial. From these data, we can calculate the following:

$$\begin{aligned} f_{00} &= 280/1000 & q_{00} &= 500/1000 & p_1 &= 700/1000 \\ f_{01} &= 10/1000 & q_{01} &= 50/1000 & p_0 &= 200/1000 \\ f_{10} &= 180/1000 & q_{10} &= 200/1000 & p_1 - p_0 &= 0.5 \\ f_{11} &= 210/1000 & q_{11} &= 450/1000 \end{aligned}$$

It follows that, from equations (11) and (12), respectively

$$\begin{aligned} \widehat{\beta}_{C0} &= \frac{(280 - 180)}{(500 - 200)} = 1/3 \\ \widehat{\beta}_{C1} &= \frac{(210 - 10)}{(450 - 50)} = 1/2 \end{aligned}$$

The estimated risk difference (CACE) is $\widehat{\beta}_{C1} - \widehat{\beta}_{C0} = 1/2 - 1/3 = 1/6$. The corresponding OR (CACE) = 0.50 (i.e., logOR = -0.693).

Use of Negative Weights

Our statistical models specify that the number of negative outcomes for the Always Takers is the same, on average, in both arms of the trial. This is true under latent ignorability (but not necessarily other missing data mechanisms) and therefore also when there are no missing outcome data. The models also specify that the number of positive outcomes for the Always Takers is the same, on average, in both arms of the trial. And the same two equalities also apply to the Never Takers. Now, it may be the case that we have not used an exact 1:1 randomization. In this case, we simply make the necessary allowance for the lack of equality in the sizes of the two arms.

The method used here simply involves a straightforward regression (or *logistic regression* (see **Logistic Regression**), and so on, depending on the outcome) of outcome on treatment received where known noncompliers (the Never Takers in the treatment arm and the Always Takers in the control arm) are given a weight of -1. Everyone else is assigned a weight of +1. In the case of unequal randomization the weight of -1 for the known Never Takers is replaced by $-N_0/N_1$ and, similarly, the weight of -1 for the Always Takers is replaced by $-N_1/-N_0$, where N_1 and N_0 are the total number of participants

randomized to the treatment and control arms, respectively. Essentially the weights for the known Always Takers in the control are being used to “cancel out” or combine with the equivalent number of hidden Always Takers in the treatment arm. Similarly, the known Never Takers in the treatment arm cancel out the hidden ones in the control arm. What is left leads to an estimate of the effect of treatment in the compliers (i.e., the CACE). For further details, see [9] or [21]. SEs and CIs for the CACE estimates are obtained through the use of the bootstrap. Returning to the surgery trial summarized in Table 1, the use of this simple procedure using negative weights yields a risk difference of -0.031 (as before). The use of a thousand bootstrap replications yields an estimated SE of 0.023 and a 95% CI (percentile method) for the CACE of (-0.076, +0.015). A similar logistic regression yielded an estimated logOR of -0.602 (SE 0.058) with a 95% CI of (-1.713, +0.252). The point estimate is identical to that found earlier.

If we apply the same method to the hypothetical community care trial, we obtain a risk difference of +0.167 (SE 0.058) and a logOR of 0.693 (SE 0.281) with 95% CIs of (+0.057, +0.283) and (+0.265, +1.366), respectively. Again, the point estimates are those that we obtained earlier.

Potentially Invalid Estimators – as Treated and per Protocol

Most attempts made to allow for noncompliance use an “as-treated” or a “per-protocol” analysis. In the first, we simply look at the differences in outcomes for those who receive treatment and those who do not, irrespective of randomization. Per protocol looks at the same differences; however, data from those participants known to be noncompliers is first discarded. These methods only provide a valid effect of the treatment (efficacy) if compliers and noncompliers do not differ systematically in their disease state or prognosis. In practice, this is unlikely to be the case, so selection bias occurs. Heart disease patients who comply with their prescribed medication, for example, are also those who are likely to improve their diet or take more exercise and these changes, in turn, are likely to lead to a better outcome. Selection bias may often be reduced by adjustment for baseline covariates, but there is still no guarantee of an unbiased analysis. Returning to the heart surgery data in

Table 1, the estimated as-treated risk difference is $17/417 - 33/349 = -0.054$. The estimated per protocol risk difference is $15/369 - 27/323 = -0.043$. Both methods appear to be overestimating the impact of treatment.

Treatment Effect Heterogeneity

Potential treatment-effect heterogeneity is a well-known problem in generalizing the findings (ITT effects, for example) from participants in a given trial to the population as a whole. Treatment noncompliance, however, introduces more subtle problems for inferences concerning generalizability. We have demonstrated above how one might use data from a randomized clinical trial to estimate the effect of an intervention or treatment when there is a substantial amount of noncompliance with the allocated intervention. We have achieved this by concentrating on the average effect of randomization (ITT effect) on those who only receive the treatment if and only if they are allocated to receive it (i.e., the potential compliers). We have shown how to estimate this CACE using maximum likelihood both when there are no missing outcome data and when the missing data are latently ignorable. It is important that the reader realizes that we have done exactly what “CACE” implies we have done – we have estimated the effect of receipt of treatment in the compliers. We have no idea what the effect of treatment was in the Always Takers – we have no data on these people in the untreated condition. We cannot estimate the average effect of treatment in the Always Takers unless we are willing to assume that is the same as that in the compliers and what about the Never Takers? Again, we have no informative data. What if we carry out a second randomized controlled trial (RCT), and this time change our procedures so that we get much better compliance with the allocated interventions? What if we achieve complete compliance? Assuming that the treatment has not changed in any way, should we expect the treatment effect in the compliers in this second trial to be the same as in the first? Possibly not. Perhaps the treatment effect is associated with the participant’s propensity to adhere to the offered treatment. If we unrealistically assume that we have treatment-effect homogeneity (the treatment has the same effect in everyone), then we can use CACE estimates to infer the average effect of treatment for everyone in the

trial if, contrary to fact, we had persuaded everyone to adhere with their treatment. If we relax this unrealistic assumption and admit that the treatment effects vary from one participant to another, and, in particular, are likely to be associated with a participant’s willingness to adhere to the offered treatment, then we have to be very wary about making such inferences.

Acknowledgment

The author is a member of the UK Mental Health Research Network (MHRN)-funded Methodology Research Group.

References

- [1] Hirano, K., Imbens, G.W., Rubin, D.B. & Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design, *Biostatistics* **1**, 69–88.
- [2] Permutt, T. & Hebel, J.R. (1989). Simultaneous-equation estimation in a clinical trial of the effect of smoking and birth weight, *Biometrics* **45**, 619–622.
- [3] Cuzick, J., Edwards, R. & Segnan, N. (1997). Adjusting for non-compliance and contamination in randomized clinical trials, *Statistics in Medicine* **16**, 1017–1029.
- [4] Cuzick, J. (2001). Contamination and non-compliance in screening trials, in *Quantitative Methods for the Evaluation of Cancer Screening*, S.W. Duffy, C. Hill & J. Estève, eds, Arnold, London, pp. 26–33.
- [5] Mealli, F., Imbens, G.W., Ferro, S. & Biggeri, A. (2004). Analyzing a randomized trial on breast self-examination with non-compliance and missing outcomes, *Biostatistics* **5**, 207–222.
- [6] Zelen, M. (1979). A new design for randomized clinical trials, *New England Journal of Medicine* **300**, 1242–1245.
- [7] Zelen, M. (1990). Randomized consent designs for clinical trials: an update, *Statistics in Medicine* **9**, 645–656.
- [8] Bloom, H.S. (1984). Accounting for no-shows in experimental evaluation designs, *Evaluation Review* **8**, 225–246.
- [9] White, I.R. (2005). Uses and limitations of randomization-based efficacy estimators, *Statistical Methods in Medical Research* **14**, 327–347.
- [10] Matthews, J.N.S. (2000). *An Introduction to Randomized Controlled Clinical Trials*, Arnold, London.
- [11] European Coronary Surgery Study Group (1979). Coronary-artery bypass surgery in stable angina pectoris: survival at two years, *The Lancet* **1**, 889–893.
- [12] Creed, F., Mbaya, P., Lancashire, S., Tomenson, B., Williams, B. & Holme, S. (1997). Cost effectiveness of day and inpatient psychiatric treatment: results of a randomized controlled trial, *British Medical Journal* **314**, 1381–1385.

- [13] Vinokur, A.D., Price, R.H. & Schul, Y. (1995). Impact of the JOBS intervention on unemployed workers varying in risk for depression, *American Journal of Community Psychology* **23**, 39–74.
- [14] Dowrick, C., Dunn, G., Ayuso-Mateos, J.-L., Dalgard, O.S., Page, H., Lehtinen, V., Casey, P., Wilkinson, C., Vázquez-Barquero, J.-L., Wilkinson, G. & The Outcomes of Depression International Network (ODIN) Group (2000). Problem solving treatment and group psychoeducation for depression: multicentre randomised controlled trial, *British Medical Journal* **321**, 1450–1445.
- [15] Dunn, G., Maracy, M., Dowrick, C., Ayuso-Mateos, J.-L., Dalgard, O.S., Page, H., Lehtinen, V., Casey, P., Wilkinson, C., Vázquez-Barquero, J.-L., Wilkinson, G. & The Outcomes of Depression International Network (ODIN) Group (2003). Estimating psychological treatment effects from an RCT with both non-compliance and loss to follow-up, *British Journal of Psychiatry* **183**, 323–331.
- [16] Frangakis, C.E. & Rubin, D.B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes, *Biometrika* **86**, 365–379.
- [17] Angrist, J.D., Imbens, G.W. & Rubin, D.B. (1996). Identification of causal effects using instrumental variables (with discussion), *Journal of the American Statistical Association* **91**, 444–472.
- [18] Baker, S.G. & Kramer, B.S. (2005). Simple maximum likelihood estimates of efficacy in randomized trials and before-and-after studies, with implications for meta-analysis, *Statistical Methods in Medical Research* **14**, 349–367.
- [19] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, London.
- [20] Dunn, G. (2002). Estimating the causal effects of treatment, *Epidemiologia e Psichiatria Sociale* **11**, 206–215.
- [21] Kim, L.G. & White, I.R. (2004). Compliance-adjusted treatment effects in survival data, *STATA Journal* **4**, 257–264.

Related Articles

Causality/Causation

Effect Modification and Interaction

GRAHAM DUNN

Computer Security: A Historical Perspective

Computer systems have evolved from single isolated machines to billions of machines of myriad sizes, with massive interconnection over pervasive global networks. Security alternatives for computers and their networks have grown from external physical protective measures toward relying on the computers themselves to protect information essential to the vital interests, if not the survival, of enterprises. Today the threat of professional attack is serious and growing, and computers are increasingly a major source of risk.

The problem of security was not really noticed during the first two decades of use of electronic digital computers: the entire system was dedicated to a single user. Where early computers were used to process sensitive information, requirements for its security were addressed outside the computer with physical security for the facility. The computer itself was not really part of the security problem, nor was it part of its solution.

The demand for more efficient use of computers in the 1960s gave birth to multiplexing technologies, resource sharing operating systems, virtual machines, multiprogramming, and various other techniques that are still widely used. These computers were generally still used in so-called benign environments. In the 1970s and 1980s, the need to share not only resources but also information intensified. Often, this sharing was among users with disparate authorization for information access. This made controls internal to the computer system essential to mediate access of people to information in order to prevent unwanted modification and disclosure. These internal controls are the core of what we call *computer security* (see **Epidemiology as Legal Evidence; Game Theoretic Methods**).

As the 1980s approached, a published Consensus Report [1] stated, “It is a fact, demonstrable by any of several studies, that no existing commercially-produced computer system can be counted upon to protect any of its moderately knowledgeable users from having complete and undetectable access to any information in the system, no matter what kinds of so-called security features or mechanisms have been built into the system.” That truth has persisted into the twenty-first century, yet the emergence of the

Internet and the “information age” have dramatically increased our societal and economic dependence on computers.

Furthermore, computers are an increasingly critical part of marketed security solutions, with offerings like firewalls, virtual private networks, intrusion detection, file encryption devices, etc. However, the risks from using computers continue to grow, largely unabated, in spite of the billions of dollars (across the industry) spent on such solutions. So it is increasingly important to be able to assess the risks relating to computer security. The current (early twenty-first century) state of computer (in)security and associated risks is the focus of the rest of this article.

Theory of Computer Security

Computer security plays an important role in overall *system security* (see **Use of Decision Support Techniques for Information System Risk Management**). System security includes the broad range of *external* (to the computer) controls having to do with environment, people, procedures, and physical security practices that contribute to the security of the system. *Computer security* concerns the *internal controls* that the computer hardware and software – the trusted computing base (TCB) – provides to secure the enterprise information entrusted to its care. *Assurance* deals with the confidence, and the reasons for that confidence, one can have that the internal controls provided by a computer will provide the computer security expected – especially when faced with a deliberate, malicious professional attack.

Security of information inside a computer is very difficult to ensure. This is true because the “game of wits” between attacker and defender is unbalanced: an attacker bent on modification or disclosure has a substantial advantage. The system designer must anticipate every possible way to circumvent security and correct (prevent) them all; the attacker is interested in finding and using one. This is why computer security is fundamentally “hard”.

Numerous efforts to design reliable internal security controls by various *ad hoc* means have proved unsuccessful. Furthermore, technology investigations, including those by Anderson [2] and Schell [3], have consistently concluded that it is impractical to ensure that traditional internal controls contain no security

flaws. In fact, Harrison *et al.* [4] showed that it is theoretically impossible (i.e., mathematically “undecidable”), in general, to determine whether an arbitrary protection system can prevent a (possibly malicious) process from gaining unwanted access to information. This helps underscore the futility of relying (only) on testing to identify security holes: testing can only prove the presence of flaws and not their absence. Similarly, Karger and Schell [5] showed that depending on an audit log of accesses to secure a computer system is futile, because the attacker is likely to eliminate or mask any record of the attack on the security controls. Such log scrubbing is now routine, and indeed, there are automated kits to do so for major operating systems, which make them easy for even amateur attackers to use.

Not only do we want to create a computer that is secure, but we also need to know whether we have succeeded. In 1969 and 1970, a distinguished panel, headed by Dr Willis Ware of Rand Corporation, developed and published a seminal report [6] concluding that it is very difficult to evaluate a system and determine whether it provides adequate security. The previous paragraph summarized the major findings of the somewhat later Air Force panel headed by E. L. Glaser run under a contract with James P. Anderson (both members of the panel headed by Ware). This later panel published a report [2] that identified research aimed at solutions to problems identified by the Ware panel. The objective of this research was to understand the nature of the computer security problem and provide a framework for solutions. This, in turn, should make it practical to systematically evaluate the security of a system and use this evaluation as a basis for a risk assessment when using a system created within this framework. We will next examine the results of this research and development.

Nature of the Problem

A given system is secure only with respect to some specific security policy. A security policy is the set of rules that prescribe what information each person can access. The security policy is conceptualized in a way that is independent of (i.e., external to) computers.

Internal security controls are those rules enforced by computer hardware and software to govern access to system data and processing resources. In the 1970s,

the computer security profession identified a taxonomy of computer security functionality with four principal elements. These elements are elaborated on in the section titled “Framework for Solutions”. To support systematic risk assessment, these security function categories were embodied in the Trusted Computer System Evaluation Criteria (TCSEC) [7] and are still in widespread use today:

1. *Mandatory access controls* (rules that apply to all users in all circumstances).
2. *Discretionary access controls* (rules that can be changed by users and administrators authorized to do so, while the system is running).
3. *Identification and authentication* (I&A) (rules that control who may interact with the system and how their identity is verified).
4. *Audit* (rules that require certain events, trends, or changes in system state or configuration be recorded for later analysis and reporting).

Development of early military systems concluded that some portions of the system require particularly strong security enforcement. Specifically, this enforcement was necessary to protect data whose loss would cause extremely grave damage to the nation. Systems that handled such data, and that simultaneously included interfaces and users not authorized to access such data, came to be known as *multi-level*. This term was never intended to be limited in meaning to “hierarchical”, as it also applied to non-hierarchical, well-defined (but unrelated) user groups. This security policy is what is called *mandatory access control* (MAC) policy in the policy taxonomy, above. There are elements analogous to MAC policy in the commercial context that distinguish information essential to the crucial interests, or even the survival, of enterprises.

Much of the nature of the problem of computer security can be illustrated by examining common methods used to enforce the four policy elements, and, in particular, the MAC. These methods are reflected in a common categorization of computer control modes, using terms from the military context:

1. Dedicated mode

All users allowed access to the system are cleared for the highest level of information contained in the system and have a need-to-know for all the information in the system (i.e., it is dedicated to processing for users with a uniform need-to-know

for information at a given single security level). All users, equipment, and information reside within a protective boundary or security perimeter. Protection or security is enforced by physical means external to the computer (fences, guards, locked doors, etc.).

2. System high mode

The computer not only provides computation, but also must internally provide mechanisms that separate information from some users. This is because not all users of the system have a need-to-know for all the information contained in the system (but all are cleared for the highest level of information in the system). Discretionary access control (DAC), I&A, and audit controls are available in order to avoid (or recover from) user error.

3. Multilevel mode

The computer must internally provide mechanisms that distinguish levels of information and user authorization (i.e., clearance). This implies enforcement of a MAC policy. In this case, not all users of the system are cleared for the highest level of information contained in the system. This mode intrinsically requires that the computer operating system and hardware are what the military would term multilevel secure (MLS).

Each of these three modes of operation faces a very different computer security problem, i.e., requirement for internal controls.

1. The *dedicated mode* has no computer security responsibilities. Everything within the security perimeter in which the computer is located is considered benign. The computer system is not expected to seriously “defend” information from any of its users because they are considered non-malicious by virtue of their security clearances and need-to-know.
2. The computer security requirements for the *system high mode* are not very demanding, because the risk from vulnerabilities in these non-MAC controls is constrained. As in the dedicated mode, all users are considered benign, and while not every user may have need-to-know access to all data, all users are, nevertheless, cleared to see all data. DAC is therefore considered adequate to assist users in adhering to need-to-know policies.
3. The most serious risk (e.g., extremely grave damage to the nation) and the most difficult problem

are in providing adequate assurance of the MAC controls for the *multilevel mode*. Here, the computer system must protect the information from the user who is not cleared for it and who may attempt to install and execute *malicious software*. In effect, the computer system must become part of the security perimeter. The internal protection mechanisms must “assume the role” of the guards, fences, etc., that are indicative of the external security perimeter. Anything outside the security perimeter (including software) should be considered suspicious, since it may be malicious.

The remainder of this article will have a significant focus on computer security for this MAC policy.

Framework for Solutions

As noted above, the threats to computer security can be countered by providing access control over information on a computer to ensure that only specifically authorized users are allowed access. What we desire is a set of methods that make it possible to build a relatively small part of the system in such a way that one can even allow a clever attacker who employs malicious software to build the rest of the system and its applications, and it would still be secure. The theory of computer security gives us this.

Understanding computer security involves understanding and distinguishing three fundamental notions:

1. a security *policy*, stating the laws, rules, and practices that regulate how an organization manages, protects, and distributes sensitive information;
2. the *functionality* of internal mechanisms to enforce that security policy; and
3. *assurance* that the mechanisms do enforce the security policy.

The following introduces these three important notions and describes how they pertain to the reference monitor concept, which provides us with a set of principles that can be applied both to the design or selection of security features and to their implementation in ways that provide a high degree of resistance to malicious software. We then proceed to introduce the reference monitor concept as we apply it in designing secure computer systems.

Precise Policy Specification – What Does “Secure” Mean? We have already noted that a basic principle of computer security is that a given system can only be said to be “secure” with respect to some specific security policy, stated in terms of controlling access of persons to information. It is critical to understand the distinction between security policy and security mechanisms that enforce the security policy within a given computer system. For example, mechanisms might include type enforcement, segmentation, or protection rings [8]. These are all mechanisms that may be used within a computer system to help enforce a security policy that controls access of persons to information, but none of these is itself a security policy. Such mechanisms provide functionality that enables the implementation of access control within the computer system, but they do not directly represent rules in the security policy world of persons and information.

As noted earlier, a useful taxonomy for policy uses four principal elements of computer security functionality, and these in turn can be considered in two classes: access control policy is that portion of the security policy that specifies the rules for access control that are necessary for the security policy to be enforced. Supporting policy is the part that specifies the rules for associating humans with the actions that subjects take as surrogates for them in computers to access controlled information, most notably audit, and identification and authentication. We do not deal with supporting policy in greater detail.

The access control policies in turn fall into two classes: discretionary and mandatory. These two classes were described early on by Saltzer and Schroeder [9]. Both have historically been considered necessary for commercial as well as military security. We may characterize one control pattern as discretionary, implying that a user may, at his own discretion, determine who is authorized to access the objects he creates. In a variety of situations, discretionary access may not be acceptable and must be limited or prohibited. This is often described as enforcing “need-to-know” and in other cases reflects “role-based” controls.

A MAC policy contains mandatory security rules that are imposed on all users. It provides an overriding constraint on the access to information, with the highest assurance of protection possible, even in the face of Trojan horses and other forms of malicious software. This class of security policy can always be

represented as a (partially ordered) set of sensitivity labels for information (“classification”) and users (“clearance”).

As a practical matter, the choice between mandatory and DAC policies to support a particular security policy is, in most cases, tied to the penalty for which one would be liable if one violated the policy in the “paper world” – if no computers were being used. If the person responsible for protecting the information could get into “real trouble” (e.g., lose a job, get sued, be placed in jail, or even be severely reprimanded) for violating the policy in the paper world, then a MAC policy should be employed to protect the information in the computer.

Reference Monitor^a – Principles for a Security Perimeter also known as (aka TCB). The reference monitor provides the underlying “security theory” for conceptualizing the idea of protection, thereby permitting one to focus attention only on those aspects of the system that are relevant to security. The reference monitor provides a framework for extending the concept of information protection commonly used for people accessing sensitive documents to processors accessing memory. The abstract models of Lampson [10] were adapted to form the reference monitor concept, which is illustrated in Figure 1.

The reference monitor [2] is an abstraction that allows active entities called *subjects* to make reference to passive entities called *objects*, based on a set of current access authorizations. The reference monitor is interposed between the subjects and objects. The correspondence between reference monitor components and components of the computer system is reasonably clear: the *subjects* are the active entities in the computer system that operate on information on behalf of the system’s users. The subjects are processes executing in a particular domain (see below for

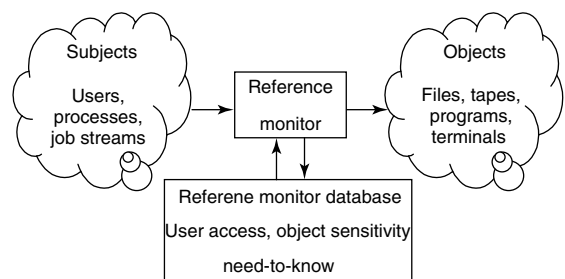


Figure 1 Reference monitor

definition) in a computer system (i.e., a $\langle$ process, domain $\rangle$ pair). Most of the subjects carry out the wishes of an individual, i.e., are a surrogate for a user.

The *objects* hold the information that the subjects may access. A domain of a process is defined to be the set of objects that the process currently has the right to access according to each access mode. Two primitive access modes, read and write, are the basis for describing the access control policy. While there are many kinds of objects in general, we can think of objects as well-defined portions of memory within the computer, most commonly memory segments. Files, records, and other types of information repositories can be built from these primitive objects, over which the reference monitor has total control. The reference monitor controls access to them by controlling the primitive operations on them – the ability to read and write them.

As a refinement on this concept, the section titled “Accreditation – Leveraging Technical Evaluation” discusses the simplifying technique of “balanced assurance” that allows constructing the named objects upon which DAC is enforced out of segments outside the reference monitor implementation. In fact, some reference monitors and their associated models address only the MAC policy.

The reference monitor itself is that most primitive portion of the computer system that we rely on to control access. We can think of implementing the reference monitor with a subset of a computer’s operating system and hardware. We shall find that, to be practical and efficient, the operating system software needs the assistance of computer hardware that is well suited to the task of providing security. A security kernel is defined [2] as the *hardware* and *software* that implement the reference monitor. (In a specific context, where the hardware context is static, “security kernel” is sometimes used in reference to just the software.)

Security Kernel as an Implementation of Reference Monitor. Employing a security kernel is the only proven way to build a highly secure computer system. Security kernel technology provides a theoretical foundation and a set of engineering principles that can be applied to the design of computer systems to effectively protect against internal attacks, including those that the designers of the system have never considered. A security kernel provides an effective security perimeter [11] inside which the information

is protected, no matter what happens outside. In fact, an implicit assumption in the design of a security kernel is that the attackers may build the remainder of the system, and yet the kernel will remain effective in prohibiting unauthorized access.

A security kernel [12] is a small, basic subset of the system software and supporting hardware that is applicable to all types of systems, wherever the highly reliable protection of information in shared-resource computer systems is required – from general-purpose multiuser operating systems to special-purpose systems such as communication processors. The implication of the term *security kernel* is that we can design a hardware/software mechanism that meets exactly the requirements for a reference monitor. This *reference monitor implementation* in a computer system must meet a set of requirements that were first identified in [2] and have been historically referred to as completeness, isolation, and verifiability:

- *Completeness* – all subjects must invoke the reference monitor on every reference to an object.
- *Isolation* – the reference monitor and its database must be protected from unauthorized alteration.
- *Verifiability* – the reference monitor must be small, well-structured, simple, and understandable so that it can be completely analyzed, tested, and verified to perform its functions properly.

The linchpin in providing the desired verifiable protection is a formal security policy model describing the core security policy to be enforced in this implementation of the reference monitor. Its properties are mathematically proved to be sound and correct. The system interface operations of the security kernel correspond to one or more rules that represent the security policy. All elements of the system must be demonstrated to be both necessary and sufficient for enforcement of the policy. The implementation is systematically traced to the policy. This includes showing that the entire code that is in the kernel is required for policy enforcement and system self-protection – this process is commonly termed *code correspondence*. The model enforced by most security kernels is derived from what is commonly referred to as the *Bell and LaPadula Model* [13]. This model provides rules that are provably sufficient for preventing unauthorized observation and modification of information.

The security kernel functions form its interface with the rest of the operating system and are derived from the formal security policy model. It turns out that, to meet the constraints of the security policy model for the MAC policy, the security kernel interface will provide a pure virtual environment for the rest of the system. This requirement results from the need to eliminate (sometimes subtle) methods of circumventing the security policy model's restrictions on the flow of information from one level to another. Frequently, specifications of the interface are expressed in a formal language capable of being analyzed by a computer program and of supporting a formal proof that the specifications conform to the requirements of the security policy model. Such a specification is termed a *formal top-level specification (FTLS)*.

Several different hardware platforms have been the target for security kernels. They each had common properties, however, including memory *segmentation* [9] and at least three hardware states for use in implementing *protection rings* [8]: one for the security kernel; one for the operating system; and one for applications. A system whose interface has been specified and proven in this manner is a system for which one has significantly increased assurance in its ability to verifiably enforce the security policy. A recent article by Anderson *et al.* [14], summarized its description of the problems addressed by the security kernel technology as follows:

As adversaries develop Information Warfare capabilities, the threat of information system subversion presents a significant risk. System subversion will be defined and characterized as a warfare tool. Through recent security incidents, it is shown that means, motive, and opportunity exist for subversion, that this threat is real, and that it represents a significant vulnerability. Mitigation of the subversion threat touches the most fundamental aspect of the security problem: proving the absence of a malicious artifact. A constructive system engineering technique to mitigate the subversion threat is identified.

A system so designed contains a security kernel and eliminates opportunities for professional attackers to use subversion to plant artifices, as any function that does not relate back to the policy model will be discovered in the "code correspondence" mapping process. This confirms there is no extraneous code and the system is shown to do what it supposed to do – and nothing more.

Taxonomy of Criteria – Risk and Security

As a practical matter, risk assessment requires that policy differentiate, or "classify", information on the basis of how much it is worth. There are two perspectives: (a) what the information is worth to an attacker and (b) what it would cost an enterprise if the secrecy or integrity of the information is lost. The value of information from the perspective of an attacker is critical to risk assessment because it is a factor in the motivation of and resources an attacker will employ, if given the opportunity. The cost of information loss from the perspective of the enterprise is critical because it yields an analytical justification for a level of expenditure to counter the risk. This is touched on further under section titled "Classification of Threats and Accepted Risk".

In the commercial world, the value of information, if measured at all, is likely to be measured in monetary terms. Metrics, in illustrative contexts, might include the average cost incurred per credit card record lost to hackers, the prospective decline in stock price (market capitalization) following a security incident, or the capitalized value of a proprietary technology advantage. In the public sector, an illustrative qualitative metric is given by regulations requiring the US government to classify information as "top secret" if its "loss would cause extremely grave damage to the nation".

In addition to the financial analyses hinted at above, the classification of data also yields the prerequisite analytical foundations for computer security. As discussed above, powerful computer security technology leverages a well-defined MAC security policy, and a MAC policy must express information classification.

Need for data classification

From a historical perspective, the use of computers to process sensitive information has clarified the value of a more systematic classification framework. In the early 1980s a "Data Security Leaders Conference" [15] recommended that managers "... assess the sensitivity of the data. Few companies examine the nature of data and rank it by sensitivity. Management must address this subject if they are going to take security seriously." The same report continues, calling this "... data classification. This is a new issue, one that has drawn little attention. But it is important, if management is going to know how much

money to spend to achieve an appropriate level of data security.” Unfortunately, although this was considered “a new issue” 25 years ago, its significance is often still unfamiliar to those assessing risk.

Need for security evaluation

Although data classification is necessary, it is not sufficient for an assessment. The same 1980s conference expressed concern for “evaluation procedures for assessing computer security. Without them, data security is a meaningless, costly exercise.” It went on to argue, “that users had too many choices for securing a system but no way of determining which would help really fend off potential data abusers.” Since that time, information security has become a multibillion dollar industry with many more choices. This cries out for criteria for assessing the “solutions” that the industry offers.

Objectives of Assessments

A taxonomy for criteria looks at the factors affecting quantitative analysis of risk in the realm of computer security. At the most general, risk assessments will be either Boolean in their ability to declare a system secure or not, or to offer some probabilistic estimate of the likelihood that the system is not secure. The ability (or inability) to state that a system is free of major security-impacting defects is precisely the factor that determines whether the system can be considered “secure”. For modern systems, a taxonomy must consider determined, intentional attacks that will often use malicious software. These attacks succeed (i.e., the system is not secure) because of the widespread lack of attention to what behaviors a system exhibits beyond those required or expected of it.

Penetration testing was among the first attempts at evaluating the security of a computer system, but the futility of substantially relying on that method for a security evaluation became known relatively early. In the early days of computer security, advocates of secure systems tried to follow a path of searching for ways to penetrate the systems’ controls. Their plan was that, failing to penetrate a system, they could then plausibly argue that there was no way to penetrate the system, since no way was known (to them). In this scenario, if a security hole is found, it can first be patched before the argument for security is made.

However, this argument suffers from both theoretical and practical difficulties. The approach presumes that one could test all possible programs to find any that led to a security penetration. If possible, this method of exhaustion would be effective, but it is far beyond the realm of feasibility. For any real computer, it would take so long that before the evaluation was finished, the sun would literally have burned out! Thus any evaluation conducted by exhaustion must be so incomplete as to be ludicrous.

The lesson that was learned is that a test can demonstrate the presence of an error but, short of exhaustion of every possible combination of inputs, it cannot demonstrate the absence of errors. Practically speaking, the effort spent in penetrate-and-patch techniques yields poor marginal return in terms of security [3].

Functional testing, penetration testing, and related quality assurance efforts can never explore all possible combinations and permutations of system state that may be used as “keys” to unlock trapdoors hidden in code. Protecting vital information from deliberate attack requires verification of what the system will *not* do, at least as much as knowing what the system *will* do. Only with a systematic mapping of the selected security policy to a formal security model, and of the model to a system (hardware and software) design (e.g., a security kernel), is it possible to determine what the system will do and *not* do. The objectives of risk assessment are related to three common aspects that are addressed next.

Certification – Reusable Technical Evaluation Results. One major component of risk assessment is systematic technical evaluation of the TCB that, as noted earlier, is at the core of computer security. The TCB is the totality of protection mechanisms within a computer system – including hardware, firmware, and software – the combination of which is responsible for enforcing a security policy. It creates a basic protection environment and provides additional user services required for a trusted computer system. The ability of a TCB to correctly enforce a security policy depends solely on the mechanisms within the TCB and on the correct input by system administrative personnel of parameters (e.g., a user’s clearance) related to the security policy [7].

A technically sound foundation for evaluation enables it to be used by a third party to accurately and objectively measure security trustworthiness of

a system. It does this in a manner which does not require the highly skilled and experienced scientists who wrote the criteria to apply it, yet which allows those less experienced who apply the criteria to come to the same conclusion as would those who wrote it. Those who apply the criteria should not be simultaneously writing and interpreting the criteria; that situation only leads to inequities in applying the criteria as interpretations differ between evaluations.

Historically, as pointed out by Bell [16], part of the original motivation for the TCSEC [7] and the Trusted Product Evaluation Program (TPEP) was to streamline and regularize the formal process of approving Department of Defense (DoD) systems for operation – certification and accreditation (C&A). While there were policies and procedures for assessing secure computer systems, there were problems with having engineering staff of each acquisition assess the underlying operating system or systems for security compliance. Not only was there no reuse of the analytical results, but also there was no guarantee of consistency across acquisitions. If all certification engineers facing a component could refer to a published summary of its security properties in the context of its TCSEC rating, then redundant analyses would be avoided and consistency across different acquisitions would be increased. Moreover, since computer security was a specialty field, the acquisition engineering staff was not always trained and current in the relevant security topics. The intention was to centralize and reuse the security evaluation results to obviate redundant analyses, to assure consistent results, and thus to provide an acquisition task with a solid, published product evaluation report.

Certification has been defined [17] as a comprehensive assessment of the management, operational, and technical security controls in an information system, made in support of security accreditation (see the next section), to determine the extent to which the controls are implemented correctly, operating as intended, and producing the desired outcome with respect to meeting the security requirements for the system. A reusable technical evaluation of the TCB can be major contribution to certification.

Accreditation – Leveraging Technical Evaluation.

Accreditation has been defined [17] as the official management decision given by a senior agency official to authorize operation of an information system and to explicitly accept the risk to agency

operations (including mission, functions, image, or reputation), agency assets, or individuals, based on the implementation of an agreed-upon set of security controls. This can be [18] made on the basis of a certification by designated technical personnel of the extent to which design and implementation of the system meet prespecified technical requirements, e.g., TCSEC, for achieving adequate data security. The management can accredit a system to operate at a higher/lower level than the risk level recommended for the certification level of the system. If the management accredits the system to operate at a higher level than is appropriate for the certification level, the management is accepting the additional risk incurred.

Historically, David Bell has noted [19] that in the DoD, the Trusted Product Evaluation Program was a program to evaluate and record the fidelity of commercial products to classes within the TCSEC, for later use by system certifiers and accreditors. Products successfully completing a “formal evaluation” were placed on the *evaluated products list (EPL)*.

Composition – Certification and Accreditation Implications.

In dealing with any complex system, an attractive strategy is one of “divide and conquer”. What is desired is an approach for dividing the trusted component of the system into simpler parts, evaluating each of the parts separately, and then validating the correctness of the way the parts are composed to form a single system enforcing a globally well-defined security policy. It is obviously desirable that the validation of the correct composition of evaluated parts be a relatively simple task.

In the general case, this is very difficult. David Bell has pointed out [16] that the growth of networking brought the “composition problem” to the fore: what are the security properties of a collection of interconnected components, each with known security properties? Efforts were made to derive methods and principles to deal with the unconstrained composition problem, with limited success. In the sense of product evaluation and eventual system certification, and accreditation, one needed to leverage the conclusions about components to reach conclusions about the whole, without having to reanalyze from first principles. Unfortunately, closed-form, engineering-free solutions were not discovered and may not be possible.

Fortunately, the simplicity of the reference monitor concept lends itself well to a “divide and conquer” evaluation strategy. In this section, two distinct strategies are overviewed. These are two strategies that are very useful in practice, but reflect constraints on composition that allow them to avoid the hard problems of unconstrained composition mentioned above. These ideas are an extension of the principles first described in [20], with the benefit of several additional years of experience and thought about how a complex TCB might be composed from a collection of simpler TCBs residing in individual protection domains.

Incremental evaluation of distinct physical components. The first strategy, the “partition evaluation” of a trusted distributed system or network, depends upon the notion that a complex system may be decomposed into independent, loosely coupled, intercommunicating processing components. This strategy is most suitable for trusted systems that have a complex physical architecture (e.g., distributed systems or networks). In such cases, it is not practical to require that an entire system be evaluated at once, so a strategy is needed for supporting incremental evaluations. The question is: how does one build a system from a set of individual components that were previously evaluated? The answer is found in two key concepts: (a) the concept of a “*partitioned TCB*”, in which individually evaluated components interact in a precisely defined fashion and (b) a “*network-security architecture*” that addresses the overall network-security policy.

These concepts enable the architecture to evolve without having to go back to reassess the role of each individual component each time a deployment consistent with the architecture is changed. This also led to the ability to recursively “compose” a set of individual components into a new single logical component that has a well-defined role within the network-security architecture and a well understood composite security rating.

TCB subsets within a single system. The second strategy, the “incremental evaluation” or “subset evaluation” of a TCB was used by Gemini Computers to internally structure the TCB of its Gemini Trusted Network Processor (GTNP) and was first publicly presented in [21]. It builds on the idea that a complex TCB may be divided into simpler TCB subsets,

each of which enforces a subset of the global security policy. The effect of such security policy decomposition, when allocated to the various TCB subsets, is to allow for a chain of simpler evaluations (each an evaluation of a subset) to be performed, leading to an overall conclusion that the composite TCB is correct. Unlike a TCB consisting of multiple partitions, incrementally evaluated TCB subsets are thought of as residing in the same, tightly coupled processing component. The subset strategy is particularly well suited for TCBs that enforce relatively complex security policies over virtual objects that are very different from the physical storage objects provided by the processor hardware. For that reason, the TCB subset strategy is particularly appropriate for the application to trusted database management systems (DBMSs).

The partition and subset evaluation strategies are compatible, and may be combined in various ways for TCBs that are complex both in architecture and security policy. This has been proposed, for example, for an embedded secure DBMS [22].

Balanced assurance. As we have noted, there are fundamental differences between the degree of protection from malicious software offered by the enforcement of a mandatory security policy and that offered by enforcing a discretionary security policy. This difference has led to the development of a useful technique, called *balanced assurance*, for enhancing assurance in the parts of the TCB where it matters most.

Using partitions or subsets, one of two positions for assurance in a TCB can be adopted. The more conservative approach requires that all assurance requirements for a particular evaluation class must be applied uniformly to the entire TCB. This is termed *uniform assurance*. The less conservative position requires the application of assurances to partitions or subsets only where the assurances are relevant to the security policy enforced by the partition or subset and provide a genuine increase in the credibility of security policy enforcement. This approach is called *balanced assurance* [23]. Balanced assurance is typically used when there is relatively little risk from low assurance for those TCB subsets not enforcing or supporting the mandatory security policy. This means that where we have a high assurance for the partitions and subsets responsible for the mandatory security policy, we can have a high assurance in the overall security of a network that

contains partitions (e.g., enforcing DAC and audit) that only meet the lower requirements for assurance. Balanced assurance is an important practical technique for achieving near-term high assurance where it counts in complex systems facing a serious threat.

The assessments discussed above are focused on the ability (or inability) to state that a system is free of major security-impacting defects, i.e., what its vulnerabilities are. The importance of vulnerabilities to risk depend on the threat a system faces, and that is what we will address next.

Classification of Threats and Accepted Risks

Global dependence on computers is growing exponentially as is the value of information assets on computers. The presence of these assets increases the threat of attacks against computers, which in turn increases the risk. We can relate threat to risk with the following notional expression:

$$\text{Risk} = \text{Threat} \times \text{Vulnerability} \quad (1)$$

We distinguish these two elements, i.e., “threat” and “vulnerability” as follows:

- *Vulnerability* – a quality or characteristic of the computer system (e.g., a “flaw”) that provides the opportunity or means of exploitation.
- *Threat* – the possible existence of one who participates in the exploitation by gaining unauthorized disclosure or modification of information such as accompanies information-oriented computer misuse.

The weak security and plethora of vulnerabilities in today’s platforms generally reflects the environments and threats of 10–20 years ago, not the current environment or that expected in 5 years. For example, a decade or two ago, personal computers performed local processing of data in relatively closed environments – as their name implies they were for “personal” use. Today, there is growing demand for personal computing platforms to perform high-value business-to-business e-commerce over the Internet.

The large number of vulnerabilities of common computing platforms poses potentially calamitous risks to high-value business transactions. This is because not only are the computers themselves relied on to protect the data, but also because there is a growing threat that the vulnerabilities will be

exploited. The commercial transactions processed by these platforms are at risk, as is sensitive data accessed by these platforms such as that targeted by industrial espionage, organized crime, and state sponsored terrorism.

Table 1 below shows a range of four classes of threats, and some characteristics of those threats. The top half of the table represents planned, concerted attacks while the bottom half represents more casual attacks. We now summarize each of those classes of threat.

All entities are trusted to behave themselves – no TCB Needed. This represents the implicitly postulated threat for what we have termed *dedicated mode* where access is not enforced by the TCB. Any information protection that is afforded is provided on an *ad hoc* basis by applications, generally as an afterthought or incidental to the primary function of the application. There may be unintended access to information. This may lead to embarrassment or inconvenience, but no security policy is violated. The threat is simply human error that leads to disclosure or modification that is basically probabilistic in nature. Such threats may involve a combination of human, hardware, and timing factors that when combined could allow an unintended (with respect to the purpose of the application) disclosure/modification of information.

Simple examples of this are a user typing sensitive information into a file that is thought to be nonsensitive, or including an unintended recipient as an addressee for E-mail. Users receiving information from this kind of disclosure or modifying information in this manner are often victims of circumstances and may not be malicious in their intent. Of course,

Table 1 Threat taxonomy

Type of attack	Threat	Effort to exploit	Motive for attack
Planned	Trapdoors (subversion)	Concerted planning	High value
	Trojan horses	Moderate	Moderate value
<i>Ad hoc</i>	Obvious holes in mechanisms	Easy	Ego boost (low value)
	Application weakness	Very easy	Vandalism

a trusted user might violate that trust and engage in some act of vandalism. They will find that easy to do, but that is not a “computer security” problem, since the computer was never designed to control such actions of users or applications. This class of threat has limited motivation and is focused on exploiting properties of the application not designed to be relied upon to enforce a security policy. The consequences are commonly called *inadvertent disclosure* and predominantly result from accidentally occurring states in the system or from human error.

Direct Probing – No Mandatory TCB Expected.

We can use the term *probing* to distinguish this class of threat from that in the above case, where an individual uses a computer system in ways that are certainly allowed, but not necessarily intended, by the system’s operators or developers. It is important to realize that the individual who is attempting probing is deliberate in his attempts. This introduces a class of “user” that computer system designers may not have seriously considered. Often, designs reflect that the systems are expected to operate in a “benign environment” where attempts to violate the system controls are presumed to be accidental, or at least amateurish and *ad hoc*. This limited design attention is the situation with most commercial systems today that offer only discretionary controls, i.e., no enforcement of a MAC policy.

This represents the implicitly postulated threat for what we have termed the *system high mode*. Because systems are presumed to be in a relatively benign environment (i.e., not under professional attack), the attacker may not have to exert much effort to succeed. By benign, we mean that this threat is typically from an amateur attacker, not a professional. A professional attacker is distinguished from the amateur by objectives, resources, access, and time. The professional is concerned about avoiding detection, whereas, amateur attackers are often motivated by a desire for notoriety or simple curiosity, as much as for gaining access.

Most of the current publicity surrounding computer and network-security breaches represent the work of such amateurs and involves frontal attacks that exploit either poor system administration or the latest hole that is not yet patched. A successful attack is commonly called a *penetration*. The threat deliberately attempts to exploit an inadvertent, preexisting flaw in the system to bypass security controls. The

penetrator will exploit bugs or other flaws to bypass or disable system security mechanisms. To succeed, the penetrator must depend upon the existence of an exploitable flaw. These frontal attacks are inexpensive to mount, and akin to a mugger in a dark alley not knowing if his next victim has any money or worse, a gun. In contrast, a professional seeking big money is not likely to use such methods given the unknown payoff and chances of being caught. We look at two classes of those professional threats next.

Probing with Malicious Software – TCB for MAC is Needed.

A professional will be well funded and have adequate resources to research and test the attack in a closed environment to make its execution flawless and therefore less likely to attract attention. This is in sharp contrast to the above *ad hoc* threats. The professional attacker understands the system life cycle and may surreptitiously construct a subverted environment by injecting artifices [24], e.g., malicious software, while participating in or controlling the efforts of a development team. In a large system with complex dependencies between modules, the opportunities for this are abundant. The well-motivated attacker will invest and plan. A system identical to the target will be obtained so it can be prodded and probed without worrying about being accidentally noticed by an alert operator, or even firewalls or intrusion detection systems. Attack methods will be tested and perfected.

The target can expect the professional to be attracted by information of value. As noted earlier this implies a need for classifying the worth of the information, and using that data classification scheme as an integral part of a MAC policy. The astute system owner will recognize the need for a TCB designed to enforce this policy. In other words, such a system will typically be a candidate to be operated in what we termed *multilevel mode*.

Finally, the professional is willing to invest significant time in both the development of the artifice as well as its use, possibly waiting years before reaping the benefits of the subversion. The subverter (who plants the artifice) may be – in fact, usually will be – an individual who is different from the attacker. A professional may have paid or persuaded someone else to perform the subversion and will at some point in the future, activate the artifice and attack the system. This may provide the professional attacker with a degree of plausible deniability not possible

with typical frontal attacks. For a state sponsored or other professional adversary such deniability is highly desirable.

The malicious software of this class of threat is termed a *Trojan horse*. The term *Trojan horse* for software is widely attributed [9] to Daniel Edwards, an early computer security pioneer, and it has become standard term in computer security. As with its mythological counterpart, it signifies a technique for attacking a system from within, rather than staging a frontal assault on well-maintained barriers; however, it does so without circumventing normal system controls. A Trojan horse is a program whose execution results in undesired side effects, generally unanticipated by the user.

A Trojan horse will most often appear to provide some desired or “normal” function. In other words, a Trojan horse will generally have both an overt function – to serve as a lure to attract the program into use by an unsuspecting user – and a covert function to perform clandestine activities. Because these programs are executing on behalf of the user, they assume all access privileges that the user has. This allows the covert function access to any information that is available to the user. The covert function is exercised concurrently with the lure function. This is essentially what the more sophisticated recent viruses do. This is a particularly effective option for the attacker owing to the fact that an authorized user is tricked into introducing the Trojan horse into the system and executing it. As far as any internal protection mechanism of the computer system is concerned there is no “illegal” actions in progress, so this type of attack largely eliminates the attacker’s exposure to discovery.

Subversion of Security Mechanism – Verifiable TCB is Needed. The final class of threat also uses malicious software in a planned professional attack, but uses it to subvert the very security mechanism itself. The distinction between a Trojan horse and this artifice as used in system subversion is important. The Trojan horse provides a function that entices the victim to use the software as well as a hidden function that carries out the malicious intentions of its designer. With the Trojan horse technique, system security mechanisms are still in place and functioning properly – the attacker’s code is executed with and limited by a legitimate user’s access permissions. In contrast, subversion does not

require action by a legitimate user. It bypasses any security mechanisms that the subverter chooses to avoid. The most common forms of artifices used in subversion are known as *trapdoors* [5].

Subversion of a computer system’s security mechanism involves the covert and methodical undermining of internal system controls to allow unauthorized and undetected access to information within the computer system. Such subversion is not limited to on-site operations, as in the case of deliberate penetration. It includes opportunities that spread over the entire life cycle of a computer system, including (a) design, (b) implementation, (c) distribution, (d) installation, and (e) use. The subverter is not an amateur. To be able to carry out subversive operations, the subverter must understand the activities that are performed during the various phases of a computer system’s life cycle. But none of these activities are beyond the skill range of the average undergraduate computer science major.

Recently, the “two-card loader” has gained notoriety. The two-card loader is named after the main-frame loader that was punched into two cards (too large to fit on one card). The hardware reads the two loader cards into a well-known location in memory, then transfers control to the first line of the loader program. The loader program, in turn, reads in the rest of the card deck and transfers control to the program contained therein.

A two-card loader subversion of an operating system reads in a malicious program as data, then transfers control to it. If a two-card loader is hidden in a commercial (or open source) operating system, then it can lie silently waiting for its trigger before doing its single, very simple job. A geek cannot find a well-written two-card loader. Exemplar subversions with a six-line “toehold” have been demonstrated [25], whereas Microsoft was unable to find an entire video game hidden in Excel before release.

Finally, a subversive artifice will typically include the capability of activation and deactivation. It was recognized long ago [3], based on experience with information warfare demonstrations, that “deliberate flaws are dormant until activated by an attacker. These errors can be placed virtually anywhere and are carefully designed to escape detection.” The artifice activation mechanism waits for the presence of some unique trigger. Examples are a particular stack state, an unlikely sequence of system calls or signals, or codes hidden in unused portions of data

structures passed into the kernel via system calls. The possibilities are endless. This trigger can be thought of as a key that can be made arbitrarily long from a cryptographic standpoint.

Security Evaluation Criteria Standards

The pervasive dependence on computers today gives us no choice, but to use them in the face of a threat from professional attackers. To responsibly manage the risk, we must not only create computers that are secure, but also verify that we have succeeded. It is essential that we have criteria for computer security evaluation.

If a platform is to withstand a planned hostile attack, a high degree of assurance in the security mechanism is required. Having the best technology in the world is of little value unless customers can know it has been effectively applied to their selected platform. Customers are often surprised at how hard that is, because of the general experience that through testing they can determine compliance in almost all areas of information technology. Unfortunately, testing is rather useless for demonstrating the security of a platform – as noted earlier, testing can only demonstrate the lack of security by exposing specific flaws. Simply offering a platform, with an absence of those specific known flaws, does not satisfy the customer's need for measurable trust.

Thirty years ago, the military identified that the greatest risks came when security tests of systems (e.g., “penetrate and patch”) found no flaws. Managers responsible for assessing risk become tempted to trust such systems despite a lack of real evidence of trustworthiness. Alternatively, citing standards and practices, e.g., of an industry alliance, may sound impressive, but does not suitably measure trust unless the standards and practices are proven and sound, and compliance is confirmed. Customers need products that are evaluated by objective third parties against proven criteria. Fortunately, various bodies, including European and US governments, have long recognized the need for standard security criteria. The resulting evaluation criteria have been evolving and have been used for over twenty years to evaluate numerous commercial products. The results of this evolution for risk assessment have been mixed, with some of the most powerful evaluation technology largely unused.

This historical evolution of security evaluation criteria has generally been along two paths that

we review below: system criteria and component criteria. The primary thread for system security has been around the TCSEC [7]. As system criteria, a range of systems issues have been integral to its use, such as secure composition and maintenance of security over the life cycle. The primary thread for component (or subsystem) evaluation has evolved to the common criteria [26]. This provides a framework for evaluations that do not necessarily answer the question “is the system secure”.

System Criteria (TCSEC)

As outlined earlier, the only known way to substantially address the threat of subversion is verifiable protection in a TCB that implements the reference monitor in a security kernel. Initially, the efforts to develop system evaluation criteria made no assumptions about the methods of the attacker. It was assumed that threat included professional attacks such as malicious software subversion. Therefore, in the early efforts the criteria for a system to be “secure” were synonymous with what later became the high end of the range of assurance levels in the TCSEC, which it called *Class A1*. Only later in the process was it asked if there were any values in defining lesser levels of assurance.

Historically, in 1977, the DoD Computer Security Initiative began, and an effort was made to consolidate the R&D gains in the theory of computer security that we have already outlined. In 1981, the DoD Computer Security Center was formed to focus DoD efforts to evaluate computer security. In 1983, following about 6 years of work by the DoD with the National Bureau of Standards (NBS), the center published the TCSEC [27]. It was published only after working examples of products meeting each requirement existed and after being “. . . subjected to much peer review and constructive technical criticism from the DoD, industrial research and development organizations, universities, and computer manufacturers” [27].

The center's charter was soon broadened, and the National Computer Security Center (NCSC) was established in 1985. The NCSC oversaw minor changes to the TCSEC, and it was established as a DoD standard in 1985 [7]. These evaluation criteria deal with trusted computer systems, which contain a TCB. The focus on *system* security was clear. The initial issue of the TCSEC explicitly states, “The

scope of these criteria is to be applied to the set of components comprising a trusted system, and is not necessarily to be applied to each system component individually [27].” It also emphasizes, “the strength of the reference monitor is such that most of the components can be completely untrusted.” The trusted components to be evaluated are the TCB.

The criteria are structured into seven distinct “evaluation classes”. These represent progressive increments in the confidence one may have in the security enforcement, and each increment is intended to reduce the risk taken by using that class of system to protect sensitive information. These increments are intended to be cumulative in the sense that each includes all the requirements of the previous. The classes and their names are:

- Class D: minimal protection
- Class C1: discretionary security
- Class C2: controlled access
- Class B1: labeled security
- Class B2: structured
- Class B3: security domains
- Class A1: verified design.

The seven classes are divided into four divisions, termed D, C, B, and A. The structure of the divisions reflects the threats they are intended to address. Division A of the TCSEC makes no assumptions about limiting the attacker. Division B hypothesizes that the attacker can subvert the applications, but not the operating system. Division C hypothesizes that the attacker uses no subversion at all. And, division D assumes that customers believe that attackers believe the vendor marketing claims. Notice that these threats correspond directly to the four classes of threat

we addressed in the section titled “Classification of Threats”. Only division A, called *verified protection*, is intended to substantially deal with the problems of subversion of security mechanism. This is shown in Table 2, which is essentially Table 1 augmented to include needed assurance.

Following initial publications of the TCSEC, the NCSC developed a number of interpretations and guidelines that assist in applying the principles of the TCSEC. These are collectively referred to as the *rainbow* series, because of their various cover colors. The TCSEC itself was commonly referred to as the *Orange Book*. One of the important topics of this series is composition.

Composition Criteria (Networks, Databases, Modules)

Commercial evaluations under the TCSEC made available trusted systems whose security properties were independently evaluated. However, there were two real-world concerns that motivated published interpretations of the TCSEC dealing with composition:

- the ability to enforce a variety of system security policies at varying levels of assurance; and,
- the ability to incrementally evaluate networks and systems based on well-defined modifications (e.g., the addition of a new subnetwork).

The initial systems evaluated were stand-alone computers, so some inaccurately assumed that use of the TCSEC was limited such that it could not be applied to a network of computers or layered applications, such as DBMSs, with their own security

Table 2 Platform trustworthiness needed to counter threats

Type of attack	Threat	Effort to exploit	Motive for attack	Assurance needed to counter
Planned	Trapdoors (subversion)	Concerted planning	High value	Division A
	Trojan horses	Moderate	Moderate value	Division B
<i>Ad hoc</i>	Obvious holes in mechanisms	Easy	Ego boost (low value)	Division C
	Application weakness	Very easy	Vandalism	Division D

policies. Since the TCSEC provides complete and reasonable criteria for evaluating systems (not just component), including networks as noted in [28] and [29], explicit interpretations and guidelines were published.

Networks. The section titled “Objectives of Assessments” has identified the major components in the strategy for a “partition evaluation” of a trusted network of independent, loosely coupled, intercommunicating processing components. The partition evaluation strategy was first introduced in [30]. It is the basis for Appendices A and B of the Trusted Network Interpretation (TNI) [18], which provides interpretations for the application of the TCSEC to networks. The TNI imposes the context of a “network-security architecture” that permits components of systems to be individually evaluated in a way that ensures that the eventual composition of the overall system will be secure.

A commercial example was the evaluation of Novell’s NetWare network under the TNI. Novell desired an evaluated system, yet was not in the business of building clients. They faced the question of how to specify the role of the client such that other vendors could produce clients that would be secure within the Novell network-security architecture. Novell completed three distinct but related evaluations: client; server; and network [31].

Balanced Assurance. The TNI Appendix A endorses the use of balanced assurance (though not by name) through its stipulation of a maximum class of C2 (with respect to assurance) for those TCB partitions not enforcing or supporting the mandatory security policy. Consistent with our earlier description of balanced assurance, this means that one can have a Class A1 network that contains partitions (e.g., enforcing DAC and audit) that only meet the Class C2 requirements for assurance.

Layered Applications. The initial focus was on developing an interpretation for a trusted DBMS. The final draft agreed to by the drafting working group of area experts recognized that the concepts applied to “extensions of the TCBs for applications that may need to provide access controls tailored to the specific design of the application subsystem as a whole”. This document [32] provided substantial rationale and

exposition of the underlying “TCB subset” concept. This is the basis for the “incremental evaluation” strategy identified in the section titled “Objectives of Assessments”. A significantly abbreviated version was published as the Trusted Database Interpretation (TDI) of the TCSEC [33], which provides interpretations for the application of the TCSEC to database DBMSs and other complex applications.

A delivered, commercial example of this was Trusted Oracle that was structured to exploit the properties of TCB subsetting. It included a mode whereby the MACs of the database were enforced by the mechanisms of the underlying operating system. This “evaluation by parts” solved the seemingly difficult problem of achieving a Class A1 database system, when neither the database vendor nor the operating system vendor was willing to provide the other with the proprietary evaluation evidence that would be necessary for a single system evaluation.

Ratings Maintenance Phase (RAMP). When changes are made to the TCB, its evaluation has to be reaccomplished. Using the TNI and TDI, it was clear that when the security architecture does not change, only the changed component, not the entire system, had to be reevaluated. This greatly simplifies maintaining the rating of the entire system. Similarly, when only selected modules internal to the TCB of a single component are changed, it should be possible to simplify the reevaluation. One of the most common and important internal changes is upgrading to new hardware as the technology changes, e.g., upgrading from a motherboard with an Intel 286 16-bit processor to an Intel Pentium 32-bit processor.

The NCSC put in place a codified rating maintenance phase (RAMP) process, so the vendor did not have to start over. The goal of evaluation is to put “secure” systems into the marketplace and into the field. With RAMP, the evaluation process served that goal, but did not supplant it. High-security systems are expensive to evaluate (tens of millions of dollars for an A1 system in the 1980s) and take a long time to evaluate (10 years from initiation to final evaluation). With RAMP, this can be reduced to a few months, or even weeks, so from the practical point of view, this is a critical tool for deploying secure systems.

Component Criteria (ITSEC, Common Criteria, FIPS 140)

After its publication, the DoD and other organizations in the United States used the TCSEC for security evaluations. In the meantime, some western European nations began developing different criteria – the Information Technology Security Evaluation Criteria (ITSEC). In 1990, a group of four nations – France, Germany, the Netherlands, and the United Kingdom – published the first draft. The draft criteria, while substantially influenced by the TCSEC, were significantly different in a number of areas [34]. Its most notable difference was that the ITSEC could be applied to components or subsystems that did not form systems.

However, its authors were insightful enough recognize the value of the TCSEC evaluation classes, and included in the ITSEC [35], an “Annex A” that defined five “functionality classes” F-C1, F-C2, F-B1, F-B2, and F-B3 that were derived from the TCSEC. These could be combined with their assurance levels E1, E2, E3, E4, E5, and E6 with the intent of expressing the same requirements as the TCSEC. For example, the criteria for “E6 F-B3” were intended to represent Class A1. In practice, several vendors at the lower assurance levels used the same evaluation evidence for evaluations in both the United States under the TCSEC and Europe under the ITSEC.

In the late 1990s, joint international efforts evolved a framework along the lines of the ITSEC into a new common criteria intended to replace both the TCSEC and the ITSEC. As with the ITSEC, common criteria evaluations need not provide a system context. The component or subsystem to be evaluated is called the *target of evaluation*. In general, evaluation of a subsystem may do the owner little good in terms of the risk to the systems, if overall system protection depends on software outside the target of evaluation. This can limit the value of a common criteria evaluation as a risk assessment tool, and the procurers of trusted systems may well have to perform much of the work of a TCSEC evaluation to do their own systems evaluation. As one example, a respected observer has pointed out [19]:

The Common Criteria does not include strong requirements for trusted distribution. Instead, it has a single delivery family, ALC DEL, which requires that a developer document procedures for delivery and use those procedures. Hence, the EAL levels do

not speak to trusted delivery at all. What one needs to prevent subversion of a high- or medium-security system is [Class] A1-style configuration management and trusted delivery.

In reality the common criteria is more of a metacriteria that provide a common language for expressing a particular criteria. It can be used to create what it defines as a “protection profile” that could come much closer to providing concrete criteria. Several security technologists and users have suggested a set of US government validated *protection profiles to mirror each of the TCSEC evaluation classes*, including the components defined in the TNI. This would be analogous to the “Annex A” of the ITSEC. So far, this suggestion for an explicit basis for applying the common criteria as a system criteria seems to have been actively resisted.

A similar challenge exists for standards for components or subsystems in other security related areas. For example, the application of cryptography has been greatly enhanced by standards such as FIPS 140-2 [36], which can make the engineering of cryptographic products like virtual private networks (VPNs) a lot more likely to be sound. Cryptography advances have indeed been considerable. However, current cryptographic products are mostly built on platforms having weak security. And, the recent neglect of the science of computer and network security suggests this will not be corrected in the near future.

As a result of this major weakness in cryptographic deployment, it has, in many cases, become what has been referred to as the *opiate of the naive*. This is not a unique insight of the authors, but has been noted by others as well. For example, one long-term expert noted [37]:

cryptography, in the form of digital signatures, public key certificates, and the like have become the “security architecture” for network based systems of the future. This trend totally ignores the fundamental fact that such encryption will only be as secure as the operating system structure in which it sits. . . . Contrary to accepted ideas, then, the use of cryptography actually enhances the need to reconsider security functionality and evaluation at the operating system and hardware levels. . . .

The common criteria and other component evaluation standards have led to a number of evaluated products, but a dearth of evaluable systems. A major contributing factor is that without something like the

suggested TCSEC/TNI equivalent protection profiles for the common criteria, there is no prescribed distinction between system evaluation and subsystem evaluations. This encourages the following:

- a willingness to make baseless assumptions about the behavior of “other” software subsystems, i.e., those falling outside the “target of evaluation”;
- a willingness to assume unenforceable prescriptions on the behavior of attackers;
- the classic logic error of treating all security policies as the same, and then concluding that certain techniques do not solve any of the problems because they fail to solve all of the problems. For example, because verifiable protection for MAC does not fully solve the problems of denial-of-service, some will overlook the fact that verifiable protection can control the risk from the procurement and deployment of a secure operating system that could be subject to subversion by a mortal enemy.

Toward Quantitative Risk Assessment

Over the past few decades, computer security has become an urgent need as computers have moved from tools for specialized use, to the dominant and essential component of the “information age”. Yet, even the casual observer would conclude that there are essentially no commercial operating systems in use today with adequate protection for information of significant value that are appropriate for use in environments where not all users are trusted (cleared) to see all information on the system. That being the case, *procurers and users would be well served by carefully assessing the risk* as part of their decision to make such deployments.

Dr David Bell, one of the most knowledgeable and experienced computer security professionals, shared this view when he recently painted [19] a bleak and alarming picture along the following lines. In today’s networks, MLS connections between networks are unavoidable. The threat of sophisticated attack or subversion makes low-security systems unacceptable for connections between isolated networks. High-security systems must be used. Very difficult computer- and network-security problems confront us today. This is best seen in the need for increased communication between different security domains and in the need

to reduce desktop clutter caused by separate workstations for each different security domain. Both of these situations call for high levels of security, both from a policy perspective and from first principles.

Dr Bell continues by noting that currently most of these critical needs are being filled by low-security components, whose deployments are contrary to security policy. Even in the absence of explicit policy, *such deployments are ill-advised, verging on irresponsible*. An attack on a weak connection no more sophisticated than the Cuckoo’s Egg [38] gives an attacker complete control. Subversion of a weak connection allows adversaries to mount attacks of their choice at any time in the future. Such attacks can steal information, alter critical information, or bring down computers, or entire networks. The worst subversion, called the *two-card loader*, cannot be prevented or even found in low-security systems [19]. Direct attacks like the Cuckoo’s Egg can only be thwarted by full mediation and self-protection, as required by B2/EAL5 or better systems. Complete prevention of subversion requires life-cycle security of the form found in Class A1 systems [19].

This state of affairs that he summarizes cries out for systematic risk assessment. Historically, computer security was viewed as a binary determination – either the system was verifiably secure, or its security was at best undecided (if not unknowable). This categorization did not come about immediately, but resulted from years of unsatisfactory attempts to use techniques like testing (e.g., “penetrate and patch”) to characterize the security of a system faced with professional adversaries. Quantitatively, the probability that commercial computer systems can be relied upon to enforce a coherent, logically complete security policy approaches zero.

In trying to quantify risk, it is tempting to try to quantify vulnerability. The above discussion indicates that most efforts to quantify vulnerabilities are not very promising. The reality of that is largely binary, and the primary uncertainty is the users awareness of vulnerabilities. It seems the implications have still not sunk in. Today, many still focus on how quickly the latest set of security “patches” is installed, when these responses to amateur attacks have little to do with the risk from serious threats, but rather with the awareness of specific vulnerabilities. There is little realistic expectation that serious vulnerability from things like maliciously installed trapdoors will show up on a patch list.

But, at the same time, it is useful to at least recognize that *data classification has moved toward quantitative risk assessment*. Recall that although quantifying vulnerability is difficult, one thing we noted that does bear distinguishing, is how “valuable” information is. This implies the need for “data classification”, which fortunately, in turn, directly supports a precise MAC policy. And we have already summarized the powerful theory for assessing the ability of a system to effectively enforce an MAC policy. Data may be classified numerically, as with a range of monetary values, or relatively, as the government classifying “secret” information as more valuable than “confidential” information. A number of sizable commercial enterprises have analogous classification schemes for sensitive data.

The use of this for risk assessment is illustrated by an example of how, in the past, the *US government quantitatively computed a risk index* when it used the TCSEC for computer security evaluation. The TCSEC was accompanied by a quantitatively framed policy [39] to guide users in selecting the minimum level of evaluation a system should meet. This defined a numerical “minimum user clearance or authorization ($R_{\min}$)” rating from 0 to 7 based on “the maximum clearance or authorization of the least cleared or authorized user.” It similarly defined a numerical “maximum data sensitivity ($R_{\max}$)” rating from 0 to 7 based on the value of the most sensitive

information being processed. It then computed the “risk index” which, in the simple case, was “determined by subtracting $R_{\min}$ from $R_{\max}$.” This was used to determine the “minimum security evaluation class” based on approximately what is represented in Table 3 below.

Although this is only an example, it shows how a technically sound evaluation scheme can be leverage for a systematic risk assessment.

The relationship between the TCSEC evaluation divisions and the threat is reflected in Figure 2 below. This figure also has a notional illustration of quantitative value to the user, of the various evaluation levels, with a rough correlation

Table 3 Minimum security class for risk index

Risk index	Minimum security class
0	C2
1	B1
2	B2
3	B3
4	A1
5–7	Beyond state of current technology

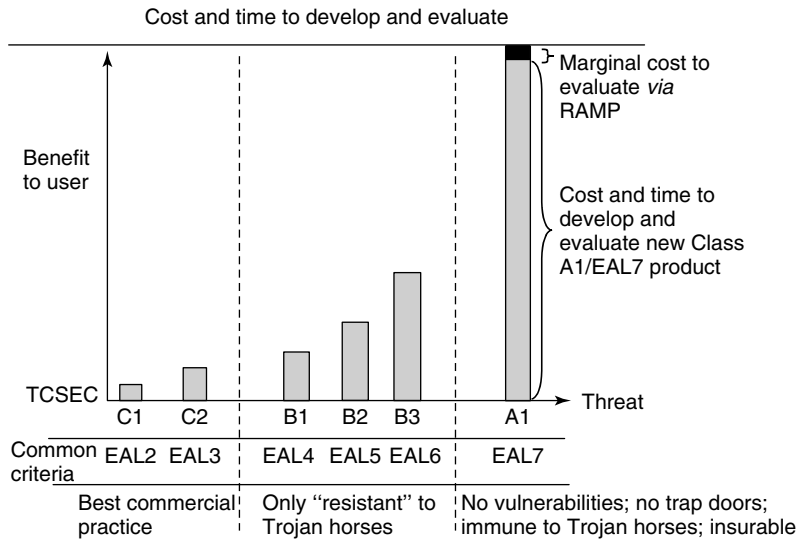


Figure 2 Cost and time to develop and evaluate trusted systems

between the TCSEC and common criteria evaluation levels.

In conclusion, we have seen that the theory of computer security is quite rich with solutions and tools. There is little reason that systematic risk assessment cannot effectively leverage the accumulated knowledge from research and practical experience of more than 30 years. The greatest past achievement of computer and network security is the demonstrated ability to build, and operationally deploy truly bulletproof systems having verifiable protection. This remains the most powerful solution available for many of today's hard computer problems.

Unfortunately, today many of those proven results for both developing and evaluating trusted systems languish and are unused. Risk assessment, as currently practiced, is largely in denial about the imminent and growing threat of professional attack using malicious software subversion, and also determinedly ignorant about the powerful technology that is available. The authors hope that the historical perspective summarized in this article will help rectify that.

End Notes

^aThis section makes substantial use of material prepared by one of the authors for "Security Kernel," contained in *Encyclopedia Of Software Engineering*, John J. Marciniak, editor-in-chief, John Wiley & Sons, 1994, pp. 1134–1136.

References

- [1] Lee, T.M.P. (1979). Consensus report, processors, operating systems and nearby peripherals, *AFIPS Conference Proceedings*, National Computer Conference, June 4–7, 1979.
- [2] Anderson, J.P. (1972). Computer security technology planning study, ESD-TR-73-51, Vol. 1, Hanscom AFB, Bedford (also available as DTIC AD-758 206).
- [3] Schell, R.R. (1979). Computer security: the Achilles' heel of the electronic air force, *Air University Review* **30**, 16–33.
- [4] Harrison, M.A., Ruzzo, W.L. & Ullman, J.D. (1976). Protection in operating systems, *Communications of the ACM* **19**(8), 461–471.
- [5] Karger, P.A. & Schell, R.R. (1974). Multics security evaluation: vulnerability analysis, ESD-TR-74-193, Vol. 2, Hanscom AFB, Bedford (also available as NTIS AD-A001120).
- [6] Ware, W.H. (1970). *Security Controls for Computer Systems: Report of Defense Science Board Task Force on Computer Security*, DTIC AD-A076-617/0, Rand Corporation, Santa Monica, reissued October 1979.
- [7] U.S. Department of Defense (1985). *Trusted Computer System Evaluation Criteria*, DOD 5200.28-STD.
- [8] Schroeder, M.D. & Saltzer, J.H. (1981). A hardware architecture for implementing protection rings, *Communications of the ACM* **15**(3), 157–170.
- [9] Saltzer, J.H. & Schroeder, M.D. (1975). The protection of information in computer systems, *Proceedings of the IEEE* **63**(9), 1278–1308.
- [10] Lampson, B.W. (1971). Protection, *Proceedings of the Fifth Princeton Symposium on Information Sciences and Systems*, Princeton University, pp. 437–443, Reprinted in *Operating Systems Review* **8**(1), 1974, 18–24.
- [11] Gasser, M. (1988). *Building A Secure Computer System*, Van Nostrand Reinhold Company, New York, pp. 162–186.
- [12] Ames Jr, S.R., Gasser, M. & Schell, R.R. (1983). Security kernel design and implementation: an introduction, *Computer* **16**(7), 14–22.
- [13] Bell, D.E. & LaPadula, L.J. (1975). Computer security model: unified exposition and multics interpretation, ESD-TR-75-306, Hanscom AFB, Bedford (also available as DTIC AD-A023588).
- [14] Anderson, E.A., Irvine, C.E. & Schell, R.R. (2004). Subversion as a threat in information warfare, *Journal of Information Warfare* **3**(2), 51–64.
- [15] IBM Education Center (1982). Information security issues for the eighties, *Data Security Leaders Conference*, San Jose, April 4–6, 1982.
- [16] Bell, D.E. (2005). Looking back at the bell-la padula model, *Proceeding ACSAC*, Tucson, 7–9 December 2005, pp. 337–351.
- [17] FIPS PUB 200 (2006). *Minimum Security Requirements for Federal Information and Information Systems*, National Institute of Standards and Technology.
- [18] National Computer Security Center (1987). *Trusted Network Interpretation of the Trusted Computer System Evaluation Criteria*, NCSC-TG-005.
- [19] Bell, D.E. (2006). *Looking Back: Addendum*, to [16].
- [20] Schaefer, M. & Schell, R.R. (1984). Toward an understanding of extensible architectures for evaluated trusted computer system products, *Proceedings of the 1984 IEEE Symposium on Security and Privacy*, Oakland, April 1984, pp. 41–49.
- [21] Shockley, W.R. & Schell, R.R. (1987). TCB subsets for incremental evaluation, *Proceedings of AIAA/ASIS/IEEE Aerospace Computer Security Conference*, Washington, DC, pp. 131–139.
- [22] Irvine, C.E., Schell, R.R. & Thompson, M.T. (1991). Using TNI concepts for the near term use of high assurance database management systems, *Proceedings of the Fourth RADC Multilevel Database Security Workshop*, Little Compton, April 22–25, 1991, pp. 107–121.
- [23] Lunt, T.F., Denning, D.E., Schell, R.R., Heckman, M. & Shockley, W.R. (1988). Element-level classification with A1 assurance, *Computers and Security* **7**, 73–82.

- [24] Myers, P.A. (1980). Subversion: the neglected aspect of computer security, Master of Science Thesis, Naval Postgraduate School, Monterey.
- [25] Lack, L. (2003). Using the bootstrap concept to build an adaptable and compact subversion artifice, Master's Thesis, Naval Postgraduate School, Monterey.
- [26] Common Criteria Implementation Board (CCIB) (1998). *Common Criteria for Information Technology Security Evaluation*, International Standard (IS) 15408, Version 2, ISO/IEC JTC 1.
- [27] DoD Computer Security Center (1983). *Trusted Computer Security Evaluation Criteria*, CSC-STD-001-83.
- [28] Shockley, W.R., Schell, R.R. & Thompson, M.F. (1987). A network of trusted systems, *Proceedings of the AIAA/ASIS/IEEE Third Aerospace Computer Security Conference*, Washington, DC.
- [29] Fellows, J., Hemenway, J., Kelem, N. & Romero, S. (1987). The architecture of a distributed trusted computing base, *Proceeding of the 10th National Computer Security Conference*, Gaithersburg.
- [30] Schell, R.R. (1985). Position statement on network security policy and models, *Proceedings of the Department of Defense Computer Security Center Invitational Workshop on Network Security*, New Orleans, March 1985, pp. 2–61, 2–70.
- [31] National Computer Security Center (1998). *Final Evaluation Report, Novell, Incorporated Netware 4.1.1 Server*.
- [32] National Computer Security Center (1988). *Trusted DBMS Interpretation of the Trusted Computer System Evaluation Criteria*, DRAFT.
- [33] National Computer Security Center (1991). *Trusted Database Management System Interpretation of the Trusted Computer System Evaluation Criteria*, NCSC-TG-021.
- [34] Branstad, M.A., Pfleeger, C.P. Brewer, D. Jahl, C. & Kurth, H. (1991). Apparent differences between the U. S. TCSEC and the European ITSEC, *Proceedings of the 14th National Computer Security Conference*, Washington, DC, pp. 45–58.
- [35] European Communities – Commission (1991). *Information Technology Security Evaluation Criteria*, Version 1.2, Office for Official Publications of the European Communities, Luxembourg.
- [36] FIPS PUB 140-2 (2001). *Security Requirements For Cryptographic Modules*, National Institute of Standards and Technology.
- [37] Caelli, W.J. (2002). *Relearning “Trusted Systems” in an Age of NIIP: Lessons from the Past for the Future*, *Colloquium for Information Systems Security Education*.
- [38] Stoll, C. (1989). *The Cuckoo's Egg*, Doubleday, New York, p. 29.
- [39] Department of Defense (1988). *Security Requirements for Automated Information Systems (AISs)*, DoD Directive 5200.28.

ROGER R. SCHELL AND EDWARDS E. REED

Condition Monitoring

Condition monitoring (CM) is a set of various techniques and procedures that people use in industry to measure the “parameters” (also called *features* or *indicators*) of the state/health of equipment, or to observe conditions under which the equipment is operating. The user’s main interest is in the equipment’s proper functioning (i.e., to operate as designed). The British Standards Institution Glossary gives a good and concise definition of CM: “The continuous or periodic measurement and interpretation of data to indicate the condition of an item to determine the need for maintenance” (BS 3811: 1993). CM is mainly applied for early detection of signs of malfunctioning and faults, and then for faults diagnosis and timely corrective or predictive maintenance. CM is also applied for operation/process control (e.g., to signal a jam on an assembly line and/or to stop the process), or safety control (checking the closure of a machine’s safety door) with a primary goal to prevent or reduce consequences of failures. Two common examples of CM are vibration analysis of rotating machines (e.g., centrifugal pumps, or electrical motors) and oil analysis of combustion engines (analysis of metal particles and contaminants in the lubrication oil), transmissions, and hydraulic systems. The whole combination of CM data acquisition, processing, interpretation, fault detection, and maintenance strategy is often called *CM system/program* (alternatively, *condition-based maintenance* (CBM)). The British Standards Institution Glossary’s definition of CBM is “Maintenance carried out according to need as indicated by condition monitoring”. An ideal situation would be to monitor conditions of all elements/parts of the machine, or, at least the ones most likely to develop significant problems. Complete monitoring is usually not possible technically, or is expensive, and it is thus important to select parts/elements of the system to monitor, and also select a method of monitoring. Common criteria for selection are based on experience and past information about failure modes and their frequencies, consequences of failures, such as downtime and cost, lost production, low quality of products, and so on, and availability of appropriate techniques. The main purposes of implementing a CM system are to be cost-effective by optimizing the maintenance program, and/or to avoid the

consequences of inadequate functioning and failures. CM is either an “off-line” procedure, when measurements/samples are taken and analyzed at predetermined moments of time (or when convenient), or an “on-line” procedure when the measurements are taken (and often analyzed) continuously, or at short intervals by the sensors permanently mounted on the equipment. Often, CM is a combination of various off-line and on-line procedures. A typical example of an off-line procedure is oil analysis, and of an on-line procedure is vibration analysis. Vibration monitoring is still commonly used as an “off-line” technique if the equipment deteriorates gradually. Now, due to advanced technology, oil analysis can, for some cases, be applied on-line (e.g., using wear debris light detectors).

The most common CM techniques/methods are vibration analysis, tribology (oil/debris analysis), visual inspections, current monitoring, conductivity testing, performance (process parameters) monitoring, thermal monitoring, corrosion monitoring, and acoustic (sound/noise) monitoring.

Monitored parameters/features can be direct, such as thickness (e.g., for brakes), amount of wear, corrosion, or cracks; or indirect, such as pressure, temperature, efficiency, vibration, and infrared and ultrasound images; or others, such as operating age. The parameters could also be operational (pressure, temperature, flow rate etc.), or diagnostic (vibration, amount and/or shape of metal particles in oil, water content in oil). Note that parameters/features are aggregated CM indicators calculated from collected raw CM data.

The methods of data/equipment condition assessment can be simple, such as measurement value checking, trending against time, or comparison with templates. They can be more advanced, such as mathematical models of deterioration and risk of failure, and artificial intelligence (AI) methods, such as neural networks and expert systems (ES).

Instruments/sensors for CM data collection/acquisition could be portable or mounted. Some instruments originated long time ago, such as temperature sensors, stroboscope (1830s), and piezoelectric accelerometer (1920s). Some instruments are more recent, such as fiber-optic laser-diode-based displacement sensors (late 1970s), laser counters combined with image analysis technology, or on-line transducers for wear particle analysis (1991). A lot of new instruments, now in use, have implemented software

for data processing, analysis, display, or wireless storage into a database.

Short History of Condition Monitoring

The development of CM technology was closely connected with the development of electronic instruments, transducers, microprocessor technology, software, mathematical modeling, and maintenance strategies. Following [1], the history of CM after its initial steps may be briefly separated into the following four stages: (a) From the 1960s to the mid-1970s, simple methods were used, combining practical experience and elementary instrumentation. (b) In the 1970s, the development of analog instrumentation was combined with the development of mainframe computers. At that time, clumsy vibrometers came into practice to measure and record vibration. Tape recorders were used to transfer data to computers, where the data was analyzed and interpreted. (c) From the late 1970s to the early 1980s, rapid development of microprocessors made possible development of much more convenient digital instrumentation which was able to collect the data, analyze it, and store the results. (d) In the mid-1980s, instruments became much smaller, faster, and the data was routinely stored on PCs for long-term use and development of maintenance strategies. Using CM was still a choice of advanced companies. Now, every more sophisticated piece of equipment arrives with built-in sensors/monitoring devices, and capability for data analysis, problem diagnosis, warning, and even maintenance recommendation. An everyday example is a new model of the private automobile.

Combination of emerging CM techniques, development of mathematical reliability methods, and new approaches to maintenance resulted in the development of new CM strategies. Initially, people predominantly used failure (breakdown/corrective)-based maintenance. Then people started using preventive (time-based) maintenance, and then, with introduction of CM methods, predictive maintenance (or CBM). This now resulted in many sophisticated and effective (but sometimes expensive) CM systems.

Implementation of Condition Monitoring

People usually apply CM to systems where faults and problems develop gradually, so they are able to

make timely maintenance decisions, such as (a) to stop the operation immediately (due to an imminent failure with significant consequences), (b) to stop at the closest convenient time (at the next planned shutdown), (c) to continue normal operation up to the next planned monitoring, without any particular action. People use collected CM data, also for the following: (a) prediction of CM parameters/features and estimation of remaining operating life, (b) long-term planning of further maintenance activities and need for spare parts, (c) fault detection and diagnostics. The obvious advantages of using CM are to be in much better control of operation, timely prediction of problems, reduction in downtime, reduction in maintenance costs, planning of activities, etc. Problems related to CM could be: difficulty to select an appropriate CM technique, possibility of high initial capital investment in instrumentation, implementation and education of personnel, necessity of standardized data collection, storage, analysis, and application of results, etc. Often, CM methods cannot provide very reliable results, and then engineers prefer to use their own judgment in combination with CM. CM may have only marginal benefits, particularly if applied to noncritical equipment, or applied with an inappropriate technique.

Basic Steps in Implementation and Use of CM Systems

The basic steps in CM implementation are as follows: (a) identification of critical equipment or systems, (b) selection of an appropriate technique/combination of techniques, (c) implementation of the technique (installation of instrumentation, setting baselines/alerts, and diagnostics), (d) data acquisition and processing, conditions assessment, and if necessary fault diagnostics and equipment repair, and (e) CM system review and adjustment after certain time. Selection of a CM system depends on several criteria, such as on the level of known relationship between the parameters and the conditions of the equipment, the ability of the system to provide timely warning of problems or deterioration, the availability of historical data and predefined and absolute standards for the assessment of equipment condition and fault diagnostics, and on the benefit of CM over an existing strategy.

An example of a CM implementation to the oil pump supplying a gas turbine (see [1]):

Main failure modes

Bearing, coupling or impeller wear, oil seal failures (possibly due to misalignment), out of balance, cavitation, overload, lack of lubrication or supply restriction.

Warning signs

Changes in vibration, temperature, current and performance (measured as pressures or flows), visible signs of leak.

Most critical failures

Bearing failure; damage could be significant.

Least critical failures

Oil leak and cavitation; no immediate risk.

CM techniques

Vibration analysis (prediction of failures caused by imbalance, misalignment, cavitation, wear, and lack of lubrication), general inspection (looking for leaks, noise, and changes in pump performance).

Setting up CM

Select and mark the measurement locations on the pump. Take monthly vibration readings on all motor and pump bearings, and on casing. In case of first warning signs, take readings more often. Set alarm and warning levels.

Reviewing

Review alarm and warning levels to optimize balance between failure consequences and excessive maintenance. Review regular measurement interval to decrease or increase it. Review usefulness of the whole CM system after a certain period of time by checking reduction in failure frequency, pump performance, increase/decrease in maintenance efforts, and cost-benefits.

More Details on Key Steps in CM Systems/Programs

The three major steps in a CM system are data acquisition, data processing, and data assessment in combination with decision making.

Data Acquisition (Data Information Collecting)

Data can be obtained from various sources, either by monitoring direct (thickness) or indirect (pressure, efficiency, vibration, cumulative stress parameters)

state parameters, and by using various CM techniques and instrumentation/sensors. Methods of data acquisition are local inspections, local instrumentation, process computers, portable monitoring equipment, and built-in monitoring equipment/sensors.

The most common CM techniques are as follows:

Vibration analysis (the most used and most convenient technique for on-line CM in industry today; intends to predict imbalance, eccentricity, looseness, misalignment, wear/damage, and so on).

Temperature analysis (monitoring of operational/surface temperature emission – that is, infrared energy sources using optical pyrometers, thermocouples, thermography, resistance thermometers).

Tribology (oil/wear debris analysis of the lubricating and hydraulic oil).

Performance/process parameters monitoring (measuring operating efficiency; possibly the most serious limiting factor in production).

Visual/aural inspections (visual signs of problems by a trained eye, such as overheating, leaks, noise, smell, decay; video surveillance of operation, utilization of visual instruments; these methods are usually cheap and easy to implement).

Others techniques, such as current monitoring, conductivity testing, corrosion monitoring, and acoustic (sound/noise) monitoring.

Data Processing

In the data processing step, the acquired raw CM data is validated and then transformed to a convenient form. After validation, the data may be either used in a raw form, such as from temperature, pressure, number and form of metal particles in oil, or in a transformed form (from vibration data, thermal images, and acoustic data). The data can be collected as a direct value (*value type*) of the measured parameter (oil data, temperature, and performance parameter), a time series (*waveform type*), where one measurement consists of information from a (usually short) time interval (vibration, acoustic data), and space-specific (*multidimension type*), where one measurement collects information over an area or volume (visual images, infrared thermographs, X-ray images, and ultrasound images). The *value type data* is either used directly, or some simple transformations is applied to it, such as rates of change, cumulative values, ratios between different variables, and

various performance measures and health indicators. The raw waveform and multidimension type data is always transformed (a procedure called *signal processing* or *feature/parameter extraction*) into a form convenient for diagnostic and prognostic purposes. The *multidimension type data* is most often transformed into plain images (infrared photo), or into some overall features of the image, such as density of points. The *waveform type data* is processed in time domain, frequency domain, or combined time–frequency domain. Some overall characteristics of the raw signal (e.g., vibration) are calculated in time domain, such as the mean, peak, peak-to-pick, standard deviation, crest factor, skewness, and kurtosis. In frequency domain, the contribution of different waves to the whole signal is analyzed (characterized by their frequencies and “weights” – amplitudes). A typical method of signal processing is to use frequency analysis (called *spectral analysis* or *vibration signature analysis*) by applying the Fast Fourier Transform (FFT) to the signal to calculate power spectrum. The overall measures of the spectrum are then used (overall root mean square (RMS)), as well as the measures of different frequency bands. Different frequency bands are indicative of different failure modes, or problems with different components of the system. This feature of the vibration signal is one which makes the vibration analysis the most popular of all CM techniques. The analysis must be adjusted to the operating conditions, notably to revolution per minute (RPM) for rotating machinery. For example, the changes of amplitudes in the frequency band of $1 \times \text{RPM}$ may be indicative of motor imbalance, and in $2 \times \text{RPM}$ that of mechanical looseness. In combined time–frequency domain, time and frequency characteristics of the signal are analyzed simultaneously using joint time–frequency distributions, or more advanced methods such as wavelet transforms.

Data Management and Interpretation and Use in Decision Support

The final step in a CM system is the analysis and interpretation of the extracted parameters. The parameters can be used for *on-line decisions*, as described in the implementation section above, for *fault diagnostics* (see **Fault Detection and Diagnosis**) (detection, isolation, and identifications of faults when they

occur), and for *prognostics* (see **Reliability Integrated Engineering Using Physics of Failure**) (prediction of time and type of potential faults).

Fault Diagnostics. For fault diagnostics, people use various statistical methods (statistical process control (SPC), pattern recognition, cluster analysis), system parameters, and AI methods. The most common method for on-line fault detection is to use *signal levels* for monitored parameters, that is, predetermined control values (such as in oil analysis and vibration). This method falls into the category of SPC that originates from statistical quality control. Typical signal levels are selected to show normal state, warning state, and alarming state of operation. Different parameters should be indicative of different problems, e.g., in oil analysis depending on metallurgy of components, or in vibration analysis depending on vibration frequencies of different parts. The signal levels are established either by manufacturer’s recommendations, or from theory, or found by experience and experimentation. The other possibility is to use *pattern recognition*, when a suitable parameter is recorded over time and compared with templates for normal operation and different faults, often using some methods of automatic recognition. The *system parameters* method is used when the system can be described by a mathematical model with system parameters being directly related to system conditions. The template (or *normal*) parameters of the system are estimated from the past data of healthy systems. Changes in current parameter values indicate changes in system conditions and/or development of certain faults. AI (see **Scenario-Based Risk Management and Simulation Optimization; Risk in Credit Granting and Lending Decisions: Credit Scoring**) methods are used for fault diagnostics, particularly in larger and more complicated systems. These include artificial neural networks (ANNs), fuzzy logic systems (FLS) (see **Early Warning Systems (EWSs) for Predicting Financial Crisis**) ES (see **Uncertainty Analysis and Dependence Modeling; Probabilistic Risk Assessment**), case-based reasoning (CBR), and other emerging techniques. The application of AI in fault diagnostics and other areas of CM plays a leading role in the development of intelligent manufacturing systems (IMS), which are now increasingly in use. Practical applications of AI methods are still not widespread due to the need for large amounts of past data (measurement and

faults histories) and qualified judgments from experts required for system training.

Prognostics. *Prognostics* use past and current CM data to predict the future behavior of the equipment by forecasting parameters/features, or estimating remaining useful life (RUL—expected useful life). Also, estimation of the probability of failure before the next inspection is of great interest, particularly when safety issues are important (e.g., in the nuclear industry). The main methods for prognostics are, as for fault diagnostics, statistical, model-based, and those based on AI. Of the statistical methods, *trending* is the most popular and simple method. The users extrapolate current and past measurements of parameters (e.g., using linear, or exponential trends) to predict when the parameters will cross warning (or alarm limits) to be able to prepare remedial actions. This method works well when measurements show clear, monotonic trend, but is less useful when measurements show large variations (for example, in oil analysis, from the authors' experience). Mathematical models of risk in the form of *hazard function* or *risk of failure* are a useful tool for risk prediction. Hazard function is particularly useful for short-term risk predictions when it includes operating time and measurements (often called *covariates* in this area), e.g., in a form of proportional-hazards model (PHM). It is also useful for long-term predictions of probability of failure and RUL when combined with a dynamic probabilistic model for parameters. The current hazard value can also be successfully used as a decision variable by comparison with warning/alarming hazard levels. *Model-based* methods use a mathematical model for the equipment's operation in time on the basis of its structure, physical properties, and measurements (e.g., Kalman filter (*see Nonlife Loss Reserving*) for wear prediction, or prediction of fatigue crack dynamics). The risk-based and model-based methods are often combined with economic consequences for optimization of maintenance activities.

Final Comments and Recommended Readings

The rapid development and widespread use of CM can be easily observed from the Internet. A search for *condition monitor* (industry) in product category

results showed 50 categories of products with about 7000 companies and products. For example, the category *Machine and Process Monitors and CM Systems* listed 214 companies, 148 for *CM and Machine Maintenance Services*, 706 for *Non-destructive testing (NDT)*, and 34 for *Oil Sensors/Analyzers*. Following [2], the *next generation of CM systems* are likely to focus on continuous monitoring and automatic diagnostic and prognostics, and thus on the design of intelligent devices (e.g., micro-electro-mechanical systems technology), which will be able to monitor their own health using on-line data acquisition, signal processing and diagnostics tools.

Suggested reading for a quick introduction to CM and CBM are [1, 3, 4]. For more information, see [5–7] (mostly on vibration) and [8]. Practical guides and overviews of CM techniques and instrumentation (though some of the instrumentation information is out of date) are [9, 10]. A good introduction to multiple sensor data fusion is [11]. For a more advanced introduction to model-based fault diagnostics see [12]. Various practical implementations of CM can be found in the previous references, and also in [13]. For an overview of AI methods with some applications, see [14]. A good introductory *review article* of CM, with application to machine tools, is [15]. For a detailed overview of CM data processing, diagnostics and prognostics, see [2], and for applications of PHM in CM, see [16]. For an overview of the *history* of CM (with emphasis on vibration analysis), see [17].

References

- [1] Barron, R. (ed) (1996). *Engineering Condition Monitoring: Practice, Methods and Applications*, Longman, Essex.
- [2] Jardine, A.K.S., Daming, L. & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance, *Mechanical Systems and Signal Processing* **20**, 1483–1510.
- [3] Williams, J.H., Davies, A. & Drake, P.R. (1994). *Condition-based Maintenance and Machine Diagnostics*, Chapman & Hall, London.
- [4] Wild, P. (1994). *Industrial Sensors and Applications for Condition Monitoring*, Mechanical Engineering Publications, London.
- [5] Collacott, R.A. (1977). *Mechanical Fault Diagnosis*, Chapman & Hall, London.
- [6] Davies, A. (1998). *Handbook of Condition Monitoring: Techniques and Methodology*, Chapman & Hall, London.

- [7] Mitchell, J.S. (1993). *Machinery Analysis and Monitoring*, 2nd Edition, PennWell Books, Tulsa.
- [8] Mobley, K. (2002). *An Introduction to Predictive Maintenance*, 2nd Edition, Butterworth-Heinemann, Amsterdam.
- [9] Bloch, H.P. & Geitner, F.K. (1994). *Machinery Failure Analysis and Troubleshooting*, 2nd Edition, Gulf Publishing Company, Houston.
- [10] Bøving, K.G. (1989). *NDE Handbook: Non-Destructive Examination Methods for Condition Monitoring*, Butterworths, London.
- [11] Hall, D.L. & Llinas, J. (2001). *Handbook of Multisensor Data Fusion*, CRC Press, Boca Raton.
- [12] Simani, S., Fantuzzi, C. & Patton, R.J. (2003). *Model-based Fault Diagnosis in Dynamic Systems Using Identification Techniques*, Springer-Verlag, London.
- [13] Rao, B.K.N. (1993). *Profitable Condition Monitoring*, Kluwer, Dordrecht.
- [14] Wang, K. (2003). *Intelligent Condition Monitoring and Diagnosis Systems*, IOS Press, Amsterdam.
- [15] Martin, K.F. (1994). A review by discussion of condition monitoring and fault diagnosis in machine tools, *International Journal of Tools Manufacturing* **34**, 527–551.
- [16] Jardine, A.K.S. & Banjevic, D. (2005). Interpretation of condition monitoring data, in *Modern Statistical and Mathematical Methods*, A. Reliability, S. Wilson, Y.A. Keller-McNulty & N. Limnios, eds, World Scientific, New Jersey, pp. 263–278.
- [17] Mitchell, J.S. (2007). From Vibration Measurements to Condition Based Maintenance, *Sound and Vibration*, January, pp. 62–75.

DRAGAN BANJEVIC AND ANDREW K.S.
JARDINE

Conflicts, Choices, and Solutions: Informing Risky Choices

In managing risky activities, decision makers (DMs) may have different available choices, payoffs or utilities (*see Utility Function*) for outcomes, as well as cultural expectations and ethical or moral values constraining how they believe they should behave. For example, a confrontation occurs in an international forum, such as in the development of certain terms for a convention or a treaty, where some stakeholders want to determine their best payoff, while minimizing that of another, before or during their negotiations (*see Considerations in Planning for Successful Risk Communication; Role of Risk Communication in a Comprehensive Risk Management Approach*). Similarly, an industry may assess what it can pay, before a fee or penalty schedule is codified. In these and many other situations, it is advantageous to describe alternatives or choices; positive or negative outcomes associated with each choice; probable effects of each choice, modeled by evaluating the probable consequences of different combinations of choices made by the DMs or stakeholders. The study of strategies and counterstrategies clarifies what a *solution* to a conflict (or partial conflict) situation is, when one exists, whether it is unique, why and whether it is stable (so that it is not plausible for players to deviate from it), conditions under which communicating with an opponent is or is not advisable, and ways to systematically change assumptions to test the sensitivity of results to those changes.

This article deals with key aspects of decisions by one or two stakeholders that confront alternative choices. In this confrontation, a *strategy* is a feasible act or option (including decision rules specifying what to do contingent upon information available at the time a choice must be made). Each stakeholder is assumed to have a discrete set of possible strategies that are mutually exclusive and fully exhaustive; these strategy sets are assumed to be common knowledge. A measure of value associated with each outcome is the *payoff* that each of the individuals has. Payoffs can be measured in units of money (usually transferable), *utility* (not necessarily transferable among players), or physical units (e.g.,

number of deaths or life-years lost), depending on the application.

Framework for Analysis

A Decision Maker with a Single Objective

Table 1 shows a problem with three mutually exclusive and collectively exhaustive choices (the rows) that a single DM must choose from. However, *Nature* – representing all factors over which the DM has no control – may affect the net gains or losses (in the cells of the table, the c_{ij}) by either taking on state 1 (s_1) or state 2 (s_2), with probability 0.50 each in this example. Note that the two states are mutually exclusive and collectively exhaustive: exactly one occurs. We can account for uncertainty by assigning probabilities to each of these two states, perhaps using historical data or expert judgment. The last column of Table 1 shows the calculation of the expected values of the acts/rows (defined by adding the values for each possible outcome, weighted by their respective probabilities). According to these calculations, a single DM who seeks to maximize expected value should choose act a_2 .

Before making a final choice, it is useful to measure the value of payoffs that might result from each choice. One obvious choice of scale is money. Another is *von-Neumann–Morgenstern (NM) utility*, which justifies the use of expected values (i.e., expected utilities) to choose among acts.

Two or More Decision Makers with Different Objectives

To depict a situation in which parties confront each other, one can replace “Nature” in the preceding example by a second rational, deliberating player. Options and payoffs for each player can still be tabulated. For example, here is a 3×2 , read *three-by-two*, representation (a matrix) of a hypothetical confrontation between two players, imaginatively called *Row* and *Column* (*blue* and *red* are also often used, especially in military applications).

	c_1	c_2
r_1	4	1
r_2	5	4
r_3	2	3

Table 1 Representation of choices affected by a neutral antagonist called *Nature*

Decision maker's alternative actions	States of Nature		Expected values
	s_1 (pr = 0.50)	s_2 (pr = 0.50)	
No action, a_1	$c_{11} = 200$	$c_{12} = -180$	$(0.5)(200) + (0.5)(-180) = 10$
Low cost intervention, a_2	$c_{21} = 100$	$c_{22} = -20$	$(0.5)(100) + (0.5)(-20) = 40$
High cost intervention, a_3	$c_{31} = 0$	$c_{32} = 0$	$(0.5)(0) + (0.5)(0) = 0$

The table shows Row's payoffs, measured in some units, such as dollars or utilities. If Row chooses the first row and Column chooses the first column, then Row receives a payoff of four units. In this type of zero-sum game, whatever Row wins, Column loses (and *vice versa*); thus, Row prefers large numbers and Column prefers small ones.

The largest of the row minima (*maximin*) is 4, and the smallest of the column maxima (*minimax*) is also 4: thus, the pair of choices (r_2, c_2) has the property that neither player can do better by deviating from (i.e., unilaterally choosing a different act from) this pair. The fact that the minimax (also called the *security level* for a player) and the maximin are the same identifies a "saddle point" solution (since, like the lowest point on the long axis of a saddle-shaped surface this point is simultaneously a minimum in one direction – it is the smallest payoff in row r_2 , and hence the smallest one that Column can get when Row selects r_2 – but is a maximum in the other (largest number in its column, and hence is the largest number that Row can achieve when Column selects c_2)). This is the "best" solution for both players, in the sense that neither can obtain a preferred payoff by a unilateral change of strategy (i.e., Row cannot do better by choosing a different row; and Column cannot do better by choosing a different column.) The celebrated minimax theorem [1] shows that any zero-sum game has a saddle point solution, provided that players are able to choose among alternative acts using any desired set of probabilities. In many practical situations, however, there is no saddle point among "pure" (nonrandomized) strategies. For example, consider the following 2×2 zero-sum game:

	c_1	c_2
r_1	2	-3
r_2	0	3

Because the minimax differs from the maximin, this game does not have a saddle point among pure

strategies. Further, no strategy dominates the other for either player (i.e., no strategy yields a preferred payoff for all choices that the other player could make). Nonetheless, if payoffs are measured using NM utilities, then a minimax strategy necessarily exists among mixed strategies.

Numerical solution techniques for zero-sum games (e.g., linear programming for general zero-sum games (Luce and Raiffa [2, Appendix 5]), Brown's method of fictitious play (Luce and Raiffa [2, Appendix A6.9]), the method of *oddmments* (Straffin [3] and Williams [4]), and various geometric methods) are well developed. For the game,

	c_1	c_2
r_1	2	-3
r_2	0	3

a simple geometric technique suffices: the two payoff lines represent Row's optimal probabilities for mixed strategy. The approach is depicted in Figure 1, as seen by Row, where the vertical axes measure Columns' payoffs and the oblique lines show the strategies for Row.

The solution (obtained algebraically) is 0.375 and thus Row 1 should be played with probability 0.375 and Row 2 with probability 0.625. Column will play strategy 1 with a probability that equals 0.75 and strategy 2 with probability that equals 0.25. Observe that the original game did not have a saddle; the approach requires assessing that the 2×2 game does not have either a saddle or dominance by a strategy, and then proceed as developed. The importance of this technique is that it allows for the solution of $n \times k$ zero-sum games by assessing mixed strategies in 2×2 subgames of the original $n \times k$ game (a result that is guaranteed by the minimax theorem (Straffin [3]), due to John von Neumann).

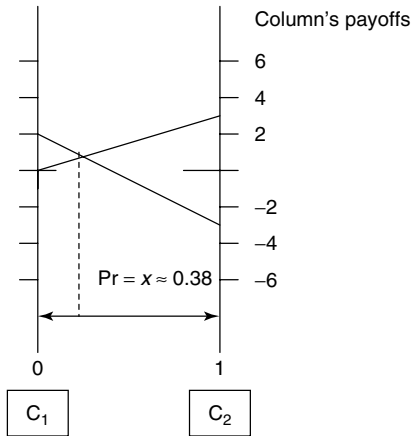


Figure 1 Graphical solution of two-person games: finding Row's probabilities

Beyond the 2×2 Game

A DM can use alternative criteria for choosing one course of action over another. We further discuss the *minimax*, *maximin*, and *expected value* using a situation in which Row now faces Nature in a 4×4 game (Straffin [3]):

	n_1	n_2	n_3	n_4
r_1	2	2	0	1
r_2	1	1	1	1
r_3	0	4	0	0
r_4	1	3	0	0

There is no row dominance. That is, no row has values that are at least as great in every column (and strictly greater in at least one column) as the values in all other rows. However, column 2 dominates all other columns. The minima of the rows are $[0, 1, 0, 0]$ with Row's maximin equal 1. The maxima of the columns are $[2, 4, 1, 1]'$ with Column's minimax equal 1 utility (' denotes transpose, i.e., writing a column as a row or *vice versa*). There are two saddle points at r_2, n_3 and at r_2, n_4 . If Row believes that all four columns are equally probable, then the strategy with the largest expected utility is r_1 , with $E(r_1) = 0.25 \times 2 + 0.25 \times 2 + 0.25 \times 0 + 0.25 \times 1 = 1.25$.

We can also calculate the *regrets* for Row, defined as the difference between the best payoff and the alternative payoffs, conditioned on each n_j , as shown

below:

	n_1	n_2	n_3	n_4
r_1	$2 - 2 = 0$	$4 - 2 = 2$	$1 - 0 = 1$	$1 - 1 = 0$
r_2	$2 - 1 = 1$	$4 - 1 = 3$	$1 - 1 = 0$	$1 - 1 = 0$
r_3	$2 - 0 = 2$	$4 - 4 = 0$	$1 - 0 = 1$	$1 - 0 = 1$
r_4	$2 - 1 = 1$	$4 - 3 = 1$	$1 - 0 = 1$	$1 - 0 = 1$

In this case, the row maxima are $[2, 3, 2, 1]'$ and, if we were to adopt the *minimax regret criterion*, then the optimal choice for Row by this criterion would be r_4 . If the expected regret was calculated using equal probabilities (pr) for the four states of nature ($\text{pr}(n_j) = 0.25$), then the result would be ambiguous (they would be r_2, r_3 , and r_4) because there are three equal and largest expected utility values.

Because different decision criteria lead to different optimal choices, several axioms have been developed to guide their selection. Some of those axioms are summarized in Table 2 and are related to the criteria for justifying that choice.

Searching for a Common Ground: Nash Equilibrium

The process of understanding rational decisions can be shown by the same normal form representation of strategies, $S_{i,j}$, consequences, $C_{i,j}$, and probabilities adopted in the previous section with the added understanding of the possibility of choices of strategies that are optimal. We use 2×2 games in which DMs can either cooperate or not. Some games do not allow sequential choices, while others do. Strategies, as course of actions, say cooperate, defect, attack, retreat, and so on, can be either pure (no probabilistic assessment is used: deterministic numbers are used) or mixed (probabilities play a role in assessing the eventual outcome from the situation confronting the DMs). A 2×2 game has two strategies per DM: for example, either defect or cooperate, but not both. We will discuss games that are effectuated simultaneously; further information is given in Burkett [5].

Pure Strategy Games

We begin with situations in which pure strategies are used to study what might happen. Consider, for example, the following situation in which DM1 and DM2 cannot cooperate – speak to one another (Table 3). The consequences (outcomes or payoffs)

Table 2 Selected axioms for decisional criteria^(a)

Axiom	Description	Criterion met by	Criterion not met by
Symmetry	Permuting rows or columns should not change the best choice	Maximin, expected value, regret	NR ^(b)
Dominance	When every row element is greater than another, the latter is inferior and can be ignored	Maximin, expected value, regret	NR
Linearity	Multiplying by a positive constant and/or adding a constant to all entries in the payoff matrix should not change the best choice	Maximin, expected value, regret	NR
Duplication of columns	Repeating columns should not change the best choice	Maximin, regret	Expected value
Addition of row	Repeating row or rows should not change the best choice	Maximin, expected value	Regret
Invariance to a change in utility	A constant amount of utility or disutility does not change the choice	Expected value, regret	Maximin
Comparability	Adding a weakly dominated row should not change the best choice of act	Expected value, regret	Maximin

^(a) Adapted from Straffin [3]^(b) Not reported**Table 3** Relationship between decision makers, actions and payoffs (large numbers are preferred to small numbers by both)

		Decision maker 1	
		S_1	S_2
Decision maker 2	S_1	4, 4	0, 6
	S_2	6, 0	3, 3

for two DMs are measured by numbers such that a number of large magnitude is preferred to one of smaller magnitude (for example, 3 is preferred to 2). If the numbers were to represent disbenefits, then the opposite interpretation is true. The first number in a cell is that for DM1 and the second is that for DM2:

What would these two DMs do in this situation? They would most likely select their dominant strategies. The dominant strategy for DM1 is $S_2 = 6$, 3 and the dominant strategy for DM2 is $S_2 = 6$, 3. DM1 and DM2 are aware of the consequences and, because they do not know what the other will do, will settle for the strategy that dominates as seen from the vantage point of the individual DM. Thus the solution will be 3, 3: neither of the two will be induced to deviate from his/her choice of dominant strategy. Clearly,

however, the optimal solution is 4, 4; if the two DMs could communicate, they would select S_1 as their solution to the game.

The solution identified by the couple 3, 3 is known as the *Nash equilibrium* for this game: it is the strategy that does not provide any incentive to move elsewhere from that solution. For the example, it is a unique Nash equilibrium. Some games have no Nash equilibrium while others can have multiple Nash equilibriums. Consider the game in Table 4:

This situation does not have a Nash equilibrium because the two DMs can switch strategy as there is no incentive to maintain a strategy regardless of what the other does.

Mixed Strategy Games

The distinguishing feature of mixed strategies is that each is characterized by probabilities. Continuing with the 2×2 representation, if strategy for DM1 has two outcomes, then one outcome has probability pr and the other has probability $(1 - pr)$. For the other DM, however, we label the probabilities as q and $(1 - q)$. Consider the earlier game as in Table 5.

We can calculate the expected payoffs as follows:

Table 4 Example of game without a Nash equilibrium

		Decision maker 1	
		S_1	S_2
Decision maker 2	S_1	Profits, does not profit	Does not profit, profits
	S_2	Does not profit, profits	Profits, does not profit

Table 5 Example for mixed strategy (also see Table 3) using arbitrary probabilities and leading to a unique Nash equilibrium

		Decision maker 1	
		S_1	S_2
Decision maker 2	S_1	4, 4	0, 6
	S_2	6, 0	3, 3

For DM2 and S_1 : $(pr)(4) + (1 - pr)(0) = (pr)(4)$.
 For DM2 and S_2 : $(pr)(6) + (1 - pr)(3) = (pr)(3) + 3$.

For DM1 and S_1 : $(q)(4) + (1 - q)(0) = (pr)(4)$.
 For DM2 and S_2 : $(q)(6) + (1 - q)(3) = (q)(3) + 3$.

Putting arbitrary probabilities (0.5) in these expressions, we obtain:

For DM2 and S_1 : $(0.5)(4) + (1 - 0.5)(0) = (pr)(4)$. For DM2 and S_2 : $(0.5)(6) + (1 - 0.5)(3) = (0.5)(3) + 3 = 4.5$.

For DM1 and S_1 : $(0.5)(4) + (1 - 0.5)(0) = (0.5)(4)$. For DM2 and S_2 : $(0.5)(6) + (1 - 0.5)(3) = (0.5)(3) + 3 = 4.5$.

It follows that the game characterized by the unique Nash equilibrium remains unchanged. This is not the case for the other example that did not have a Nash equilibrium (Table 6).

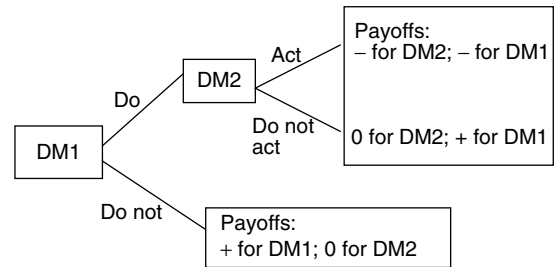
The mixed strategies approach changes the deterministic results. Putting arbitrary probabilities (0.5) in these expressions, we obtain

For DM2 and S_1 : $(0.5)(1) + (1 - 0.5)(0) = 0.5$.
 For DM2 and S_2 : $(0.5)(1) + (1 - 0.5)(0) = 0.5$.
 For DM1 and S_1 : $(0.5)(0) + (1 - 0.5)(1) = 0.5$.
 For DM2 and S_2 : $(0.5)(0) + (1 - 0.5)(1) = (0.5)(3) + 3 = 0.5$.

The result is Nash equilibrium for these probabilities: both DMs are indifferent to the strategy of the other.

The concepts of sequential games [5], where the choices made by the DMs account for the evolution of the choices, can be illustrated (Figure 2) as a decision tree in which the lines are the *branches* of the tree:

Each branch identifies a possible strategy that can be taken by a DM. This tree identifies simple binary choices and payoffs, which can be positive (+), negative (−), or zero (0). Reading the tree from left to right leads to an understanding of how the two DMs can behave and their possible final choices. DM1 would probably *do*, while DM2 might try to change the game, if it were possible.

**Figure 2** Simple example of sequential decision making involving two decision makers and two binary choices**Table 6** Example for mixed strategies showing indifference to each other's choice

		Decision maker 1	
		S_1	S_2
Decision maker 2	S_1	Profits = 1, does not profit = 0	Does not profit = 0, profits = 1
	S_2	Does not profit = 0, profits = 1	Profits = 1, does not profit = 0

Discussion

The relevance of studying situations of conflict through these methods is suggested by the following considerations:

- The structure of a game is unambiguously described.
- The payoffs, either inutilities or other suitable units, are quantitative expressions of each stakeholder's attitude.
- The methods for solving these games are theoretically sound and replicable.
- The criteria for choosing an alternative strategy over another consider both stakeholders.
- The resolution of some of the ambiguities inherent to such games may exist and be unique to that game.
- The difference between individual and societal choices can be shown in a coherent framework.

In the end, it is preferable to have more information than less (not necessarily, e.g., mutual insurance can be destroyed by free information, reducing everyone's expected utility); thus, anticipating the outcomes of a dispute provides useful information by orienting and informing a choice.

How does this theoretical framework compare with actual behavior? The basis for game theory is that DMs optimize their choices and seek more utility, rather than less. Although this basis is theoretically sound, there have been many cases of violations in simple, practical situations. Some, for example, Kahneman and Tversky (1979) found that actual behavior affects decision making to the point that the theory, while sound, required modifications, thus resulting in prospect theory. Camerer and Fehr [6] have also found that DMs inherently act altruistically and thus are not money maximizers.

Glimcker [7] reports the following. He uses the game *work* or *shirk* in a controlled setting: the employer faces the daily decision of going to work; the employer has the choice to inspect the employee's performance. For example consider Glimcker [7, p. 303, Table 12.2] (with change in terminology). The two stakeholders incur daily monetary costs and benefits in unstated monetary units, as shown in Table 7.

In general, because inspection cost is high, the employer will be allowed some shirking. If the inspection cost was lower, they could result in

Table 7 Wage–leisure trade-offs with or without cost of controlling the employees' attendance

	Employer checks attendance	Employer does not check
Employee works	Wage – leisure, revenue – wage – inspection cost (100 – 50 = 50, 125 – 100 – 50 = 25)	Wage – leisure, revenue – wage (50, 25)
Employee shirks	0, –inspection cost (0, –50)	Wage, –wage (100, –100)

more inspection and thus lead to decreasing rates of shirking: the result is Nash's equilibrium. For the employer, the hazard rates are (*shirk*) = ($\$inspec$ tion/ $\$wage$); (*inspect*) = ($\$leisure/\$wage$). The Nash equilibrium is (the result is obtained by solving for the probabilities, e.g., $50 \times x + (1 - x) \times 50 = 100 - 100 \times x$; thus $x = 0.5$) shirk 50% of the time and inspect 50% of the time [7, p. 304]. Students (unfamiliar with game theory) were enrolled in the shirk–inspect game to assess if their behavior was consistent with this theoretical equilibrium. In the first 100 (out of approximately 140) tests, the stakeholders modified their outcomes and *tended* toward the Nash equilibrium. When the payoffs were changed, the results changed as well.

The *shirk-work* game discussed in Glimcker [7] provides some interesting insights on the sequential choices made between either working or shirking, when deciding which alternative to take by the worker who is initially contemplating the choice of either working or shirking work. The results show that the worker (in an experiment involving 150 trials) appears to be acting randomly and thus keeps the employer guessing as to what course of action she will actually take. This author reports (adapted from Figure 12.2, p. 308) that the choices, made in 150 games, has the results as in Figure 3.

The decision tree shows the chronology of the decision. The initial choices are to either work or shirk. The second decision, conditioned (the conditionalization is symbolized by the symbol “[|]”), is to *work* | *work*, *shirk* | *work*, and so on, as depicted by the tree. It appears that the worker is acting randomly and thus keeps the employer guessing!

Clearly, individual choices cannot be divorced from the fact that those choices are made in the

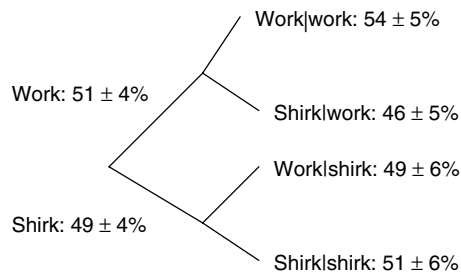


Figure 3 Decision tree depicting the alternatives in the wage-shirking work game

brain. Interestingly, it is now increasingly apparent that humans have [8] ... *at least two systems working when making moral judgments* ... *There is an emotional system that depends on (the ventromedial prefrontal cortex) ... and another that performs more utilitarian cost-benefit analyses which in these people is intact.* When that region is damaged, individuals opt for choices that are different from those not damaged. For example, in a relatively small sample of respondents, only approximately 20% of *normal* individuals answered *yes*, while approximately 80% of individuals with damaged brain answered *yes*, to the following question:

You have abandoned a sinking cruise ship and are in a crowded lifeboat that is dangerously low in the water. If nothing is done it will sink before the rescue boats arrive and everyone will die. However, there is an injured person who will not survive in any case. If you throw the person overboard, the boat will stay afloat and the remaining passengers will be saved. Would you throw this person overboard in order to save the lives of the remaining passengers?

Now consider a different choice (Carey [8]): having to divert a train by *flipping a switch*, to save five workers, knowing that such diversion would kill one other worker for sure. The research shows that those with the brain injury, normal individual, and those with a different type of brain injury would divert the train. When that certainty of killing the

single worker was not apparent, all of the three groups rejected the trade-off. Moreover, if the action did not comport flipping a switch but, rather, the act was equivalent to pushing that single individual to certain death to save several others at risk, the results by all of the three groups were still different: those with the ventromedial injury were *about twice as likely as other participants to say that they would push someone in front of the train (if that were the only option ...)*. Although the brain injury is uncommon, the responses may make some think twice about the implications of either answer.

References

- [1] Von Neumann, J. & Morgenstern, O. (1967). *The Theory of Games and Economic Behavior*, John Wiley & Sons, New York.
- [2] Luce, R.D. & Raiffa, H. (1957). *Games and Decisions*, John Wiley & Sons, New York.
- [3] Straffin, P.D. (1993). *Game Theory and Strategy*, The Math Association of America, New York.
- [4] Williams, J.D. (1986). *The Complete Strategist*, Dover, Mineola.
- [5] Burkett, J.P. (2006). *Microeconomics, Optimization, Experiments, and Behavior*, Oxford University Press, Oxford.
- [6] Camerer, C.F. & Fehr, E. (2006). When does *economic man* dominate social behavior? *Science* **311**, 47.
- [7] Glimcker, P.W. (2003). *Decision, Uncertainty, and Brain: The Science of Neuroeconomics*, MIT Press, Cambridge.
- [8] Carey, B. (2007). Brain injury is linked to moral decisions, *The New York Times*, Thursday, March 22, A19.

Further Reading

- Luce, R.D. (1992). Where does subjective utility theory fail prescriptively? *Journal of Risk and Uncertainty* **5**, 5–27.
- Luce, R.D. & Narens, L. (1987). Measurement scales on the continuum, *Science* **236**, 1527.
- Render, B., Stair, R.M. & Balakrishnan, N. (2003). *Managerial Decision Modeling with Spreadsheets*, Prentice-Hall, Upper Saddle River.

PAOLO F. RICCI

Confounding

In the presence of confounding, the old adage “association is not causation” holds even if the study population is arbitrarily large. For example, consider a very large cohort study (*see Cohort Studies*) whose goal is estimating the causal effect of receiving the treatment (the exposure) A on the risk of death (the outcome) Y . Investigators compute the risk ratio $\Pr[Y = 1|A = 1] / \Pr[Y = 1|A = 0]$, where A is 1 if the subject received the exposure and 0 otherwise, as a measure of the association between A and Y . However, what investigators would really want to compute is the causal risk ratio $\Pr[Y_{a=1} = 1] / \Pr[Y_{a=0} = 1]$, where, for each subject, Y_a is the (possibly counterfactual) outcome that would have been observed had the subject received exposure level a . We say that there is confounding if, for example, subjects with the worst prognosis received treatment $A = 1$. When there is confounding, the value of the associational risk ratio may be very different from that of the causal risk ratio, even in the absence of other sources of bias. For simplicity of presentation, let us assume throughout this chapter that other sources of bias (e.g., selection bias, measurement error, and sampling variability) are absent. This chapter is organized as follows.

First, we use causal graphs (*see Causal Diagrams*) to describe the causal structure that gives rise to confounding. Causal directed acyclic graphs (DAGs) [1–3] have been increasingly used in epidemiology [4–6] and other disciplines. Nodes in DAGs represent variables, both measured and unmeasured, that are linked by arrows. A causal DAG is one in which the arrows can be interpreted as direct causal effects, and all common causes of any pair of variables are included in the graph. Causal DAGs are acyclic because a variable cannot cause itself, either directly or through other variables. A path in a DAG is any arrow-based route between two variables in the graph. A path is causal if it follows the direction of the arrows, and noncausal otherwise. As an example, the causal DAG in Figure 1 represents the dichotomous variables A (ever smoker) and Y (lung cancer)

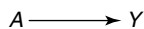


Figure 1 A simple causal DAG

linked by a single causal path. The lack of other variables in the DAG indicates that no common causes of A and Y are believed to exist. A major strength of using DAGs is that a set of simple graphical rules can be applied to determine whether two variables are unassociated under the causal assumptions encoded in the graph. These rules, known as *d-separation rules*, are summarized in the Appendix.

Second, we review a condition under which confounding can be eliminated. This condition will be referred to as a condition for the identifiability of causal effects. Third, we examine the definition of confounder that logically follows from the identifiability condition. Fourth, we describe an alternative, counterfactual-based definition of confounding that is mathematically equivalent to the structural (graphical) definition of confounding. Finally, we provide a classification of the methods to eliminate the measured confounding using data on the available confounders. Let us now turn our attention to the causal structure that underlies confounding bias.

The Structure of Confounding

Confounding is the bias that arises when the exposure and the outcome share a common cause. The structure of confounding can be represented by using causal DAGs. For example, the causal DAG in Figure 2(a) depicts an exposure A , an outcome Y , and their common cause L . This DAG shows two sources of association between exposure and outcome: (i) the path $A \rightarrow Y$ that represents the causal effect of A on Y and (ii) the path $A \leftarrow L \rightarrow Y$ between A and Y that is mediated by the common cause L . In graph theory, a path like $A \leftarrow L \rightarrow Y$ that links A and Y through their common cause L is referred to as a *backdoor path*. More generally, in a causal DAG, a backdoor path is a noncausal path between exposure and outcome that does not include any variable causally affected by (or, in graph-theoretic terms, any descendant of) the exposure.

If the common cause L did not exist in Figure 2(a), then the only path between exposure and outcome would be $A \rightarrow Y$, and thus the entire association between A and Y would be due to the causal effect of A on Y . That is, association would be causation; the associational risk ratio $\Pr[Y = 1|A = 1] / \Pr[Y = 1|A = 0]$ would equal the causal risk ratio $\Pr[Y_{a=1} = 1] / \Pr[Y_{a=0} = 1]$.

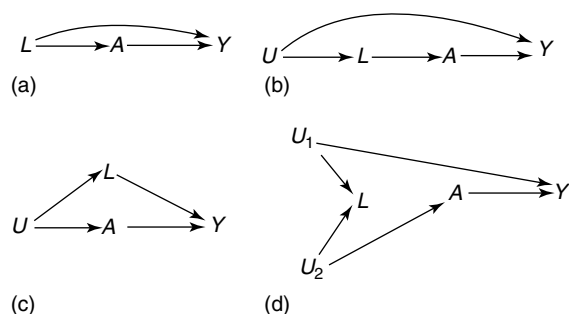


Figure 2 Examples of traditional “confounders”

But the presence of the common cause L creates an additional source of association between the exposure A and the outcome Y , which we refer to as confounding for the effect of A on Y . Because of confounding, association is not causation; the associational risk ratio does not equal the causal risk ratio.

Examples of confounding abound in observational research. Consider the following examples of confounding for various kinds of exposures in relation to health outcomes:

Occupational factors

- The effect of working as a firefighter A on the risk of death Y will be confounded if “being physically fit” L is a cause of both being an active firefighter and having a lower mortality risk. This bias, depicted in the causal DAG in Figure 2(a), is often referred to as a *healthy worker bias*.

Drug use and other clinical decisions

- The effect of use of drug A (say, aspirin) on the risk of disease Y (say, stroke) will be confounded if the drug is more likely to be prescribed to individuals with certain condition L (say, heart disease) that is both an indication for treatment and a risk factor for the disease. Heart disease L may be a risk factor for stroke Y because L has a direct causal effect on Y as in Figure 2(a) or, as represented in Figure 2(b), because both A and Y are caused by atherosclerosis U , an unmeasured variable. This bias is known as *confounding by indication* or *channeling*, the last term often being reserved to describe the bias created by patient-specific risk factors L that encourage doctors to use certain drug A within a class of drugs.

Lifestyle

- The effect of behavior A (say, heavy alcohol intake) on the risk of Y (say, death) will be confounded if the behavior is associated with another behavior L (say, cigarette smoking) that has a causal effect on Y and tends to co-occur with A . The structure of the variables L , A , and Y is depicted in the causal DAG in Figure 2(c), in which the unmeasured variable U represents the sort of personality that leads to both heavy drinking and smoking.

Genetic factors

- The effect of a DNA sequence A on the risk of developing certain trait Y will be confounded if there exists a DNA sequence L that has a causal effect on Y and is more common among people carrying A . This bias, also represented by the causal DAG in Figure 2(c), is known as *linkage disequilibrium* or *population stratification*, the last term often being reserved to describe the bias arising from conducting studies in a mixture of individuals from different ethnic groups. Thus the variable U can stand for ethnicity or other factors that result in linkage of DNA sequences.

Social factors

- The effect of income at age 65, A , on the level of disability at age 75 Y will be confounded if the level of disability at age 55 L affects both future income and disability level. This bias may be depicted by the causal DAG in Figure 2(a).

Environmental exposures

- The effect of airborne particulate matter A on the risk of coronary heart disease Y will be confounded if other pollutants L whose levels covary with those of A cause coronary heart disease. This bias is also represented by the causal DAG in Figure 2(c), in which the unmeasured variable U represent weather conditions that affect the levels of all types of air pollution.

In all these cases, the bias has the same structure: it is due to the presence of a common cause (L or U) of the exposure A and the outcome Y , or, equivalently, to the presence of an unblocked backdoor path between A and Y . We refer to the bias caused by common causes as confounding. We use other names

to refer to biases caused by structural reasons other than the presence of common causes. For example, we say that selection bias is the result of conditioning on common effects [7].

Confounding and Identifiability of Causal Effects

Once confounding is structurally defined as the bias resulting from the presence of common causes of exposure and outcome, the key question is, under what conditions can confounding be eliminated? In other words, in the absence of measurement error and selection bias, under what conditions can the causal effect of exposure A on outcome Y be (nonparametrically) identified? An important result from graph theory, known as *the backdoor criterion* [1], is that the causal effect of exposure A on the outcome Y is identifiable if all backdoor paths between them can be blocked. Thus the two settings in which causal effects are identifiable are as follows:

1. *No common causes*
If, as in the causal DAG of Figure 1, there are no common causes of exposure and outcome, and hence no backdoor paths that need to be blocked, we say that there is no confounding.
2. *Common causes but enough measured variables to block all backdoor paths*
If, as in the causal DAG of Figure 2(a), the backdoor path through the common cause L can be blocked by conditioning on the measured covariates (in this example, L itself), we say that there is confounding but no unmeasured confounding.

The first setting is expected in simple randomized experiments in which all subjects have the same probability of receiving exposure. In these experiments, confounding is not expected to occur because exposure is determined by the flip of a coin – or its computerized upgrade, the random number generator – and the flip of the coin cannot be a cause of the outcome.

The second setting is expected in randomized experiments in which the probability of receiving exposure is the same for all subjects with the same value of risk factor L but, by design, this probability varies across values of L . The design of these experiments guarantees the presence of confounding,

because L is a common cause of exposure and outcome, but in these experiments confounding is not expected conditional on – within levels of – the covariates L . This second setting is also what one hopes for in observational studies in which many variables L have been measured.

The backdoor criterion answers three questions: (i) does confounding exist? (ii) can confounding be eliminated? and (iii) what variables are necessary to eliminate the confounding? The answer to the first question is affirmative if there exist unblocked backdoor paths between exposure and outcome; the answer to the second question is affirmative if all those backdoor paths can be blocked by conditioning on the measured variables; the answer to the third question is the minimal set of variables that, when conditioned on, block all backdoor paths. The backdoor criterion, however, does not answer questions regarding the magnitude of the confounding. It is logically possible that some unblocked backdoor paths are weak (e.g., if L does not have a large effect on either A or Y) and thus induce little bias, or that several strong backdoor paths induce bias in opposite directions and thus result in a weak net bias.

Note that, even though the application of the backdoor criterion requires conditioning on the variables L , the backdoor criterion does not imply that conditioning or restriction is the only possible method to eliminate confounding. Conditioning is only a graphical device to determine whether a certain variable in L can be used to block the backdoor path. The section titled “Methods to Adjust for Confounding” provides an overview of the methods that eliminate confounding using data on the confounders. First let us review the definition of confounder.

Confounding and Confounders

Confounding is the bias that results from the presence of common causes of exposure and outcome. A *confounder* is any variable that can be used to reduce such bias, that is, any variable that can be used to block a backdoor path between exposure and outcome. In contrast with this structural definition, confounder was traditionally defined as any variable that meets the following three conditions: (i) it is associated with the exposure, (ii) it is associated with the outcome conditional on exposure (with “conditional on exposure” often replaced by “in the

unexposed”), and (iii) it does not lie on a causal pathway between exposure and outcome. According to this traditional definition, all confounders so defined should be adjusted for in the analysis. However, this traditional definition of confounder may lead to inappropriate adjustment for confounding [2, 4]. To see why, let us compare the structural and traditional definitions of confounder in Figure 2(a–d).

In Figure 2(a) there is confounding because the exposure A and the outcome Y share the common cause L , i.e., because there is a backdoor path between A and Y through L . However, this backdoor path can be blocked by conditioning on L . Thus, if the investigators collected data on L for all individuals, there is no unmeasured confounding given L . We say that L is a confounder because it is needed to eliminate confounding. Let us now turn to the traditional definition of confounder. The variable L is associated with the exposure (because it has a causal effect on A), it is associated with the outcome conditional on the exposure (because it has a direct effect on Y), and it does not lie on the causal pathway between exposure and outcome. Then, according to the traditional definition, L is a confounder and it should be adjusted for. There is no discrepancy between the structural and traditional definitions of confounder for the causal DAG in Figure 2(a).

In Figure 2(b) there is confounding because the exposure A and the outcome Y share the common cause U , i.e., there is a backdoor path between A and Y through U . (Unlike the variables L , A , and Y , we suppose that the variable U was not measured by the investigators.) This backdoor path could be theoretically blocked, and thus confounding eliminated, by conditioning on U , had data on this variable been collected. However, this backdoor path can also be blocked by conditioning on L . Thus, there is no unmeasured confounding given L . We say that L is a confounder because it is needed to eliminate confounding, even though the confounding resulted from the presence of U . Let us now turn to the traditional definition of confounder. The variable L is associated with the exposure (because it has a causal effect on A), it is associated with the outcome conditional on the exposure (because it shares the common cause U with Y), and it does not lie on the causal pathway between exposure and outcome. Then, according to the traditional definition, L is a confounder and it should be adjusted for. Again, there is no discrepancy between the structural and

traditional definitions of confounder for the causal DAG in Figure 2(b).

In Figure 2(c) also there is confounding because the exposure A and the outcome Y share the common cause U , and the backdoor path can also be blocked by conditioning on L . Therefore, there is no unmeasured confounding given L , and we say that L is a confounder. Because L is associated with the exposure (it shares the common cause U with A), it is associated with the outcome conditional on the exposure (it has a causal effect on Y), and it does not lie on the causal pathway between exposure and outcome, then, according to the traditional definition, L is a confounder and it should be adjusted for. Again, there is no discrepancy between the structural and traditional definitions of confounder for the causal DAG in Figure 2(c).

The key figure is Figure 2(d). In this causal DAG there are no common causes of exposure A and outcome Y , and therefore no confounding. In graph-theoretical terms, the backdoor between A and Y through L ($A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$) is blocked because L is a collider on that path (see Appendix). Association is causation. There is no need to adjust for L in any way. In fact, conventional adjustment for L would introduce (selection) bias because conditioning on L would open the otherwise blocked backdoor path between A and Y . However, L is associated with the exposure (it shares the common cause U_2 with A), it is associated with the outcome conditional on the exposure (it shares the common cause U_1 with Y), and it does not lie on the causal pathway between exposure and outcome. Hence, according to the traditional definition, L is a confounder and it should be adjusted for. In this example, the traditional definition labels L as a confounder even in the absence of confounding bias. The result of trying to adjust for this nonexistent confounding is selection bias. This particular example of selection bias has been referred to as *M-bias* [8] because the structure of the variables involved in it – U_2, L, U_1 – resembles a letter M (lying on its side in Figure 2d). Some authors (e.g., Greenland, Poole, personal communication) refer to this type of M-bias, or to any bias that results from selection on preexposure factors, as confounding rather than as selection bias. This choice of terminology slightly extends the meaning of the term “confounding” to some situations in which no common causes exist, but this extension has no practical consequences:

adjusting for L in Figure 2(d) creates bias, regardless of its name.

We have described an example in which the standard definition of confounder fails because it misleads investigators into adjusting for a variable when adjustment for such a variable is not only superfluous but also harmful. This problem arises because the standard definition treats the concept of confounder, rather than that of confounding, as the primary concept. In contrast, the structural definition first characterizes confounding as the bias resulting from the presence of common causes and then characterizes confounder as any variable that is necessary to eliminate the bias in the analysis. For all practical purposes, confounding is an absolute concept – common causes of exposure and the outcome either exist or do not exist in our region of the universe – whereas confounder is a relative concept – a variable L may be needed to block a backdoor path only when another variable U is not measured. A related problem is that the standard definition defines confounder in terms of statistical associations (a variable associated with the exposure, etc.) that may exist even when confounding (due to common causes) does not exist as shown in Figure 2(d). Confounding is a causal concept that cannot be reduced to statistical terms.

Consider now the causal DAG in Figure 3. In this graph there is confounding for the effect of A on Y because of the presence of the unmeasured common cause U . The measured variable L is a proxy or surrogate for the common cause U . For example, the unmeasured variable socioeconomic status U may confound the effect of physical activity A on the risk of cardiovascular disease Y . Income L is a surrogate for the often ill-defined variable socioeconomic status. Should we consider the variable L as a confounder? On the one hand, L cannot be used to block the backdoor path between A and Y because it does not lie on that path. On the other hand, adjusting for L would indirectly adjust for some of the confounding caused by U because L is correlated with U . In the extreme, if L were perfectly correlated with U ,

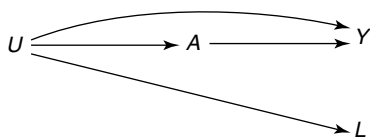


Figure 3 A surrogate confounder

then it would make no difference whether we one conditions on L or on U . That is, conditioning on L results in a partial blockage of the backdoor path $A \leftarrow U \rightarrow Y$, and we will usually prefer to adjust, rather than not to adjust, for L . We refer to variables like L as *surrogate confounders*.

For pedagogic purposes, all causal DAGs represented in this chapter include non-time-varying exposures only. However, the above definitions of confounding and confounders can be generalized to the case of time-varying exposures. Pearl and Robins [9] described the generalized backdoor criterion for the identifiability of the effects of time-varying and generalized exposures. A time-varying confounder is any time-varying variable that is needed to identify the effect of a time-varying exposure. Settings with time-varying confounders and exposures make it even clearer why the traditional definition of confounding, and the conventional methods for confounding adjustment, may result in selection bias. See [7] for an informal introduction to the problems associated with time-varying exposures and confounders, and [10] for a more formal treatment.

Confounding and Exchangeability

So far, we have defined confounding from a structural standpoint: the bias induced by common causes of exposure and outcome. However, it is also possible to provide a precise definition of confounding based on the implications of its structure, with no explicit reference to common causes.

Take our second example of confounding above: the individuals exposed to aspirin are different from the unexposed because, if the exposed had remained unexposed, their risk of stroke would have been higher than that of those that did actually remain unexposed – because the exposed had a higher frequency of heart disease, a common cause. Thus the consequence of the existence of common causes of exposure and outcome is that the exposed and the unexposed are not “comparable” or, more technically, that they are not exchangeable [11]. Exchangeability can be represented mathematically by using counterfactuals (see, for example [12]). For a dichotomous exposure A and outcome Y , exchangeability means that the equality $\Pr[Y_a = 1] = \Pr[Y = 1|A = a]$ holds for all a .

Confounding can then be defined as the bias that results from the presence of common causes of exposure and outcome (a structural definition) or as the bias that results from lack of exchangeability of the exposed and the unexposed (a counterfactual-based definition). In the absence of bias caused by selection or mismeasurement, both definitions are mathematically equivalent. When the exposure is randomly assigned, the exposed and the unexposed are expected to be *exchangeable* because no common causes exist (no confounding). When the exposure is not randomly assigned, investigators measure many variables in an attempt to ensure that the exposed and the unexposed are *conditionally exchangeable* given the measured covariates L because, even though common causes may exist (confounding), investigators assume that the variables L are sufficient to block all backdoor paths (no unmeasured confounding).

However, an exchangeability-based definition of confounding that does not explicitly specify the common cause(s) responsible for the bias has limited practical utility in observational research. This is so because methods to adjust for confounding in observational studies require adjustment for variables (i.e., L in Figure 2a–c) other than exposure A and outcome Y , and therefore detailed knowledge about the structure of the bias is needed to decide on which variables to measure. In the next section, we review the methods that can be used to eliminate confounding when enough confounders L are available to block all backdoor paths between exposure and outcome.

Methods to Adjust for Confounding

Randomization is the preferred method to control for confounding because a random assignment of exposure is expected to produce exchangeability of the exposed and the unexposed (*see Randomized Controlled Trials*). When the exposure is randomly assigned in large samples, investigators need not wonder about common causes of exposure and outcome because randomization precludes their existence. As a consequence, subject-matter knowledge to identify appropriate adjustment variables is unnecessary. In contrast, all other methods to eliminate confounding require expert knowledge – because the judgment as to which variables on a DAG cause which others must in general be based on subject

matter considerations – and succeed only under the uncheckable assumption that this knowledge is sufficient to identify and block all backdoor paths between exposure and outcome.

The goal of all methods to eliminate measured confounding is estimation of the effect of an exposure A on an outcome Y in a certain population. These methods can be classified into the two following categories:

1. Methods that estimate the causal effect in one or more subsets of the population in which no measured confounding exists.

These methods – including restriction, stratification, and matching – compare the distribution of the outcome between exposed and unexposed in such subset(s) of the population.

(a) Restriction

Limiting the analysis to one stratum of the covariates in L , a method known as *restriction*, estimates the effect of the exposure on the outcome in the selected stratum, which may differ from the effect in the population if effect measure modification exists. The investigators may restrict their analysis of aspirin A and stroke Y to the subset of subjects without a history of heart disease ($L = 0$). By conditioning on $L = 0$, an action that can be graphically represented as a box around variable L in Figure 2(a), the backdoor path from A to Y is blocked and thus the association, if any, between exposure A and outcome Y can be entirely attributed to the causal effect of A on Y (under the assumption of no unmeasured confounding given L). In other words, the exposed and the unexposed are exchangeable in the stratum $L = 0$ or, equivalently, the distribution of the outcome Y in the exposed (unexposed) in the stratum $L = 0$ equals the distribution of the counterfactual outcome if all subjects in the stratum $L = 0$ had been exposed (unexposed).

(b) Stratification

In the above example, a stratified analysis is simply an analysis conducted separately in each of the two strata of history of heart disease L (0: no, 1: yes). Thus stratification is simply the application of restriction to several comprehensive and mutually exclusive

subsets of the population. The conditioning on each value of L , which can be represented as a box around variable L in Figure 2(a), results in two association measures between A and Y , one per stratum of L , that can be interpreted as the causal effect of A on Y in each stratum (under the assumption of no unmeasured confounding given L). Often, when effect measure modification is not believed to exist, the stratum-specific effect measures are pooled to increase the statistical efficiency of the estimation. Conventional parametric or semiparametric regression models are a sophisticated version of stratification in which conditional effect measures are estimated within levels of all the covariates in the model. Not including interaction terms between exposure and covariates in the regression model is equivalent to assuming that all conditional effect measures (i.e., stratum-specific for discrete variables) are equal.

(c) Matching

Another method for the elimination of measured confounding is that of selecting the exposed and the unexposed so that both groups have the same distribution of the variables in L . For example, in our aspirin and stroke example, one could find one (or more) unexposed subject with $L = 0$ for each exposed subject with $L = 0$. In the resulting matched population, a subset of the entire population, a history of heart disease L does not predict whether the patients are exposed to aspirin A and thus there is no measured confounding by L . This method estimates the effect of A on Y in a subset of the population with the same distribution of L as the exposed so that, under the assumption of no unmeasured confounding given L , this method estimates the effect in the exposed.

A common variation of restriction, stratification, and matching replaces the variable L by the probability of receiving exposure $\Pr[A = 1|L]$, usually referred to as the *propensity score*. When L is high dimensional (say, a vector of 20 covariates), propensity score-based methods can be seen as a dimension reduction strategy that allows investigators to work with one scalar variable, the propensity score, which

is in itself sufficient to block the same backdoor paths that can be blocked by the multiple variables in L [13].

2. Methods that estimate the causal effect in the entire population or in any subset of it.

These methods – including standardization, inverse probability weighting, and g-estimation – simulate the distribution of the (counterfactual) outcomes in the population if no measured confounding had existed. A detailed description of these methods is beyond the scope of this chapter, but they all use the assumption that there is no unmeasured confounding within strata of the variables in L (i.e., that all backdoor paths can be blocked) to simulate the association between A and Y in the population if there had actually been no measured confounding (no backdoor paths mediated through L). Hence, under the assumption of no unmeasured confounding given L , the simulated association between A and Y can be entirely attributed to the effect of A on Y . For example, in the analysis of aspirin A and stroke Y , these methods would simulate the distribution of the outcomes if the probability of receiving aspirin had been the same regardless of a subject's heart disease history L . This can be represented in Figure 2(a) by deleting the arrow from L to A . In other words, the exposed and the unexposed are exchangeable in the simulated population or, equivalently, the distribution of the outcome Y in the exposed (unexposed) equals the distribution of the counterfactual outcome if all subjects in the population had been exposed (unexposed). These methods can also be applied to a subset of the population, if so desired. The parametric and semiparametric extensions of these methods are the parametric g-formula for standardization [14], marginal structural models for inverse probability weighting [15], and nested structural models for g-estimation [16]. Unlike the methods based on restriction, stratification, or matching, these methods can appropriately adjust for measured time-dependent confounding [10].

All the methods listed above require the assumption of no unmeasured confounding given L for the effect of exposure A on outcome Y , that is, the assumption that investigators have measured enough variables L to block all backdoor paths from A to Y . (Technically, g-estimation requires the slightly

weaker assumption that the amount of unmeasured confounding given L is known, of which the assumption of no unmeasured confounding is a particular case.) Other methods not listed above – instrumental variable methods – use a different set of assumptions to estimate the causal effect of A on Y in observational studies. Among other things, these methods require the measurement of a variable I , the instrument, such that there is no unmeasured confounding for the effect of I on Y and that the effect of I on Y is entirely mediated through A . For a review of the assumptions underlying instrumental variable methods, see [17]. The assumptions required by instrumental variable methods may be stronger than the assumption of no unmeasured confounding for the effect of A on Y . However, because all these assumptions are not empirically verifiable, it will be often unclear which set of assumptions is more likely to hold in a particular setting.

Note that the existence of common causes of exposure and outcome, and thus the definition of confounding, does not depend on the method used to adjust for it or on the results obtained from these methods. Hence we do not say that measured confounding exists simply because the adjusted estimate is different from the unadjusted estimate. In fact, one would expect that the presence of measured confounding lead to a change in the estimate, but not necessarily the other way around. Changes in estimates may occur for reasons other than confounding, including the introduction of selection bias when adjusting for nonconfounders [18] and the use of noncollapsible effect measures like the odds ratio (see **Odds and Odds Ratio**) in a closed cohort study (see **Cohort Studies**) [19]. Attempts to define confounding based on change in estimates have been long abandoned because of these problems.

Conclusion

The structural definition of confounding emphasizes that causal inference from observational data requires *a priori* causal assumptions or beliefs, which must be derived from subject-matter knowledge, and not only from statistical associations detected in the data. As shown in Figure 2(d), statistical criteria are insufficient to characterize confounding and can lead to inconsistencies between beliefs and actions in data analysis.

Knowledge of the causal structure is a prerequisite to determine the existence of confounding and to accurately label a variable as a confounder. In practice, however, this requirement may impose such an unrealistically high standard on the investigator that many studies simply cannot be done at all. A more realistic recommendation for researchers is to avoid adjusting for a variable L unless they believe it may possibly be a confounder, i.e., it may block a backdoor path. At the very least, investigators should generally avoid stratifying on variables affected by either the exposure or the outcome. Of course, thoughtful and knowledgeable investigators could believe that two or more causal structures, possibly leading to different conclusions regarding confounding and confounders, are equally plausible. In that case they would perform multiple analyses and explicitly state the assumptions about causal structure required for the validity of each. Unfortunately, one can never be certain that the set of causal structures under consideration includes the true one; this uncertainty is unavoidable with observational data.

Appendix: d-Separation

In a DAG, each path is blocked according to the following graphical rules [1]:

1. If there are no variables being conditioned on, a path is blocked if and only if two arrowheads on the path collide at some variable (known as a *collider*) on the path.
2. Any path that contains a noncollider that has been conditioned on is blocked. We use a square box around a variable to indicate that we are conditioning on it.
3. A collider that has been conditioned on does not block a path.
4. A collider that has a descendant that has been conditioned on does not block a path.

Rules 1–4 can be summarized as follows. A path is blocked if and only if it contains a noncollider that has been conditioned on, or it contains a collider that has not been conditioned on and has no descendants that have been conditioned on. Two variables are d-separated if all paths between them are blocked (otherwise they are d-connected). Some conclusions that follow from the method of d-separation are that causes (ancestors) are not independent of their effects

(descendants) and *vice versa*, and that generally two variables are associated if they share a common cause. Another important conclusion is that sharing a common effect does not imply that two causes are associated. Intuitively, whether two variables (the common causes) are correlated cannot be influenced by an event in the future (their effect), but two causes of a given effect generally become associated once we stratify on the common effect.

Finally, two variables that are not d-separated may actually be statistically independent. The reason is that it is logically possible that causal effects in opposite directions may exactly cancel out. Because exact cancellation of causal effects is probably a very rare event in natural circumstances, d-separation and independence may be treated in practice as equivalent concepts with little risk. In the probably rare occasions in which two variables are simultaneously d-connected and statistically independent, we say that the joint distribution of the variables in the DAG is not faithful to the DAG [3].

Acknowledgment

The author thanks Sander Greenland for his comments to an earlier version of the manuscript. This work was supported by NIH grant R01 HL080644.

References

- [1] Pearl, J. (1995). Causal diagrams for empirical research, *Biometrika* **82**, 669–710.
- [2] Pearl, J. (2000). *Causality: Models, Reasoning and Inference*, Cambridge University Press, Cambridge.
- [3] Spirtes, P., Glymour, C. & Scheines, R. (2001). *Causation, Prediction, and Search*, 2nd Edition, The MIT Press, Cambridge.
- [4] Greenland, S., Pearl, J. & Robins, J.M. (1999). Causal diagrams for epidemiologic research, *Epidemiology* **10**, 37–48.
- [5] Robins, J.M. (2001). Data, design, and background knowledge in etiologic inference, *Epidemiology* **12**, 313–320.
- [6] Glymour, M.M. & Greenland, S. (2008). Causal diagrams, in *Modern Epidemiology*, 3rd Edition, K.J. Rothman, S. Greenland & T.L. Lash, eds, Lippincott Williams & Wilkins, Philadelphia (in press).
- [7] Hernán, M.A., Hernández-Díaz, S. & Robins, J.M. (2004). A structural approach to selection bias, *Epidemiology* **15**, 615–625.
- [8] Greenland, S. (2003). Quantifying biases in causal models: classical confounding versus collider-stratification bias, *Epidemiology* **14**, 300–306.
- [9] Pearl, J. & Robins, J.M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables, in *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, Montreal, Quebec, Canada, pp. 444–453.
- [10] Robins, J.M. & Hernán, M.A. (2007). Estimation of the effects of time-varying exposures, in *Advances in Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke & G. Molenberghs, eds, Chapman & Hall/CRC Press, New York (in press).
- [11] Greenland, S. & Robins, J.M. (1986). Identifiability, exchangeability, and epidemiological confounding, *International Journal of Epidemiology* **15**, 413–419.
- [12] Hernán, M.A. (2004). A definition of causal effect for epidemiological research, *Journal of Epidemiology and Community Health* **58**, 265–271.
- [13] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.
- [14] Robins, J.M., Hernán, M.A. & Siebert, U. (2004). Effects of multiple interventions, in *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors*, M. Ezzati, A.D. Lopez, A. Rodgers & C.J.L. Murray, eds, World Health Organization, Geneva, Vol. II.
- [15] Robins, J.M. (1998). Marginal structural models, in *1997 Proceedings of the Section on Bayesian Statistical Science*, American Statistical Association, Alexandria, pp. 1–10.
- [16] Robins, J.M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies, in *Health Service Research Methodology: A Focus on AIDS*, L. Sechrest, H. Freeman & A. Mulley, eds, U.S. Public Health Service, National Center for Health Services Research, Washington, DC, pp. 113–159.
- [17] Hernán, M.A. & Robins, J.M. (2006). Instruments for causal inference: an epidemiologist’s dream? *Epidemiology* **17**, 360–372.
- [18] Hernán, M.A., Hernández-Díaz, S., Werler, M.M. & Mitchell, A.A. (2002). Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology, *American Journal of Epidemiology* **155**, 176–184.
- [19] Greenland, S., Robins, J.M. & Pearl, J. (1999). Confounding and collapsibility in causal inference, *Statistical Science* **14**, 29–46.

Related Articles

Causality/Causation

Compliance with Treatment Allocation

MIGUEL A. HERNÁN

Considerations in Planning for Successful Risk Communication

As has been previously discussed, risk communication is an integral and vital component of risk assessment and risk management (*see* **Role of Risk Communication in a Comprehensive Risk Management Approach**). However, successful risk communication does not “just happen” – it is the result of careful planning and consideration of multiple needs and objectives.

Much guidance has been provided on conducting risk communication. This has involved several different approaches over the last 20 years, including: (a) the “Message Transmission Model for Risk Communication” [1]; (b) the Communication Process Model [2]; (c) the “Improving Risk Communication” approach of the US National Research Council [3]; (d) the Seven Cardinal Rules of Risk Communication [4, 5]; (e) the “Hazard *versus* Outrage” approach [6]; (f) the “Mental Models” approach [7–9]; (g) the “Analytic–Deliberative Process” approach [10]; and (h) the considerations outlined by the US Presidential/Congressional Commission on Risk Assessment and Risk Management [11]. All of these approaches have provided insights into the risk communication process and advice on improving the effectiveness of risk dialogues.

However, it is important to recognize there is no “magic” set of rules that, if followed, will always ensure effective risk communication and accepted risk-management decisions. Nonetheless, there are certain considerations that can provide guidance in planning and executing risk communication efforts to enhance opportunities for productive dialogues.

Successful Risk Communication Considerations

There is no “cookbook” formula for risk communication

This is the first and foremost consideration in good risk communication. Risk communication efforts must be based on the specific circumstances surrounding each risk situation. A rigid or “one size fits

all” protocol for public participation and communication cannot possibly accommodate the diverse circumstances of all cases [12]. Attempts to reduce risk communication to a rote procedure will inevitably result in a less than optimal process and ineffective (if not outright damaging) results.

Planning is required for any effort to be effective

Notwithstanding the above, risk communication requires a certain degree of planning to be effective. A risk communication plan will help focus efforts, ensure that nothing is overlooked and keep everyone informed. Risk communication frameworks usually proceed in a linear, but iterative, fashion starting with problem formulation and proceeding through to communications strategy design and implementation. Evaluation plays a very important role in the framework, and should be conducted at the beginning, during, and at the end of the program (*see* **Evaluation of Risk Communication Efforts**). Any risk communication plan must be considered a “living” document that grows and evolves throughout a risk assessment program [3, 10, 11].

All interested and affected parties should be given the opportunity to be involved at the outset of, and throughout, the risk process

Stakeholder involvement is a critical component in risk communication. Involving the appropriate people from the beginning of the process, when the problem is being formulated, helps to ensure that the right problem is being addressed in the right way [10, 11].

Achieving and maintaining mutual trust and credibility are critical to communication success

As noted by Slovic [13, 14], trust is not automatic, nor everlasting – it is difficult to gain, even harder to maintain, and once lost, almost impossible to regain. Incorporation of a fair, open process of public participation and dialogue has been advocated as one important means of increasing public trust [12, 15].

Communication must be seen as a process, and not simply as effects

Risk communication is not simply relaying a one-way message about risk to those potentially affected by the risk situation. It is instead a process of mutual learning, discourse, and negotiation on both the nature of the risk and the best means of managing the risk in a manner acceptable to all involved [3].

All communication should be open, honest, and accurate

Many agencies still ascribe to the notion that a premature release of uncertain information will cause undue worry and distress. However, we know that the exact opposite is true – people become more concerned if they discover that important information has been withheld, on the basis that they cannot trust agencies to provide them with timely information. Contrary to popular belief, most people are fully capable of understanding uncertainty and accepting that information may change, as knowledge about the risk situation improves. They are usually much more accepting of information that is continually updated than they are of information that is seen to be kept “secret” by risk agencies [12].

Communications should be balanced

Risk is seldom a “black and white” issue, with some chemicals or circumstances being universally “dangerous” and others being universally “safe”. Indeed, risk is more often an exercise in moderation. For example, consuming aspirin at accepted doses is acknowledged to be beneficial for everything from relieving pain to preventing heart attacks. However, at higher doses aspirin can be lethal. This is the basis of the toxicological principal first espoused by Paracelsus in the 1500s, “All substances are poisons; there is none which is not a poison. The right dose differentiates a poison from a remedy”. Despite this knowledge, communication about risks often focuses solely on the negative aspects of the risk, without acknowledging possible benefits. One of the most notable examples of this imbalance was in northern Canada, where indigenous residents were initially advised to stop eating their traditional food because of contamination with polychlorinated biphenyls (PCBs). It was later recognized that the benefits of consuming traditional foods (including, but not limited to, a greatly reduced rate of heart disease) needed to be balanced against the relatively small exposure to a toxic substance for which the risk information is still considered equivocal for cancer related effects [16].

Two-way communication involves listening as well as speaking

As noted elsewhere (*see Role of Risk Communication in a Comprehensive Risk Management Approach*), risk communication is now acknowledged to be “an interactive process of exchange

of information and opinion among individuals” that “raises the level of understanding of relevant issues or actions for those involved and satisfies them that they are adequately informed within the limits of available knowledge” [3, p. 2]. True two-dialogue cannot be achieved unless both parties are committed to listening and learning from diverse viewpoints.

People cannot communicate what they do not know or understand

This seemingly intuitive principle of risk communication was listed by Berlo [17] as one of the key elements of his “Source-Message-Channel-Receiver (SMCR)” model of communication. However, it is still an established practice in many agencies to relegate formal risk communication activities to people in their communication or public relations departments who have little specific knowledge of the more technical aspects of the risk. Conversely, if this responsibility is relegated to those with knowledge of the risk assessment process, they may lack knowledge of communication processes and considerations. Either situation usually results in confusion and frustration in risk communication efforts. Risk communicators need to both understand the technical aspects of the risk and have knowledge of risk communication processes.

Communication should be meaningful

If agencies embark on a risk communication process, they must be prepared to ensure that it is adequately and comprehensively conducted. If information on a risk is considered necessary for people to make informed decisions on a risk, then that information should be made available to all potentially affected people. For example, issuing information on a fish consumption advisory, but not ensuring that it reaches subsistent consumers of fish is not meaningful communication [12]. Likewise, involving people in a dialogue on risk means that agencies must be prepared to listen and learn from multiple viewpoints, and incorporate different types of knowledge into the risk discourse.

Agencies must accept responsibility for risk issues and devote adequate resources to their assessment and communication

Risk issues will never be adequately dealt with if no one assumes responsibility for their assessment, communication, and management [18]. This cannot be accomplished unless sufficient resources are made

available to do this in a comprehensive and sustained manner. In addition, responsible organizations must be careful to adequately balance their efforts and resource allocation. Most agencies, in dealing with risks, expend most of their resources on the technical assessment of the risks. However, the social discourse on risk is perhaps ultimately the most important component of risk management – if risk assessment information is not being adequately communicated and discussed in formulating appropriate risk decisions, ultimately it is for naught. Ideally, the resources devoted to the social management of risk should be equal to those devoted to the technical assessment of risks [19].

Summary

Successful risk communication will not occur serendipitously. Careful consideration must be paid to advance planning and to the many lessons that have been learned through previous (mostly unsuccessful) risk communication efforts. The considerations presented here may help in this process. However, it must be stressed that this list should not be considered either fully comprehensive or exhaustive. The best overall guidance is to try to view the risk through the eyes of those who may legitimately and rationally view the nature and consequences of the risk differently. Evoking an empathetic understanding of different perspectives of risk, together with the application of good judgment and plain old-fashioned common sense, is ultimately the key to initiating a productive dialogue on risks.

References

- [1] Covello, V.T., von Winterfeldt, D. & Slovic, P. (1987). Communicating scientific information about health and environmental risks: problems and opportunities from a social and behavioral perspective, in *Uncertainty in Risk Assessment, Risk Management, and Decision-Making*, V.T. Covello, L.B. Lave, A. Moghissi & V.R.R. Uppuluri, eds, Plenum Press, New York, pp. 221–239.
- [2] Leiss, W. & Krewski, D. (1989). Risk communication: theory and practice, in *Prospects and Problems in Risk Communication*, W. Leiss, ed, University of Waterloo Press, Waterloo, pp. 89–112.
- [3] U.S. National Research Council (1989). *Improving Risk Communication*, National Academy Press, Washington, DC.
- [4] U.S. Environmental Protection Agency (1992). *Seven Cardinal Rules of Risk Communication*. EPA 230-K-92-001, May 1992.
- [5] Covello, V.T., Sandman, P.M. & Slovic, P. (1988). Risk communication, in *Carcinogenic Risk Assessment*, C.C. Travis, ed, Plenum Press, New York, pp. 193–207.
- [6] Sandman, P.M. (1989). Hazard versus outrage in the public perception of risk, in *Effective Risk Communication: The Role and Responsibility of Government and Nongovernment Organizations*, V.T. Covello, D.B. McCallum & M.T. Pavlova, eds, Plenum Press, New York, pp. 45–49.
- [7] Morgan, M.G., Fischhoff, B., Bostrom, A., Lave, L. & Atman, C.J. (1992). Communicating risk to the public, *Environmental Science and Technology* **26**(11), 2048–2056.
- [8] Atman, C.J., Bostrom, A., Fischhoff, B. & Morgan, G.M. (1994). Designing risk communications: completing and correcting mental models of hazardous processes, Part I, *Risk Analysis* **14**(5), 779–788.
- [9] Morgan, M.G., Fischhoff, B., Bostrom, A. & Atman, C.J. (2002). *Risk Communication: A Mental Models Approach*, Cambridge University Press, New York.
- [10] U.S. National Research Council (1996). *Understanding Risk: Informing Decisions in a Democratic Society*, National Research Council, National Academy Press, Washington, DC.
- [11] U.S. Presidential/Congressional Commission on Risk Assessment and Risk Management (1997). *Framework for Environmental Health Risk Management*, Final Report, Washington, DC, Vol. 1 and 2, at <http://www.risk-world.com/Nreports/1997/risk-rpt/pdf/EPAJAN.PDF>.
- [12] Jardine, C.G. (2003). Development of a public participation and communication protocol for establishing fish consumption advisories, *Risk Analysis* **23**(3), 461–471.
- [13] Slovic, P. (1986). Informing and educating the public about risk, *Risk Analysis* **6**(4), 403–415.
- [14] Slovic, P. (1993). Perceived risk, trust, and democracy, *Risk Analysis* **13**(6), 675–682.
- [15] Bradbury, J.A., Kristi, K.M. & Focht, W. (1999). Trust and public participation in risk policy issues, in *Social Trust and the Management of Risk*, G. Cvetkovich & R.E. Löfstedt, eds, Earthscan Publication, London, pp. 117–127.
- [16] Van Oostdam, J., Donaldson, S.G., Feeley, M., Arnold, D., Ayotte, P., Bondy, G., Chan, L., Dewailly, É., Furgal, C.M., Kuhnlein, H., Loring, E., Muckle, G., Myles, E., Receveur, O., Tracy, B., Gill, U. & Kalhok, S. (2005). Human health implications of environmental contaminants in Arctic Canada: a review, *Science of the Total Environment* **351–352**, 165–246.
- [17] Berlo, D.K. (1960). *The Process of Communication: An Introduction to Theory and Practice*, Holt, Rinehard, and Winston, New York.
- [18] Powell, D. & Leiss, W. (2004). *Mad Cows and Mother's Milk*, 2nd Edition, McGill-Queen's University Press, Montreal.
- [19] Leiss, W. (2001). *In the Chamber of Risks: Understanding Risk Controversies*, McGill-Queen's University Press, Montreal.

Related Articles

Risk and the Media

**Scientific Uncertainty in Social Debates Around
Risk**

**Stakeholder Participation in Risk Management
Decision Making**

CYNTHIA G. JARDINE

Continuous-Time Asset Allocation

Asset allocation (*see Actuary; Asset–Liability Management for Nonlife Insurers; Longevity Risk and Life Annuities*) is the distribution of one’s wealth among several securities so as to achieve a financial goal according to certain preferences or criteria. In a continuous-time environment, such a distribution is allowed to be rebalanced (*see Role of Alternative Assets in Portfolio Construction*) continuously over the entire investment horizon. Expected utility maximization (EUM) (*see Clinical Dose–Response Assessment*) and mean–variance (MV) portfolio (*see Scenario-Based Risk Management and Simulation Optimization; Risk Measures and Economic Capital for (Re)insurers; Extreme Value Theory in Finance*) selection, both in a given, finite investment horizon, are two predominant models for continuous-time asset allocation (*see Equity-Linked Life Insurance; Repeated Measures Analyses*). One of the two widely used approaches in the study of these models is formulating them as (dynamic) stochastic control problems (*see Role of Alternative Assets in Portfolio Construction*), and then solving them *via* standard control techniques such as dynamic programming, maximum principle, and linear–quadratic control. This approach is referred to as the *primal* or *forward* one, since it derives forward optimal strategies all at once. The other approach is to first identify the optimal terminal wealth (at the end of the horizon) by solving certain static optimization problems, and then finding optimal strategies by replicating the optimal terminal wealth. This is referred to as the *dual* or *backward* approach because the investment strategies are obtained in a reversed (in time) fashion, on the basis of the terminal wealth. This approach intimately relates to the hedging of contingent claims; and hence it reveals an inherent connection between asset allocation and hedging/pricing.

In this article, the primal and dual approaches in continuous-time asset allocation, for each of the EUM and MV models, are described.

The Market

Throughout this article, $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ is a filtered probability space (satisfying standard assumptions),

which represents the underlying random world and the progressive revelation of information with the passage of time. A standard m -dimensional Brownian motion (*see Insurance Pricing/Nonlife; Simulation in Risk Management*) $\{W(t); t \geq 0\}$, which is the source of the uncertainty, exists; it is assumed that $\mathcal{F}_t$ is generated by this Brownian motion. Also, the time horizon under consideration is $[0, T]$ throughout, where $T > 0$ is given and fixed.

Suppose there is a market in which $m + 1$ assets (or securities) are traded continuously. One of the assets is a *bank account* or a *bond* whose value $S_0(t)$ is subject to the following ordinary differential equation (ODE):

$$dS_0(t) = r(t)S_0(t) dt; \quad S_0(0) = s_0 > 0 \quad (1)$$

where the *interest rate* $r(t) \geq 0$ is a measurable and uniformly bounded function of t . The other m assets are *stocks* whose price processes $S_1(t), \dots, S_m(t)$ satisfy the following stochastic differential equation (SDE) (*see Statistics for Environmental Toxicity*):

$$dS_i(t) = S_i(t) \left[b_i(t) dt + \sum_{j=1}^m \sigma_{ij}(t) dW^j(t) \right];$$

$$S_i(0) = s_i > 0, \quad i = 1, 2, \dots, m \quad (2)$$

where $b_i(t)$ and $\sigma_{ij}(t)$ are respectively the *appreciation rates* and *dispersion* (or *volatility*) rates of the stocks. Once again, we assume that $b_i(t)$ and $\sigma_{ij}(t)$ are measurable and uniformly bounded functions of t .

Define the *volatility matrix* $\sigma(t) := (\sigma_{ij}(t))_{m \times m}$ and the *excess rate of return vector* $B(t) = (b_1(t) - r(t), \dots, b_m(t) - r(t))$. The basic assumption throughout is, for some $\delta > 0$ (*see Risk Attitude*),

$$\sigma(t)\sigma(t)' \geq \delta I, \quad \forall t \in [0, T] \quad (3)$$

Consider an agent whose total wealth at time $t \geq 0$ is denoted by $x(t)$. Assume that the trading of shares takes place continuously in a self-financed fashion, and transaction cost is not considered. Then $x(\cdot)$ satisfies the *wealth equation*:

$$dx(t) = [r(t)x(t) + B(t)\pi(t)] dt$$

$$+ \pi(t)' \sigma(t) dW(t); \quad x(0) = x_0 \quad (4)$$

where $\pi(\cdot) := (\pi_1(\cdot), \dots, \pi_m(\cdot))'$, with $\pi_i(t)$, $i = 1, 2, \dots, m$, denoting the total market value of the

agent's wealth in the i th asset at time t , is a *portfolio* or *investment strategy* of the agent.

A portfolio $\pi(\cdot)$ is called *admissible* if $\pi(\cdot)$ is $\mathcal{F}_t$ -progressively measurable and square integrable. Clearly, the SDE equation (4) has a unique solution $x(\cdot)$ corresponding to each admissible $\pi(\cdot)$, and we refer to $(x(\cdot), \pi(\cdot))$ as an *admissible (wealth–portfolio) pair*. The set of all admissible portfolios is denoted by Π .

Some preliminaries are in order. Define $\theta(t) = \sigma^{-1}(t)B(t)'$, the *market price of risk*, as well as a *pricing kernel* $\rho(\cdot)$ via the following SDE:

$$d\rho(t) = \rho(t)[-r(t)dt - \theta(t)'dW(t)]; \quad \rho(0) = 1 \quad (5)$$

Itô's formula shows that

$$x(t) = \rho(t)^{-1}E(\rho(T)x(T)|\mathcal{F}_t), \quad \forall t \in [0, T], \text{ a.s.} \quad (6)$$

and, in particular,

$$E(\rho(T)x(T)) = x_0 \quad (7)$$

which is termed the *budget constraint*. This single constraint is substituted for the dynamic budget constraint (the wealth equation), which also specifies the range of possible terminal wealth values (as random variables) induced by admissible portfolios, given the initial budget x_0 .

Expected Utility Maximization Model

In an EUM model, an agent makes asset allocation decisions based on some preferences, which are captured by a *utility function* $U(\cdot): \Re \mapsto \Re^+$, on the terminal wealth. This function is typically assumed to be concave, to reflect the *risk-averse* nature of a rational investor. More technical assumptions on $U(\cdot)$ (such as the *Inada* conditions) may be imposed in order for the EUM model to be solvable.

Given a utility function $U(\cdot)$ and an initial budget x_0 , the EUM is to

$$\begin{aligned} &\text{Maximize} && J_{\text{EUM}}(\pi(\cdot)) = EU(x(T)) \\ &\text{subject to} && (x(\cdot), \pi(\cdot)) \text{ admissible, satisfying} \\ &&& \text{equation (4).} \end{aligned} \quad (8)$$

The primal (forward) approach to solve the problem, which is essentially that of the stochastic control, is described first. Let v be the *value function* corresponding to equation (8), that is, $v(t, x)$ is the optimal value of equation (8) if the initial time is t (instead of 0) and the initial budget is x (instead of x_0). Then v satisfies the *Hamilton–Jacobi–Bellman* (HJB) equation:

$$\begin{cases} v_t + \sup_{\pi \in \mathbb{R}^m} \left(\frac{1}{2} \pi' \sigma \sigma' \pi v_{xx} + B \pi v_x \right) \\ \quad + r x v_x = 0, \quad (t, x) \in [0, T) \times \Re \\ v(T, x) = U(x) \end{cases} \quad (9)$$

The *verification theorem* in stochastic control dictates that the optimal portfolio (control) is the one that achieves the supremum above, that is,

$$\pi^*(t, x) = -(\sigma(t)')^{-1} \theta(t) \frac{v_x(t, x)}{v_{xx}(t, x)} \quad (10)$$

The above optimal strategy is expressed in a *feedback* form as a function of the calendar time t and the wealth x , which is easy to use (if v is known). To get v , plugging this expression back to equation (9) we obtain the following partial differential equation (PDE) that v must satisfy:

$$\begin{cases} v_t - \frac{1}{2} |\theta|^2 \frac{(v_x)^2}{v_{xx}} + r x v_x = 0, \\ \quad (t, x) \in [0, T) \times \Re \\ v(T, x) = U(x) \end{cases} \quad (11)$$

The solution to the above PDE depends on the choice of the utility function U , and could be obtained analytically for certain classes of utility functions. The general procedure of the primal approach is, therefore, (a) to solve PDE equation (11) first; and (b) to obtain the optimal feedback strategy π^* via equation (10).

The second approach, the dual (backward) one, involves the pricing kernel ρ defined by equation (5). In view of the budget constraint equation (7), one solves first a static optimization problem in terms of the terminal wealth, X :

Maximize $EU(X)$
 subject to $E[\rho(T)X] = x_0$; X is an
 $\mathcal{F}_T$ -measurable random variable (12)

This is a constrained convex optimization problem. To solve it, one introduces a Lagrange multiplier λ to eliminate the linear constraint, leading to the following problem

Minimize $E[U(X) - \lambda\rho(T)X]$ (13)

The solution is, obviously, $X^* = (U')^{-1}(\lambda\rho(T))$, where λ is determined in turn by the original constraint

$$E[\rho(T)(U')^{-1}(\lambda\rho(T))] = x_0 \quad (14)$$

The optimal portfolio to equation (8) and the corresponding wealth process can be determined by replicating X^* . This is realized by solving the following backward stochastic differential equation (BSDE) in $(x^*(\cdot), z^*(\cdot))$:

$$\begin{aligned} dx^*(t) &= [r(t)x^*(t) + \theta(t)'z^*(t)]dt + z^*(t)'dW(t), \\ x^*(T) &= (U')^{-1}(\lambda\rho(T)) \end{aligned} \quad (15)$$

and setting

$$\pi^*(t) = (\sigma(t)')^{-1}z^*(t) \quad (16)$$

Then $(x^*(\cdot), \pi^*(\cdot))$ is the optimal pair.

It remains to solve the BSDE equation (15). One way is to employ the so-called four-step scheme that, in the current setting, starts with conjecturing $x^*(t) = f(t, \rho(t))$ for some function f . Applying Itô's formula and noting equation (5), we obtain

$$\begin{aligned} dx^*(t) &= \left(f_t - r\rho f_\rho + \frac{1}{2}|\theta|^2\rho^2 f_{\rho\rho} \right) dt \\ &\quad - \rho f_\rho \theta' dW(t) \end{aligned} \quad (17)$$

Comparing the drift and diffusion terms between equations (15) and (17), we derive the following PDE that f must satisfy

$$\begin{cases} f_t + \frac{1}{2}|\theta|^2\rho^2 f_{\rho\rho} + (|\theta|^2 - r)\rho f_\rho - rf = 0, \\ (t, \rho) \in [0, T] \times \mathbb{R}^+ \\ f(T, \rho) = (U')^{-1}(\lambda\rho) \end{cases} \quad (18)$$

This is actually a *Black-Scholes equation* (see **Risk-Neutral Pricing: Importance and Relevance**;

Weather Derivatives) arising in option pricing. The optimal strategy is, therefore,

$$\begin{aligned} \pi^*(t) &= (\sigma(t)')^{-1}z^*(t) \\ &= -(\sigma(t)')^{-1}\theta(t)\rho(t)f_\rho(t, \rho(t)) \end{aligned} \quad (19)$$

To recapture, one carries out the following steps in the dual approach: (a) solving equation (14) to get λ ; (b) solving equation (18) to obtain the function f ; and (c) applying equation (19) to obtain the optimal strategy.

Mean-Variance Portfolio Selection Model

An MV model is to minimize the variance of the terminal wealth after a target expected terminal return is specified. Mathematically, it is formulated as a constrained stochastic optimization problem, parameterized by $a \geq x_0 e^{\int_0^T r(t) dt}$:

Minimize $J_{MV}(\pi(\cdot)) = \text{Var}(x(T))$
 subject to $Ex(T) = a$; $(x(\cdot), \pi(\cdot))$ admissible,
 satisfying equation (4) (20)

The optimal strategy of the above problem is called *efficient*.

To handle the constraint $Ex(T) = a$, we apply the Lagrange multiplier technique to minimize the following cost functional

$$\begin{aligned} J(\pi(\cdot), \lambda) &:= E\{|x(T)|^2 - a^2 - 2\lambda[x(T) - a]\} \\ &= E|x(T) - \lambda|^2 - (\lambda - a)^2 \end{aligned} \quad (21)$$

for each fixed $\lambda \in \mathbb{R}$.

To solve the preceding problem (ignoring the term $-(\lambda - a)^2$ in the cost), we use the *completion-of-square* technique common in stochastic linear-quadratic control. Applying Itô's formula we obtain

$$\begin{aligned} d \left[e^{-\int_t^T (|\theta(s)|^2 - 2r(s)) ds} \left(x(t) - \lambda e^{-\int_t^T r(s) ds} \right)^2 \right] \\ = e^{-\int_t^T (|\theta(s)|^2 - 2r(s)) ds} \\ \times \left[\left| \sigma(t)' \pi(t) + \theta(t) \left(x(t) - \lambda e^{-\int_t^T r(s) ds} \right) \right|^2 dt \right. \\ \left. + 2 \left(x(t) - \lambda e^{-\int_t^T r(s) ds} \right) \pi(t)' \sigma(t) dW(t) \right] \end{aligned} \quad (22)$$

Integrating in time and taking expectation, and employing a standard stopping time technique, we have

$$E|x(T) - \lambda|^2 \geq e^{-\int_0^T (|\theta(s)|^2 - 2r(s)) ds} \times \left(x_0 - \lambda e^{-\int_0^T r(s) ds} \right) \quad (23)$$

and the equality holds if and only if

$$\pi(t) = -(\sigma(t)')^{-1} \theta(t) \left(x(t) - \lambda e^{-\int_t^T r(s) ds} \right) \quad (24)$$

It remains to determine the value of λ . Notice we have shown above that the minimum value of $J(\pi(\cdot), \lambda)$ over Π for each fixed λ is $J^*(\lambda) = e^{-\int_0^T (|\theta(s)|^2 - 2r(s)) ds} (x_0 - \lambda e^{-\int_0^T r(s) ds}) - (\lambda - a)^2$. Hence the convex duality theorem indicates that the value of λ for solving equation (20) is the one that maximizes $J^*(\lambda)$, which is

$$\lambda = \frac{a - x_0 e^{\int_0^T [r(t) - |\theta(t)|^2] dt}}{1 - e^{-\int_0^T |\theta(t)|^2 dt}} \quad (25)$$

To summarize, the optimal strategy for solving equation (20) is given by equation (24) (which is actually in a feedback form), where λ is given by equation (25). Now, treating $Ex(T) = a$ as a parameter, we can easily obtain the optimal value of equation (20) as follows:

$$\text{Var}(x(T)) = \frac{1}{e^{\int_0^T |\theta(t)|^2 dt} - 1} \left[Ex(T) - x_0 e^{\int_0^T r(t) dt} \right]^2, \\ Ex(T) \geq x_0 e^{\int_0^T r(t) dt} \quad (26)$$

This gives a precise trade-off between the mean (return) and variance (risk) that an efficient strategy could achieve, in the classical spirit of Markowitz. Equation (26) gives rise to the *efficient frontier*. It is a straight line if plotted on a mean–standard deviation plane.

In addition to the primal approach depicted above, the MV model (20) can also be solved by the dual approach. Consider the following optimization problem in terms of the terminal wealth:

$$\begin{aligned} &\text{Minimize} \quad EX^2 - a^2 \\ &\text{subject to} \quad E[\rho(T)X] = x_0; \quad E[X] = a; \quad X \text{ is an} \\ &\quad \mathcal{F}_T\text{-measurable random variable} \end{aligned} \quad (27)$$

There are two constraints, calling for two Lagrange multipliers $(2\mu, -2\lambda)$ so that one solves

$$\text{Minimize}_{X \text{ is } \mathcal{F}_T\text{-measurable}} E[X^2 - 2\lambda X + 2\mu\rho(T)X] \quad (28)$$

The optimal solution to such a relaxed problem is $X^* = \lambda - \mu\rho(T)$, with the constants λ and μ satisfying

$$E[\rho(T)(\lambda - \mu\rho(T))] = x_0, \quad E[\lambda - \mu\rho(T)] = a \quad (29)$$

Since these equations are linear, the solution is immediate:

$$\begin{aligned} \lambda &= \frac{aE[\rho(T)^2] - x_0E[\rho(T)]}{\text{Var}(\rho(T))}, \\ \mu &= \frac{aE[\rho(T)] - x_0}{\text{Var}(\rho(T))} \end{aligned} \quad (30)$$

To obtain the replicating portfolio and the corresponding wealth process $(x^*(\cdot), \pi^*(\cdot))$, we use equation (6):

$$\begin{aligned} x^*(t) &= \rho(t)^{-1} E\left[(\lambda - \mu\rho(T))\rho(T) | \mathcal{F}_t\right] \\ &= \lambda e^{-\int_t^T r(s) ds} - \mu e^{-\int_t^T (2r(s) - |\theta(s)|^2) ds} \rho(t) \end{aligned} \quad (31)$$

A direct computation on the above, using equation (5), yields

$$\begin{aligned} dx^*(t) &= \left[rx(t) + \mu|\theta(t)|^2 \rho(t) e^{-\int_t^T (2r(s) - |\theta(s)|^2) ds} \right] dt \\ &\quad + \mu e^{-\int_t^T (2r(s) - |\theta(s)|^2) ds} \rho(t) \theta(t)' dW(t) \end{aligned} \quad (32)$$

Comparing the above with the wealth equation (4), we conclude that

$$\begin{aligned} \pi^*(t) &= (\sigma(t)')^{-1} \theta(t) \mu e^{-\int_t^T (2r(s) - |\theta(s)|^2) ds} \rho(t) \\ &= -(\sigma(t)')^{-1} \theta(t) \left(x^*(t) - \lambda e^{-\int_t^T r(s) ds} \right) \end{aligned} \quad (33)$$

This, certainly, agrees with equation (24) derived *via* the forward approach.

Notes

Bachelier [1] was the first to use Brownian motion to model the dynamics of security prices, whereas Mirrlees [2, 3] is probably the first to employ the Itô calculus to study continuous-time asset allocation. A continuous-time asset allocation model is significantly different from its single-period counterpart because the possibility of continuous trading effectively changes the investment set from one with a finite number of securities to one with an infinite number of attainable contingent claims [22]. The stochastic control (or forward/primal) approach described in the section titled “Expected Utility Maximization Model”, including the use of the HJB equation, was put forward and developed by Merton [4, 5] for solving EUM. Latest and comprehensive sources for general stochastic control theory and applications are Fleming and Soner [6] and Yong and Zhou [7]. Extensions following Merton’s method have abounded; refer to Duffie [8] and Karatzas and Shreve [9] for more historical notes. The backward/dual approach (also known as the *martingale approach*), described in the section “Expected Utility Maximization Model”, was developed by Harrison and Kreps [10], Harrison and Pliska [11], Pliska [12], Cox and Huang [13], and Karatzas *et al.* [14], except that in these works, the martingale representation theorem is used instead of BSDEs. This approach has its roots in Bismut [15, 16] who happens to be the father of the BSDE theory. (Interestingly, Bismut’s linear BSDE is originally wedded to the maximum principle in stochastic control – this shows an intrinsic connection between the primal and dual approaches in the context of asset allocation.) See Yong and Zhou [7] and Ma and Yong [17] for more on BSDEs (including extension to nonlinear BSDEs and the four-step scheme mentioned in the section “Expected Utility Maximization Model”) and the related historical notes, and Duffie [8] and Karatzas and Shreve [9] for more references on the dual approach for EUM. It is intriguing that the dual approach leads to a Black–Scholes type equation (18), which suggests a profound primal–dual relation between asset allocation and pricing. On the other hand, in a recent paper Jin *et al.* [18] have shown, through various counterexamples, that some of the standing assumptions in this approach, such as the existence of the Lagrange multiplier and the well posedness of the underlying model,

may not be valid, and sufficient conditions are presented to ensure the solvability of a general EUM model.

Markowitz’s (single-period) MV model [19] marked the start of the modern quantitative finance theory. Perversely enough, extensions to the dynamic – especially continuous time – setting in the asset allocation literature have been dominated by the EUM models, making a considerable departure from the MV model. While the utility approach was theoretically justified by von Neumann and Morgenstern [20], in practice “few if any investors know their utility functions; nor do the functions which financial engineers and financial economists find analytically convenient necessarily represent a particular investor’s attitude towards risk and return” (Markowitz, H. *Private communication*, 2004). On the other hand, there are technical difficulties in treating dynamic MV models, primarily that of the incompatibility with the dynamic programming principle owing to the variance term involved. Richardson [21] is probably the earliest paper that studies a faithful extension of the MV model to the continuous-time setting (albeit in the context of a single stock with a constant risk-free rate), followed by Bajoux-Besnainou and Portait [22]. Li and Ng [23] developed an embedding technique to cope with the nonapplicability of dynamic programming for a discrete-time MV model, which was extended by Zhou and Li [24], along with a stochastic LQ control approach described in the section titled “Mean–Variance Portfolio Selection Model”, to the continuous-time case. Further extensions and improvements are carried out in, among many others, Lim and Zhou [25], Lim [26], Bielecki *et al.* [27], Xia and Yan [28], and Jin and Zhou [29].

Following the Nobel prize winning work of Kahneman and Tversky [30], known as the *prospect theory* (PT), that brings human emotions and psychology into decision making, there have been burgeoning research interests in incorporating the PT into portfolio choice; nonetheless, they have been hitherto overwhelmingly limited to the single-period setting. There are only a couple of papers that deal with behavioral asset allocation in continuous time, including Berkelaar *et al.* [31] and Jin and Zhou [32]. The latter have obtained fairly explicit solutions for a general problem featuring both *S*-shaped utility functions and distorted probability.

References

- [1] Bachelier, L. (1900). *Théorie de la Speculation*, *Annales Scientifiques de L'École Normale Supérieure*, 3rd series, **17**, 21–88.
- [2] Mirrlees, J.A. (1965). *Optimal Accumulation Under Uncertainty*, unpublished paper.
- [3] Mirrlees, J.A. (1971). Optimal accumulation under uncertainty: the case of stationary returns to investment, in *Allocation Under Uncertainty: Equilibrium and Optimality*, J. Drèze, ed, John Wiley & Sons, New York.
- [4] Merton, R. (1969). Lifetime portfolio selection under uncertainty: the continuous time case, *The Review of Economics and Statistics* **51**, 247–257.
- [5] Merton, R. (1971). Optimum consumption and portfolio rules in a continuous time model, *Journal of Economic Theory* **3**, 373–413.
- [6] Fleming, W.H. & Soner, H.M. (2006). *Controlled Markov Processes and Viscosity Solutions*, 2nd Edition, Springer-Verlag, New York.
- [7] Yong, J. & Zhou, X.Y. (1999). *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York.
- [8] Duffie, D. (1996). *Dynamic Asset Pricing Theory*, 2nd Edition, Princeton University Press, Princeton.
- [9] Karatzas, I. & Shreve, S.E. (1998). *Methods of Mathematical Finance*, Springer-Verlag, New York.
- [10] Harrison, J.M. & Kreps, D. (1979). Martingales and multiperiod securities market, *Journal of Economic Theory* **20**, 381–408.
- [11] Harrison, J.M. & Pliska, S. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and their Applications* **11**, 215–260.
- [12] Pliska, S.R. (1986). A stochastic calculus model of continuous trading: optimal portfolios, *Mathematics Operations Research* **11**, 371–384.
- [13] Cox, J. & Huang, C.-F. (1989). Optimal consumption and portfolio policies when asset follows a diffusion process, *Journal of Economic Theory* **49**, 33–83.
- [14] Karatzas, I., Lehoczky, J. & Shreve, S.E. (1987). Optimal portfolio and consumption decisions for a small investor on a finite horizon, *SIAM Journal on Control and Optimization* **25**, 1157–1186.
- [15] Bismut, J.M. (1973). Conjugate convex functions in optimal stochastic control, *Journal of Mathematical Analysis and Applications* **44**, 384–404.
- [16] Bismut, J.M. (1975). Growth and optimal intertemporal allocations of risks, *Journal of Economic Theory* **10**, 239–287.
- [17] Ma, J. & Yong, J. (1999). *Forward-Backward Stochastic Differential Equations and Their Applications*, *Lecture Notes in Mathematics 1702*, Springer-Verlag, Berlin-Heidelberg.
- [18] Jin, H., Xu, Z. & Zhou, X.Y. (2007). A convex stochastic optimization problem arising from portfolio selection, *Mathematical Finance* to appear.
- [19] Markowitz, H. (1952). Portfolio selection, *Journal of Finance* **7**, 77–91.
- [20] von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, 2nd Edition, Princeton University Press, Princeton.
- [21] Richardson, H.R. (1989). A minimum variance result in continuous trading portfolio optimization, *Management Science* **35**, 1045–1055.
- [22] Bajeux-Besnainou, L. & Portait, R. (1998). Dynamic asset allocation in a mean – variance framework, *Management Science* **44**, 79–95.
- [23] Li, D. & Ng, W.L. (2000). Optimal dynamic portfolio selection: multiperiod mean–variance formulation, *Mathematical Finance* **10**, 387–406.
- [24] Zhou, X.Y. & Li, D. (2000). Continuous time mean–variance portfolio selection: a stochastic LQ framework, *Applied Mathematics and Optimization* **42**, 19–33.
- [25] Lim, A.E.B. & Zhou, X.Y. (2002). Mean–variance portfolio selection with random parameters, *Mathematics of Operations Research* **27**, 101–120.
- [26] Lim, A.E.B. (2004). Quadratic hedging and mean–variance portfolio selection with random parameters in an incomplete market, *Mathematics of Operations Research* **29**, 132–161.
- [27] Bielecki, T.R., Jin, H., Pliska, S.R. & Zhou, X.Y. (2005). Continuous-time mean–variance portfolio selection with bankruptcy prohibition, *Mathematical Finance* **15**, 213–244.
- [28] Xia, J. & Yan, J.-A. (2006). Markowitz's portfolio optimization in an incomplete market, *Mathematical Finance* **16**, 203–216.
- [29] Jin, H. & Zhou, X.Y. (2007). Continuous-time Markowitz's problems in an incomplete market, with no-shorting portfolios, in *Stochastic Analysis and Applications – A Symposium in Honor of Kiyosi Itô*, G. Di Nunno, ed, Springer-Verlag, Berlin Heidelberg.
- [30] Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk, *Econometrica* **47**, 263–291.
- [31] Berkelaar, A.B., Kouwenberg, R. & Post, T. (2004). Optimal portfolio choice under loss aversion, *The Review of Economics and Statistics* **86**, 973–987.
- [32] Jin, H. & Zhou, X.Y. (2007). Behavioral portfolio selection in continuous time, *Mathematical Finance* to appear.

Further Reading

- Cvitanic, J. & Karatzas, I. (1992). Convex duality in constrained portfolio optimization, *Annals of Applied Probability* **2**, 767–818.

XUN YU ZHOU

Copulas and Other Measures of Dependency

The field of probability includes the study of a wide class of univariate probability functions. The class of multivariate probability functions is less well studied and is mostly restricted to the multivariate normal distribution and its related functions such as the multivariate t and the Wishart distributions. The invention and study of copulas are motivated by the need to enlarge the class of multivariate distributions.

Copulas were proposed by Sklar [1, 2] as invariant transformations to combine marginal probability functions to form multivariate distributions. The word “copula” comes from Latin and means to connect or join. Sklar introduced this word and concept for linking univariate marginal distributions to form multivariate distributions. Specifically, copulas are measures of the dependent structure of the marginal distributions. Sklar’s proposal for developing multivariate distributions was to first determine the univariate marginal distributions for the variables of interest, transform the marginal cumulative distribution functions (*cdf*) to uniform variables on the $[0,1]$ interval, and form the multivariate distribution using a copula, which is the multivariate distribution of uniform marginals.

Sklar’s Theorem for Continuous Marginal Distributions

Let $F_i(x_i)$ be the *cdf* for X_i $i = 1, \dots, n$, and let $H(x_1, \dots, x_n)$ be the multivariate *cdf* with marginal $F_i(x_i)$. Since the probability integral $U_i = F_i(x_i)$ is a uniform random variable on $[0,1]$, there exists a unique copula, which is the multivariate *cdf* of uniform random variables. Thus, $C(u_1, \dots, u_n)$ on the n dimensional cube $[0, 1]^n$ is the *cdf* so that all marginals are uniform. Therefore,

$$\begin{aligned} C(u_1, \dots, u_n) &= P(U_1 \leq u_1, \dots, U_n \leq u_n) \\ &= H(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n)) \end{aligned} \quad (1)$$

Conversely,

$$\begin{aligned} H(x_1, \dots, x_n) &= C(F_1(x_1), \dots, F_n(x_n)) \\ &\text{for all } (x_1, \dots, x_n) \end{aligned} \quad (2)$$

Thus $H(x_1, \dots, x_n)$ is a multivariate distribution with marginals $F_1(x_1), \dots, F_n(x_n)$. Note that if $F_i(x_i)$ $i = 1, \dots, n$ are not all continuous, the copula is unique on the range of positive values for the marginal distributions.

Copulas as Measures of Dependence

Since most applications have been based on bivariate distributions, the following discussion will assume $n = 2$.

The importance of copulas is that they provide a general description of complex dependent structures among the variables. For the multivariate normal distribution, and more generally, the elliptical distributions, the correlation coefficient, ρ , measures the dependency of random variables as a pairwise measure of linear relationships. It is invariant for strictly increasing/decreasing linear transformations. The bivariate random variables (X_1, X_2) are strictly independent when $\rho = 0$. However, ρ is not invariant to more general transformations, it does not capture tail dependencies, nor is it defined for probability functions that do not have finite variances. Further, ρ does not describe dependent structure in nonelliptical multivariate distributions. A more general strategy for generating independent random variables is with a copula $C(u_1, u_2) = u_1 u_2$. Note

$$\begin{aligned} C(u_1, u_2) &= P(U_1 \leq u_1, U_2 \leq u_2) \\ &= u_1 u_2 \\ &= H(F_1^{-1}(u_1) F_2^{-1}(u_2)) \end{aligned} \quad (3)$$

so that

$$H(x_1, x_2) = F_1(x_1) F_2(x_2) \quad (4)$$

If the random variables are not independent, their relationship is summarized in the joint *cdf* $H(x_1, x_2)$ or, equivalently, with the two-dimensional copula

$$H(x_1, x_2) = C(F_1(x_1) F_2(x_2)) \quad (5)$$

Both parametric and nonparametric methods have been suggested in the literature for estimating copulas. Parametric methods are often based on the method of maximum-likelihood estimation and nonparametric procedures have been proposed, for example, by Genest and Rivest [3].

The normal copula is a popular copula for generating a bivariate normal distribution with correlation ρ . Let $\Phi(x_1)$ be the *cdf* of the standard normal distribution and let $\Phi_\rho(x_1, x_2)$ be the *cdf* of the standard bivariate normal and $\Theta_\rho(x_1, x_2)$ be the probability density function of the standard bivariate normal. Then,

$$\begin{aligned} C(u_1, u_2) &= \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \Theta_\rho(x_1, x_2) dx_1 dx_2 \\ &= \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \end{aligned} \quad (6)$$

Since many physical phenomena exhibit characteristics not described by a normal distribution such as skewness, fat tails, and outliers, the normal copula has been generalized to the form

$$C(u_1, u_2) = H_\delta(F_1^{-1}(u_1), F_2^{-1}(u_2)) \quad (7)$$

where the marginal distributions can come from different families of distributions and the dependent relationships between random variables may be more complex than linear correlation. For this reason, alternative methods of dependence have been studied.

The two most popular measures for Kendall's τ and Spearman's ρ that will be denoted as ρ_s . Both these measures lie in the $[-1, 1]$ interval for continuous random variables, can be written as a function of copulas (see Schweizer and Wolff [4]), and are invariant under monotone transformations applied to each marginal distribution. Also, for the bivariate normal, Pearson's correlation coefficient, ρ_p , relates these nonparametric measures as follows:

$$\rho_p = \text{Sin}\left(\frac{\pi}{2}\rho_s\right) \quad (8)$$

$$\rho_p = 2\text{Sin}\left(\frac{\pi}{6}\tau\right) \quad (9)$$

In fact, the normal copula is easily extended beyond the bivariate case because it is not restricted to a correlation matrix consisting of Pearson correlation coefficients (see Frees and Valdez [5]).

To digress, Johnson and Tenenbein [6] proposed a method for constructing continuous bivariate distributions with specified marginals and dependence measures based on Spearman's ρ_s and Kendall's τ .

Generating Copulas

Copulas are important for constructing multivariate distributions by providing measures of dependence

for combining marginal distributions. Archimedean copulas, described in Nelsen [7, Chapter 4], are a large class of copulas generated by an additive, continuous, decreasing convex function $\varphi(t)$ that maps t from the $[0, 1]$ interval, onto the $[0, \infty]$ range. Thus, a multivariate copula in this family can represent the association between variables with given marginal *cdf*'s by a univariate function. Copulas in this class are symmetrical, satisfying the relationship $C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v)) = C(v, u)$ where φ^{-1} is the inverse of the generator. Genest and Rivest [3] proposed a methodology for identifying $\varphi(t)$. The Archimedean copulas are easily constructed and appear in a simple closed form that allows for a wide variety of asymmetrical forms of tail distribution dependence. A discussion and listing of both bivariate and multivariate copulas in this class are available in Nelsen [7]. Thus, for example, using a linear combination of Archimedean copulas describing the dependency of two marginal t -distributions gives rise to a fat-tailed distribution with different tail dependencies. Canela and Collazo [8] propose using the *EM* algorithm, for determining weights to combine a Clayton copula that provides lower tail dependence with a Gumbel/Joe copula that provides upper tail dependence. Another procedure developed by Genest *et al.* [9] for developing asymmetrical copulas was based on developing a family of nonexchangeable bivariate copulas. These distributions play an important role in applications of extreme value theory to the study of risk; see Genest and Rivest [10]. Joe [11, Chapter 6] provides a summary of univariate extreme value theory with extensions to multivariate extreme distributions based on copulas (see **Extreme Value Theory in Finance**). This chapter concludes with unsolved problems in this area.

Frees and Valdez [5] presented a procedure for simulating outcomes from a multivariate distribution based on copulas. Then they illustrated their methodology for generating losses and expenses of insurance company indemnity claims that were subject to censoring.

The need for general measures of dependency is also apparent in the number of articles in this encyclopedia that are based on copulas; Dhaene *et al.* (see **Comonotonicity**) and Frees (see **Dependent Insurance Risks**) recommend copulas for modeling correlated risks as well as groups of individuals that are exposed to similar economic and physical environments. Wang (see **Credit Value at Risk**) uses

copulas to model joint default probabilities in credit portfolios, and Kapadia (*see Default Correlation*) modeled correlated default probabilities of individual firms. In a paper on simulation, McLeish and Metzler (*see Simulation in Risk Management*) proposed copula models of default times of individual obligors for studying credit risk. Other applications include modeling the probability survival of groups of individuals as well as modeling joint mortality patterns.

References

- [1] Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges, *Public Institute Statistical University, Paris* **8**, 229–231.
- [2] Sklar, A. (1973). Random variables, joint distribution functions and copulas, *Kybernetika* **9**, 449–460.
- [3] Genest, C. & Rivest, L.P. (1993). Semiparametric inference procedures for bivariate Archimedean copulas, *Journal of the American Statistical Association* **88**, 1034–1043.
- [4] Schweizer, B. & Wolff, E.F. (1981). On nonparametric measures of dependence for random variables, *The Annals of Statistics* **9**, 879–885.
- [5] Frees, E.W. & Valdez, E.A. (1998). Understanding relationships using copulas, *North American Actuarial Journal* **2**(1), 1–25.
- [6] Johnson, M. & Tenenbein, A. (1981). A bivariate distribution family with specified marginal's, *Journal of the American Statistical Association* **76**, 198–201.
- [7] Nelsen, R.B. (1998). *An Introduction to Copulas*, Springer-Verlag, New York.
- [8] Canela, M.A. & Collazo, E.P. (2005). *Modelling Dependence in Latin American Markets Using Copula Functions*, Draft.
- [9] Genest, C., Ghoudi, K. & Rivest, L.P. (1998). Discussion on “understanding relationships using copulas”, *North American Actuarial Journal* **2**(2), 143–146.
- [10] Genest, C. & Rivest, L.P. (1989). A characterization of Gumbel's family of extreme value distributions, *Statistics and Probability Letters* **8**, 207–211.
- [11] Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

EDWARD L. MELNICK AND AARON TENENBEIN

Correlated Risk

Correlated Risks from Natural Disasters

Correlated risk refers to the *simultaneous occurrence* of many losses from a single event. As pointed out earlier, natural disasters such as earthquakes, floods, and hurricanes produce highly correlated losses: many homes in the affected area are damaged and destroyed by a single event.

If a risk-averse insurer faces a highly correlated losses from one event, it may want to set a high enough premium not only to cover its expected losses but also to protect itself against the possibility of experiencing catastrophic losses. An insurer will face this problem if it has many eggs in one basket, such as providing earthquake coverage mainly to homes in Los Angeles rather than diversifying across the entire state of California.

To illustrate the impact of correlated risks on the distribution of losses, assume that there are two policies sold against a risk where $p = 0.1$, $L = \$100$. The actuarial loss for each policy is \$10. If the losses are perfectly correlated, then there will be either two losses with probability of 0.1, or no losses with a probability of 0.9. On the other hand, if the losses are independent of each other, then the chance of

two losses decreases to 0.01 (i.e., 0.1×0.1), with the probability of no losses being 0.81 (i.e., 0.9×0.9). There is also a 0.18 chance that there will be only one loss (i.e., $0.9 \times 0.1 + 0.1 \times 0.9$).

The expected loss for both the correlated and uncorrelated risks is \$20.^a However, the variance is always higher for correlated than for uncorrelated risks if each has the same expected loss. Thus, risk-averse insurers will always want to charge a higher premium for the correlated risk.^b

End Notes

a. For the correlated risk the expected loss is $0.9 \times \$0 + 0.1 \times \$200 = \$20$. For the independent risk the expected loss is $(0.81 \times \$0) + (0.18 \times \$100) + (0.01 \times \$200) = \20 .

b. For more details on this point and an illustrative example, see Hogarth and Kunreuther [1].

Reference

- [1] Hogarth, R. & Kunreuther, H. (1992). Pricing insurance and warranties: ambiguity and correlated risks, *The Geneva Papers on Risk and Insurance Theory* **17**, 35–60.

HOWARD KUNREUTHER

Cost-Effectiveness Analysis

This article presents economic evaluation, in particular, cost–effectiveness analysis (CEA), as a framework to approach and inform difficult decisions about the best use of limited resources. The article focuses on the requirements of a health care policy maker attempting to determine whether to reimburse a health care technology given the information and uncertainty surrounding the decision, and/or whether to request further research to inform the decision. The approach employs a decision analysis framework. The article starts with a brief discussion of welfare economics and cost–benefit analysis, before focusing on CEA. The article covers incremental cost–effectiveness ratios (ICERs) and decision rules for cost effectiveness, methods for handling and presenting uncertainty in cost effectiveness, and the use of value of information (VOI) methods for decision making (*see* **Decision Analysis**). The general approach and specific concepts are illustrated through the use of a stylized case study.

Welfare Economics: The Foundation of Economic Evaluation

In a resource-limited environment, it is important to determine whether a proposed health program represents a good use of scarce resources (money, labor, land, capital, etc.). Welfare economics provides a framework to address the issue of whether a proposed change is a good use of resources. It is founded on the view that individuals are welfare maximizers and are in the best position to assess their own welfare (consumer sovereignty). Proposed changes are judged in terms of welfare and a program change is deemed to be “better” if it improves social welfare. In the case where a program change makes at least one person better off and no one worse off (in terms of welfare), it is said to be a “Pareto improvement”. However, most program changes involve both gainers and losers. In this case, we ask whether the gainers could compensate the losers for their loss and still be better off (Kaldor compensation test) or if losers could compensate gainers to oppose the change and still be better off

(Hicks compensation test). In either case, there is a “potential Pareto improvement”. The compensation does not actually have to be paid for the change to be considered “better”. Implicitly this means that the current state of income distribution and the distributive impact of the program change are considered to be acceptable.

Economic evaluation is about identifying, measuring, and valuing what is given up (i.e., the costs of a program) and what is gained (i.e., the benefits of a program) in order to determine whether a program is a good use of resources. Cost–benefit analysis (CBA) involves individuals expressing their willingness to pay (in monetary terms) for the welfare changes (both positive and negative) associated with the proposed program. These valuations encapsulate all that is important to the individual, and all individuals whose welfare is affected are included within the process. The monetary valuations are then combined and compared to the costs associated with the change to determine whether the change is worthwhile. The major benefit of CBA is that, under the assumption that the distribution of income and impact of the change is acceptable, it is possible to determine that a program change is worthwhile *per se*, where the monetary valuation of the benefits exceeds the monetary value of the costs. CBA has been widely used within both environmental and transport economics to appraise projects such as road expansions, new airports, and environmental protection [1, 2]. However, within assessments of health care programs CBA has not been widely adopted although there are examples of applied CBA within the literature [3–7]. This is due mainly to the difficulty, value judgments, and ethical issues involved with asking individuals to assign monetary values to health care, which they are often not used to paying for.

Economic Evaluation in Health Care: The Move to Cost–Effectiveness Analysis

The more common view taken in the economic evaluation of health care is that of the extra welfarist. Here, health is considered to be the primary outcome of interest for health programs, an outcome so fundamental that individuals should receive it irrespective of their willingness to pay (known in economics as a *merit good*). With this approach, the societal objective becomes one of maximizing health gains (rather

than welfare) subject to the resource constraints. The methods of CEA are used to solve this constrained maximization problem. Within this framework, costs are measured in monetary terms and outputs are measured as utilities, typically using quality adjusted life years (QALYs). The aim with QALYs is to incorporate both the length and quality of remaining life, in order to reflect that a year in full health is not considered equivalent to a year in poor health. In calculating QALYs, the length of life is adjusted by a health-related quality of life score that reflects the quality of each year. Full health is usually represented by a quality weight of 1 with, for ease of calculation, death represented by a quality weight of 0. Sometimes natural units (e.g., life years) are used as a measure of output; however, this does not solve the constrained maximization problem.

In this framework, it is not possible to identify that a policy change is worthwhile *per se*. Instead, results are presented in terms of additional costs and additional outcomes associated with the program (compared to the alternative) to give a measure of the additional costs incurred, due to the program, for the additional health benefits secured [8–10]. This information is generally reported in the form of an ICER, although it can be provided graphically on the cost–effectiveness plane by plotting the costs against the effects for the program(s) investigated [11]. Equation (1) presents the ICER:

$$ICER = \frac{\text{Mean cost}_{\text{new program}} - \text{Mean cost}_{\text{current program}}}{\text{Mean QALY}_{\text{new program}} - \text{Mean QALY}_{\text{current program}}} \quad (1)$$

Figure 1 illustrates the cost–effectiveness plane. The plane is split into four quadrants by the origin (which represents the treatment comparator). The north east (NE) and south east (SE) quadrants involve positive incremental health effect associated with the treatment of interest, while the NE and north west (NW) quadrants involve positive incremental cost.

Decision Rules for CEA: Assessing Value for Money

A series of programs that apply to the same patient group, from which only one can be chosen, are

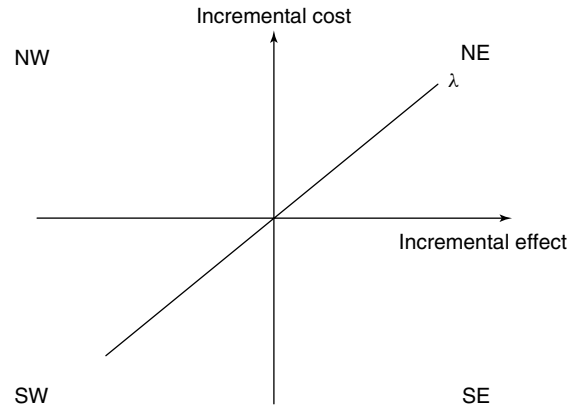


Figure 1 The incremental cost–effectiveness plane

mutually exclusive while programs that apply to different patient groups are independent (since any combination of programs can be implemented). ICERs are calculated relative to the next best, viable mutually exclusive programs; they are not calculated between independent programs. Programs are ranked in ascending order of cost, and ICERs are then calculated by dividing the additional cost by the additional health benefit involved with each successively more costly viable program [12]. On the cost–effectiveness plane, the slope of a line joining any two mutually exclusive programs denotes the ICER associated with moving from the cheaper to the more expensive program, and shallower slopes represent lower ICERs. Figure 2 illustrates an example of a series of fictional, mutually exclusive programs plotted on the cost–effectiveness plane.

On Figure 2, the program represented by point A is dominated by point B, which involves greater health benefits for a lower cost. Dominated programs should be identified and excluded from the analysis prior to calculating ICERs, to avoid spurious results. The program represented by point C is extended (or weakly) dominated, as it can be dominated by a mixed strategy involving programs B and D, represented by any point between M^1 (a strategy with identical health benefits to C that can be achieved at cheaper cost) and M^2 (a strategy involving identical costs to C that involves greater health benefits). The line OBDE presents a cost–effectiveness frontier, connecting all of the efficient programs.

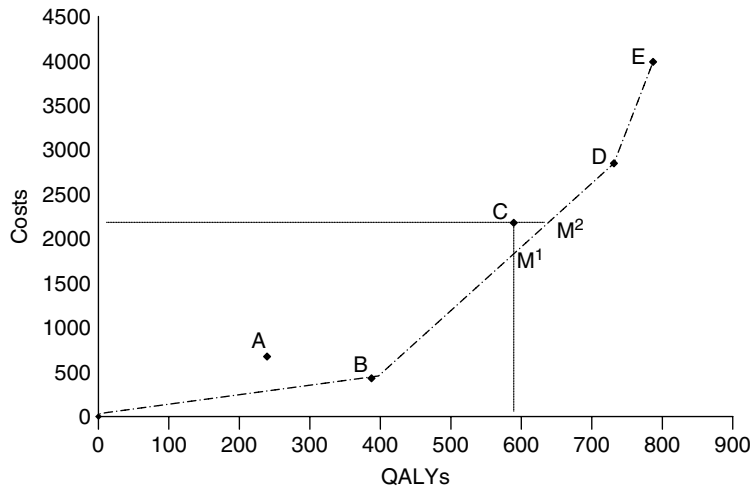


Figure 2 Mutually exclusive programs on the CE plane

Having identified the ICERs for all the efficient programs that could be implemented, a decision must be made about which program should be implemented. Traditionally, the cost-effective program has been identified as the one that generates the largest ICER that falls below an externally set cost-effectiveness threshold (λ) [12]. This threshold (λ) represents the monetary value that the policy maker assigns to the health effect generated. It can be derived as the maximum price that society is willing to pay for health benefits or as the shadow price of health benefits [12]. Within the context of a budget-constrained health service (e.g., the National Health Service (NHS) in the United Kingdom), the second approach provides the theoretically appropriate method to maximize the health outcomes from a fixed budget. However, this method involves an enormous informational requirement in order to appropriately allocate the budget [13]. The first approach with its simple accept/reject rule is more convenient for decision making, but its employment requires explicit determination of the appropriate “fixed price” for health benefits.

The process of determining the ICER can be complex, particularly when a decision involves a choice between various mutually exclusive programs. However, by explicitly incorporating the cost-effectiveness threshold (λ) and rescaling the effects into a monetary value, the costs and effects can be combined to form a measure of net benefit for each program expressed in monetary terms (NMB)

[14–18]. Equation (2) illustrates the calculation of the NMB for a program (j).

$$NMB_j = (\lambda \times QALY_j) - Cost_j \quad (2)$$

The cost-effective program is now identified simply as that with the highest NMB.

Perspective

An important element of any economic evaluation is the perspective used to calculate costs and health benefits, as this affects the extent of the costs and consequences incorporated into the evaluation. Economists advocate a societal approach, which requires measuring and valuing the impacts of the program upon everyone who is affected by it. However, narrower, more manageable perspectives (e.g., that of the NHS) can be employed if value judgments are made that this perspective captures the majority of the important effects so that the extra effort required to get a societal perspective is not necessary. A frequently used perspective is that of the health care provider (e.g., NHS or health insurance company) for costs, and the individual patient for health benefits.

Incorporating Uncertainty

Economic evaluation is concerned with measures of cost effectiveness at the population level, the

expected costs and effects associated with the program change, rather than the impact on costs and effects at an individual patient level. As such, economic evaluation is primarily concerned with uncertainty rather than patient variability. Probabilistic sensitivity analysis can be used to convert uncertainty surrounding the model parameters into uncertainty surrounding the costs and effects. The process involves specifying a probability distribution for each of the basic parameters, those estimated directly from the evidence, within the model that are considered uncertain. The probability distribution represents the range of values that the parameter can take as well as the probability that it takes any specific value. In specifying the probability distribution, all information that is currently available for the parameter should be incorporated including any logical restrictions (for example, a probability must be restricted to the range 0–1). The next step involves propagating the parameter uncertainty through the model using Monte Carlo simulation. At each iteration, a value is selected from the probability distributions for each parameter and the model is recalculated. The cost and effect results for each iteration represent a realization of the uncertainty that exists in the model, as characterized by the probability distributions. Within the Monte Carlo simulation, this process is repeated a large number of times (usually > 1000) to generate a distribution of expected costs and effects associated with the program change.

Other types of uncertainty that may be important in an economic evaluation include methodological uncertainty (uncertainty surrounding the analytical methods used in the analysis e.g., discount rates) and structural uncertainty (uncertainty surrounding the structure and the assumptions used in the analysis e.g., method of extrapolation). Both of these can be assessed through the use of sensitivity analysis e.g., one-way sensitivity analysis, multiway sensitivity analysis, extreme analysis, or threshold analysis.

The uncertainty in costs and effects is most often presented as a scatter plot on the cost–effectiveness plane or summarized using a cost–effectiveness acceptability curve (CEAC). The CEAC illustrates the (Bayesian) probability that the program change is cost effective given the data and the alternative program(s) for a range of values for the maximum cost–effectiveness threshold (λ). The complement is the probability that the program change is not cost effective given the data and alternatives (the error

probability). For more details on CEACs for decisions involving multiple treatment alternatives and the link with the error probability, please see [19, 20].

Value of Information

Given an objective to maximize health subject to a resource constraint, a program change can and should be identified as cost effective simply on the basis of the expected values (ICER or NMB). However, owing to the inevitable uncertainty surrounding the estimates of costs and effects, there is a nonnegligible possibility that a decision made on the basis of the available information will be incorrect. Bayesian VOI analysis provides a framework to formally combine the error probability with the consequences associated with making an error to assess the expected opportunity losses associated with the existing (uncertain) evidence base and, thus, value further research aimed at reducing this uncertainty (*see Bayesian Statistics in Quantitative Risk Assessment*). The techniques involve establishing the difference between the expected value of a decision made on the basis of the existing evidence and the expected value of a decision made on the basis of further information [21–24]. This difference is then compared with the cost of collecting the additional evidence; where the value of further information exceeds the costs of collecting it, the research is deemed worthwhile [15].

Expected Value of Perfect Information

Perfect information surrounding all elements of the decision would, by definition, eliminate all uncertainty. It follows that the expected value of perfect information (EVPI) is equivalent to the expected cost of the current uncertainty surrounding the decision. Hence, estimating the costs of uncertainty surrounding a decision on whether to fund a program provides a measure of the *maximum* possible value for future research. This maximum value can then be compared with the cost of gathering further information to provide a necessary (but not sufficient) condition for determining whether further research is *potentially* worthwhile [15]. Research effort can then be focused on those areas where the cost of uncertainty is high and where additional research is potentially cost effective. As information provided by research is nonrival in consumption (i.e., once information

has been generated to inform a decision, it can be used to inform the decision for all patients without reducing the information available), the societal value of research (*see Societal Decision Making*) should be calculated across the population of potential program participants [14, 15].

The EVPI can be determined directly from the distribution of the mean costs and mean effects of the program change generated from the probabilistic sensitivity analysis. Each individual value in the distribution represents a possible future resolution of the current uncertainty (i.e., a possible future realization under perfect information) for which the appropriate intervention can be determined, on the basis of maximum NMB. As it is not known at which particular realization the uncertainty will resolve, the expected value of a decision with perfect information is calculated by averaging these maximum NMB over the distribution. The EVPI is then simply the difference between the expected value of the decision taken with perfect information (i.e., the expectation of the maximum NMB) and that taken with current information (i.e., the maximum of the expected NMB) [25] as illustrated in equation (3).

$$E_{\theta} \max_j NMB(j, \theta) - \max_j E_{\theta} NMB(j, \theta) \quad (3)$$

where j = programs and θ = parameters.

EVPI for Parameters. In addition to determining the EVPI for the entire decision, the techniques can be applied to direct research toward the elements of the decision where the elimination of uncertainty is of most value through the calculation of the EVPI for parameters (EVPPI). Here, the expected value of the decision with current information is the same as for the calculation of EVPI, but the calculation of the expected value of the decision with perfect information is more computationally intensive. It involves two stages. The first involves determining the expected value of the decision on the basis of the remaining uncertainties and the resolved parameters, in the same way as the expected value of the decision with current information is calculated for the EVPI (i.e., on the basis of the maximum expected net benefits). However, as the actual value(s) that the parameter(s) resolve(s) at is unknown, this calculation must be undertaken for each possible realization of the uncertainty in the parameter(s) (i.e., for each possible value in the

distribution). The expected value of the decision with perfect information about the remaining uncertainties is then taken as the expectation of these values, in a process similar to the calculation of the expected value with perfect information [25]. Equation (4) illustrates the calculation of the EVPPI.

$$EVPPI = E_{\varphi} \max_j E_{\omega|\varphi} NMB(j, \theta) - \max_j E_{\theta} NMB(j, \theta) \quad (4)$$

where j = programs, θ = parameters, φ = parameter (s) of interest, and ω = other parameters.

Expected Value of Sample Information (EVSII)

Perfect information is not obtainable with a finite sample size, and the EVPI only provides a necessary but not sufficient condition for undertaking research to reduce uncertainty. Determining the worth of specific research requires a method to value the reduction in uncertainty actually achievable through research. This method is provided through the expected value of sample information (EVSII). Where the EVSI outweighs the costs associated with the research, there is a positive expected net benefit of sampling (ENMBS). In this situation, the specific research is considered worthwhile *per se* [15, 25].

The EVSI depends upon the extent to which uncertainty and the associated consequences are actually reduced by the information provided from research (the informativeness of the research) and is a function of the specifics of the research (e.g., sample size and allocation; length of follow-up; and endpoints of interest). For more details see [15, 25].

Case Study: Uncomplicated Urinary Tract Infection in Adult Women

The Model

The case study is a stylized representation of a decision involving different strategies for managing nonpregnant, adult women presenting to general practice with the symptoms of uncomplicated urinary tract infection (UTI) (frequency, dysuria, urgency, and nocturia). It has been adapted, for the purposes of presentation, from previous analyses undertaken by Fenwick *et al.* [26, 27]. The first strategy involves empiric antibiotic treatment on

presentation of symptoms. The remaining strategies involve the use of diagnostic tests to exclude or confirm the presence of UTI prior to antibiotic treatment, either through a near-patient dipstick test generating immediate results; or a laboratory test, involving an overnight urine culture. Furthermore, with the laboratory test, diagnosis can be delayed to allow for additional sensitivity testing. This delays treatment but better targets the antibiotics to the specific bacteria. In all, there are seven different strategies for managing these patients, including no treatment, which are detailed in Table 1.

Figure 3 presents the decision model. A proportion of the women who present with symptoms of UTI in general practice will have other disorders. Each patient management strategy branch is modeled by splitting the symptomatic population according to the

presence of UTI, using information from prevalence studies [28]. A lack of quantitative information of other possible causes of symptoms in this group has led to all non-UTI cases being treated identically within the model. The assumption is made that those with non-UTI will not benefit from any of the strategies considered, and the only possible health outcome for these patients is that symptoms will persist. However, as these patients are not immediately identifiable, resources are still used in the management of their symptoms and these are included within the analysis of each strategy.

Uncomplicated UTI tends to be a self-limiting condition, with 50% of cases resolving naturally after 3 days [29] with the remainder resolving after an average of a week [30]. Where no treatment is given to patients with UTI, either as a deliberate strategy

Table 1 Strategies employed within the model of UTI management^(a)

Strategy name	Strategy description
No treatment (N)	GPs provide general advice on relieving symptoms and inform patients that symptoms will resolve within 7 days.
Empiric treatment (E)	All individuals presenting with symptoms of UTI receive a 3-day course of general antibiotics.
Empiric treatment plus laboratory test (EL)	The laboratory test is used to supplement empiric treatment. While all patients provide a urine sample for testing, during the initial consultation, the results only affect the management of those patients with persistent symptoms. For these patients antibiotic sensitivity results will be available at the second visit to the GP for those who tested positive, which will enable the GP to prescribe a course of specific antibiotics. This gives the patients with UTI who test positive a second chance for treatment with antibiotics to clear the infection. Antibiotics will not be altered on the basis of the sensitivity results until the second consultation.
Dipstick and treatment (D)	The dipstick test is employed at the initial consultation to provide an indication of presence of disease and to restrict the use of antibiotics to those considered most likely to have UTI, as denoted by the result of the dipstick test.
Dipstick and treatment plus laboratory test (DL)	The laboratory test is used to supplement the dipstick test. While all patients with a positive dipstick result provide a urine sample for further testing during the initial consultation, the results only affect those patients with persistent symptoms. For these patients antibiotic sensitivity results will be available at the second visit to the GP for those who tested positive, which will enable the GP to prescribe a course of specific antibiotics. This gives the patients with UTI who test positive a second chance for treatment with antibiotics to clear the infection. Antibiotics will not be altered on the basis of the sensitivity results until the second consultation.
Laboratory test and wait for preliminary results (LW)	All patients provide a urine sample at the initial consultation and treatment is determined by the positive/negative result of this test. Hence treatment is delayed until this result is available.
Laboratory test and wait for sensitivities (LS)	All patients provide a urine sample at the initial consultation and treatment is determined by the sensitivity result of this test. Hence treatment is delayed until the results of the sensitivity analysis are available. As a result, specific antibiotics are given to every confirmed case of UTI as a first treatment, leaving no secondary course of treatment for those with persistent symptoms.

^(a) Reproduced from [26]. © Royal College of General Practitioners, 2000

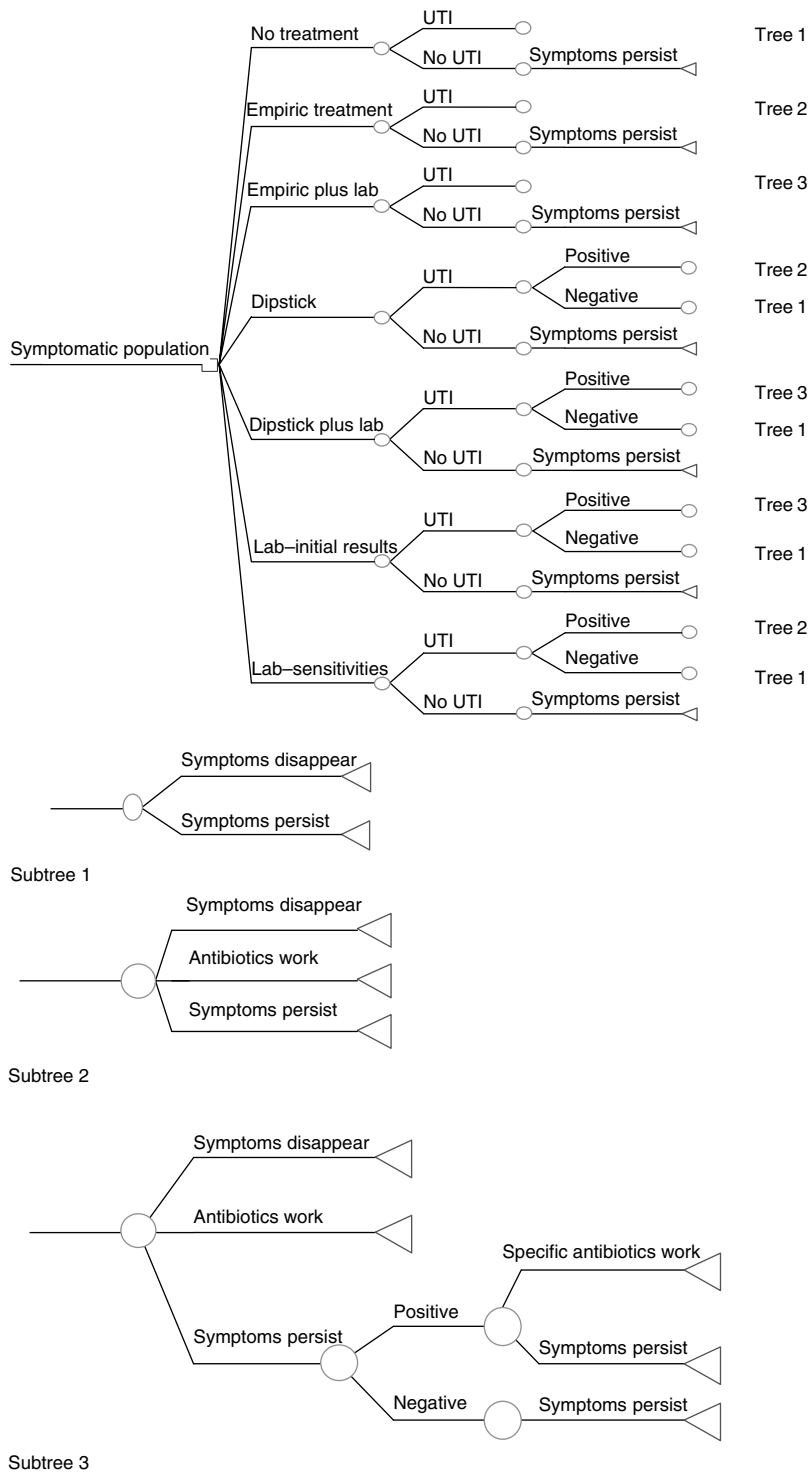


Figure 3 The patient management decision – UTI model [Reproduced from [26]. © Royal College of General Practitioners, 2000.]

or as the result of an incorrect test result, symptoms are assumed to either disappear after 3 days or to persist for 7 days (tree 1). Where UTI is the cause, the period of symptoms can be reduced through the use of antibiotics, which may resolve symptoms after 2 days from the start of the course [29]. The assumption is made that the use of antibiotics has no impact upon the probability of natural resolution, so where antibiotics are given symptoms may resolve naturally or as a result of the antibiotics, or may persist (tree 2).

When used, test results dictate the subsequent management of the patient. Treatment with antibiotics (and possibly a confirmatory laboratory culture) follows a positive result and no further treatment follows a negative one. When laboratory tests are undertaken, an initial positive/negative result can confirm the presence/absence of UTI and further analysis of positive results provides details of bacterial sensitivities that can direct prescribing. Where available, this information is used to manage patients in whom symptoms persist.

Patients given antibiotics are assumed to comply fully with the course of treatment. Eight percent are expected to experience side effects due to the antibiotics [31], which prolong the period of symptoms by an additional 2 days [32]. The assumption is made that there is no worsening of symptoms or progression to pyelonephritis due to withholding or delaying antibiotic treatment within this patient population.

This model deals with the primary management of uncomplicated UTI in women. As such, subsequent investigations in those whose symptoms persist following the completion of the management strategy are considered to be outside the scope of the model. Therefore, while patients are assumed to return to the general practitioner (GP) where symptoms persist, and the resources associated with these visits are included within the model, any further investigations are excluded from the model.

Tables 2 and 3 detail the probability distributions and data used to populate the model. Beta distributions [33] were used to represent the uncertainty concerning each of the probabilities within the model. The beta distributions were characterized using an α and β value taken from available trial data (α represents the number of patients experiencing the event and β the number not experiencing the event). For the sensitivity and specificity parameters of the dipstick, data from more than one trial was used to specify the distributions. Here the numbers of patients experiencing/not experiencing the event were added across trials to obtain the α and β values.

The event time parameters were represented as lognormal distributions, due to their positive nature and positive skew [40], characterized using a mean value and a standard error. The disutility values associated with symptoms and side effects were also represented by lognormal distributions characterized using the mean and standard error (taken as a

Table 2 Parameters for the β distributions for probability

Parameter	Base value (%)	α	β	References
Probability of UTI given symptoms	47	51	58	[28]
Sensitivity of dipstick	92	185	16	Combined data from [34, 35]
Specificity of dipstick	88	640	91	Combined data from [35, 36]
Sensitivity of laboratory test	99	104	1	Assumption
Specificity of laboratory test	99	104	1	Assumption
Probability that symptoms will resolve naturally given UTI	50	40	40	[29]
Probability that antibiotics will resolve symptoms given UTI	91	88	9	[37]
Probability that specific antibiotics will resolve symptoms given UTI	92	69	6	[38]
Probability of side effects due to antibiotic treatment	9.2	17	211	[31]
Proportion of patients suffering with continued symptoms who return to the GP when no treatment was given at initial visit	5	5	95	Assumption
Proportion of patients suffering with continued symptoms who return to the GP when treatment was given at initial visit	70	7	3	Assumption

Table 3 Parameters for the lognormal distributions for event times and disutility

Parameter	Base value	SE	References
Period of UTI infection	7	2	–
Time for antibiotics to work	2	0.6	–
Time for infection to clear up naturally	3	0.6	–
Time for laboratory test results	2	0.4	–
Time for sensitivity tests to be available	1	0.4	–
Duration of side effects	2	0.6	–
Disutility associated with persistent dysuria	0.2894	0.0724	[39]
Disutility associated with side effects	0.2894	0.0724	[39]

quarter of the mean value). Unit costs were taken from published sources including the British National Formulary and Unit Costs of Health and Social Care. Probability distributions were not assigned to the unit costs as these were considered to be variable rather than uncertain.

Results

On the basis of expected values, the least costly strategy is no treatment (N), generating an expected 0.99513 QALYs per episode of UTI at an expected cost of £8.93. The next least costly strategy involves testing with a dipstick followed by empiric treatment for positive results; this provides an expected 0.99601 QALYs per episode of UTI (an increase of 0.00088 QALYs) at a cost of £12.40 (an increase of £3.47). Thus, the dipstick strategy (D) involves an additional cost of £3492 per additional QALY (ICER). The next least costly strategy involves empiric treatment for all patients; this provides an expected 0.99603 QALYs per episode of UTI (an increase of 0.00002 QALYs) at a cost of £12.50 (an increase of £0.10). Thus, the empiric strategy (E) involves an additional cost of £4438 per additional QALY (ICER). Finally, the empiric plus laboratory (EL) strategy provides an expected 0.99606 QALYs per episode of UTI (an increase of 0.00003 QALYs) at a cost of £15.22 (an increase of £2.73). Thus, the EL strategy involves an additional cost of £14 500 per additional QALY (ICER). The two strategies employing laboratory testing prior to treatment both generate fewer QALYs at a higher cost than empiric treatment and are dominated. The strategy with a confirmatory laboratory test following a dipstick test (DL) is more effective than the empiric strategy but has a higher ICER than the EL strategy and is extended dominated. The

choice between the patient management strategies will depend crucially upon the value that the policy maker is willing to pay for additional QALYs in this patient group.

Figure 4 presents the cost–effectiveness plane for the seven management strategies illustrating the expected values for costs and effects. The uncertainty in the estimates is not shown on this figure as the cost–effectiveness plane becomes unmanageable and indecipherable with this number of interventions. Figure 5 illustrates the CEACs for the management strategies, although only five are visible over this range of values of λ . The figure shows that when the policy maker was unwilling to pay anything for an additional QALY the probability that no treatment was optimal is 1. If the policy maker was willing to pay £20 000 per QALY gained, empiric treatment (E) is considered cost effective on the basis of expected values. The probability that empiric treatment was cost effective at this value of the cost–effectiveness threshold is 0.573. These probabilities for no treatment (N), EL, and the dipstick strategy (D) are 0.016, 0.012, and 0.383 respectively. If the policy maker was willing to pay £30 000 per QALY gained, the probability that empiric treatment was cost effective is 0.544, with those for the no treatment (N), EL, and the dipstick strategy (D) being 0.009, 0.045, and 0.347 respectively. This indicates that there is considerable uncertainty surrounding the decision between the methods of patient management. For a cost–effectiveness threshold value of £20 000 per additional life year, the uncertainty (i.e., probability of making the wrong decision) associated with the choice of empiric treatment (E) is 0.427. This is much higher than would be acceptable by standard conventions of significance.

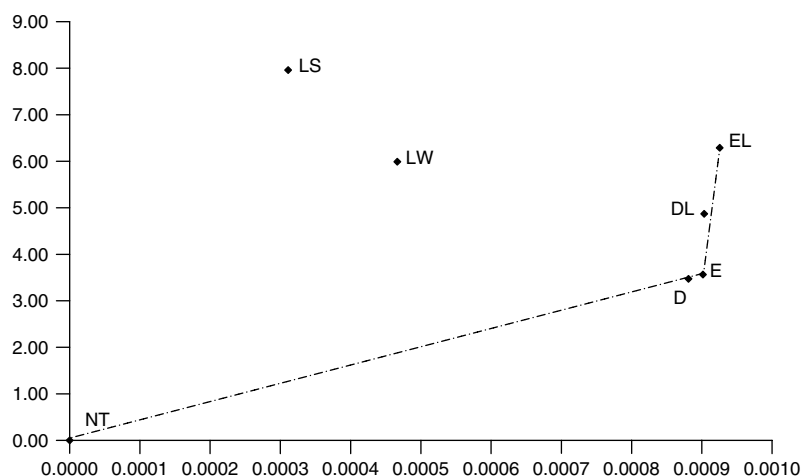


Figure 4 Cost-effectiveness plane – UTI model

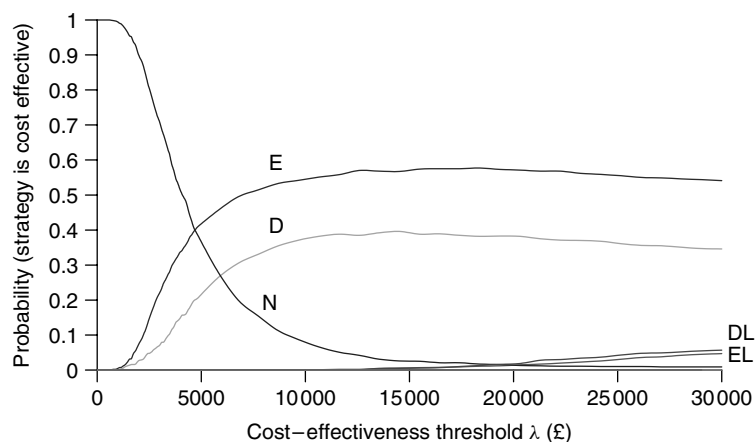


Figure 5 Cost-effectiveness acceptability curves – UTI model

The EVPI was calculated to be £0.30 per case of UTI, given a cost-effectiveness threshold of £20 000 per QALY, or £0.46 per episode for a cost-effectiveness threshold of £30 000 per QALY. These values translated into a population EVPI of £2.3 or £3.6 million respectively (see Figure 6). These values provide absolute limits on the worth of further research concerning all elements of the decision, for these values for the cost-effectiveness threshold. Thus, further research costing less than £2.3 million (£3.6 million) is potentially worthwhile, assuming that the cost-effectiveness threshold is £20 000 (£30 000) per QALY. A partial EVPI analysis

illustrated that further research should concentrate upon getting better estimates of disutility values, rather than prevalence data for UTI or side effects. Given that the problem of selection bias is not expected to be important in the estimation of disutility parameters, it would not be necessary to undertake a clinical trial.

Conclusion

This article has introduced economic evaluation for health care decision making and shown how decision analysis can be a useful tool for economic evaluation.

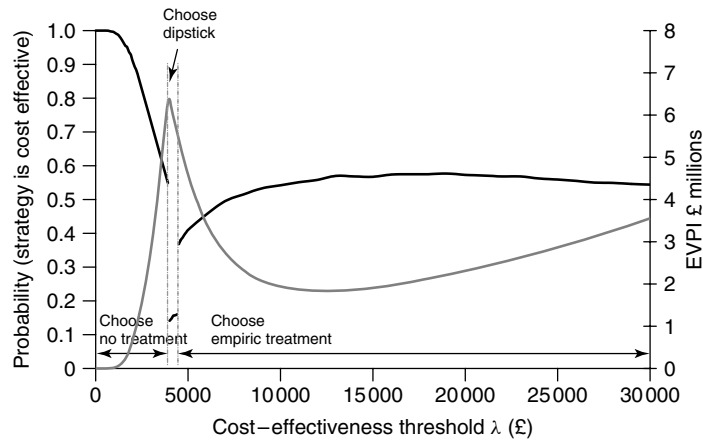


Figure 6 Expected value of perfect information for the decision (UK population) versus the uncertainty cost-effectiveness acceptability frontier (CEAF)

Within the case study, the techniques provided the cost effectiveness and uncertainty associated with each of the various methods of patient management. The EVPI analysis illustrated that there was worth in conducting further research, in particular, with respect to disutilities. Given that the problem of selection bias is not expected to be an important issue in the estimation of disutility parameters, it would not be necessary to undertake a clinical trial. As such, a decision model incorporating VOI analysis could be employed to improve the efficiency of health technology assessment. However, in order to use these techniques appropriately, it is important that the decision model captures and incorporates all of the available evidence, rather than relying on the best estimate. In addition, it must be remembered that any probabilistic analysis and/or VOI analysis is only as good as the underlying model on which it is based.

References

- [1] Bateman, I.J., Carson, R.T., Day, B., Hanemann, M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Özdemiroglu, E., Pearce, D.W., Sugden, R. & Swanson, J. (2002). *Economic Valuation with Stated Preference Techniques: A Manual*, Edward Elgar, Cheltenham.
- [2] Sugden, R. & Williams, A. (1978). *The Principles of Practical Cost-Benefit Analysis*, Oxford University Press, New York.
- [3] Clarke, P. (1998). Cost-benefit analysis and mammographic screening: a travel cost approach, *Journal of Health Economics* **17**(6), 767–787.
- [4] Clarke, P. (2000). Valuing the benefits of mobile mammographic screening units using the contingent valuation method, *Applied Economics* **32**, 1647–1655.
- [5] Haefeli, M., Elfering, A., McIntosh, E., Gray, A., Sukthankar, A. & Boos, N. (Forthcoming). A cost benefit analysis using contingent valuation techniques: a feasibility study in spinal surgery, *Value in Health*.
- [6] McIntosh, E. (2006). Using stated preference discrete choice experiments in cost-benefit analysis: some considerations, *Pharmacoeconomics* **24**(9), 855–869.
- [7] Warner, K. & Luce, B.R. (1982). *Cost-benefit and Cost-effectiveness Analysis in Health Care: Principles, Practice and Potential*, Health Administration Press, Ann Arbor.
- [8] Drummond, M.F., O'Brien, B.J., Stoddart, G.L. & Torrance, G.W. (1997). *Methods for the Economic Evaluation of Health Care Programmes*, 2nd Edition, Oxford University Press, New York.
- [9] Gold, M.R., Seigel, J.E., Russell, L.B. & Weinstein, M.C. (1997). Cost-effectiveness in health and medicine, in *Cost-effectiveness in Health and Medicine*, M.R. Gold, J.E. Seigel, L.B. Russell & M.C. Weinstein, eds, Oxford University Press, New York.
- [10] Weinstein, M.C. & Stason, W.B. (1977). Foundations of cost-effectiveness analysis for health and medical practices, *New England Journal of Medicine* **296**(13), 716–721.
- [11] Black, W.C. (1990). The CE plane: a graphic representation of cost-effectiveness, *Medical Decision Making* **10**, 212–214.
- [12] Karlsson, G. & Johannesson, M. (1996). The decision rules of cost-effectiveness analysis, *Pharmacoeconomics* **9**, 113–120.
- [13] Johannesson, M. & Meltzer, D. (1998). Some reflections on cost-effectiveness analysis, *Health Economics* **7**, 1–7.

- [14] Claxton, K. (1999). The irrelevance of inference: a decision making approach to the stochastic evaluation of health care technologies, *Journal of Health Economics* **18**, 341–364.
- [15] Claxton, K. & Posnett, J. (1996). An economic approach to clinical trial design and research priority setting, *Health Economics* **5**(6), 513–524.
- [16] Phelps, C.E. & Mushlin, A.I. (1988). Focusing technology assessment using medical decision theory, *Medical Decision Making* **8**(4), 279–289.
- [17] Stinnett, A.A. & Mullahy, J. (1998). Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making* **18**(Suppl. 2), S68–S80.
- [18] Tambour, M., Zethraeus, N. & Johannesson, M. (1998). A note of confidence intervals in cost-effectiveness analysis, *International Journal of Technology Assessment in Health Care* **14**(3), 467–471.
- [19] Fenwick, E., Claxton, K. & Sculpher, M. (2001). Representing uncertainty: the role of cost-effectiveness acceptability curves, *Health Economics* **10**, 779–787.
- [20] Fenwick, E., O'Brien, B. & Briggs, A. (2004). Cost-effectiveness acceptability curves: facts, fallacies and frequently asked questions, *Health Economics* **13**, 405–415.
- [21] Pratt, J.W., Raiffa, H. & Schlaifer, R. (1994). Introduction to statistical decision theory, in *Introduction to statistical decision theory*, J.W. Pratt, H. Raiffa & R. Schlaifer, eds, MIT press, Cambridge.
- [22] Raiffa, H. (1968). *Decision Analysis: Introductory Lectures on Choices under Uncertainty*, Addison-Wesley Publishers, New York.
- [23] Raiffa, H. & Schlaifer, R.O. (1959). *Probability and Statistics for Business Decisions*, McGraw-Hill, New York.
- [24] Raiffa, H. & Schlaifer, R.O. (1961). *Applied Statistical Decision Theory*, Harvard University Press, Cambridge.
- [25] Ades, A., Lu, G. & Claxton, K. (2004). Expected value of sample information calculations in medical decision modeling, *Medical Decision Making* **24**(2), 207–227.
- [26] Fenwick, E., Briggs, A. & Hawke, C. (2000). Management of urinary tract infection in general practice: a cost-effectiveness analysis, *British Journal of General Practice* **50**, 635–639.
- [27] Fenwick, E., Claxton, K., Sculpher, M. & Briggs, A. (2000). *Improving the Efficiency and Relevance of Health Technology Assessment: The Role of Iterative Decision Analytic Modelling*, CHE Discussion Paper 179.
- [28] Baerheim, A., Digraanes, A. & Hunskaar, S. (1999). Equal symptomatic outcome after antibacterial treatment of acute lower urinary tract infection and the acute urethral syndrome in adult women, *Scandinavian Journal of Primary Health Care* **17**, 170–173.
- [29] Brumfitt, W. & Hamilton-Miller, J.M.T. (1987). The appropriate use of diagnostic services: (xii) Investigation of urinary infections in general practice: Are we wasting facilities? *Health Bulletin* **45**, 5–10.
- [30] Medicines Resource Centre (1995). Urinary tract infection, *MeReC Bulletin* **6**, 29–32.
- [31] McCarty, J.M., Richard, G., Huck, W., Tucker, R.M., Tosiello, R.L., Shan, M., Heyd, A. & Echols, R.M. (1999). A randomized trial of short-course ciprofloxacin, ofloxacin, or trimethoprim/sulfamethoxazole for the treatment of acute urinary tract infection in women, *The American Journal of Medicine* **106**, 292–299.
- [32] Carlson, K.J. & Mulley, A.G. (1985). Management of acute dysuria. A decision-analysis model of alternative strategies, *Annals of Internal Medicine* **102**, 244–249.
- [33] Berry, D.A. & Stangl, D.K. (1996). Bayesian biostatistics, in *Bayesian Biostatistics*, D.A. Berry & D.K. Stangl, eds, Marcel Dekker, New York.
- [34] Ditchburn, R.K. & Ditchburn, J.S. (1990). A study of microscopical and chemical tests for the rapid diagnosis of urinary tract infections in general practice, *British Journal of General Practice* **40**(339), 406–408.
- [35] Pallarés, J., Casas, J., Guarga, A., Marguet, R., Grifell, E., Juvé, R. & Catellort, R. (1988). Metodos de diagnostico rapido como predictores de infeccion urinaria, *Atencion Primaria* **91**(20), 775–778.
- [36] Hallander, H., Kallner, A., Lundin, A. & Osterberg, E. (1986). Evaluation of rapid methods for the detection of bacteruria, *Acta Pathologica, Microbiologica et Immunologica Scandinavica* **94**, 39–49.
- [37] Trienekens, T.A.M., Stobberingh, E.E., Winkens, R.A.G. & Houben, A.W. (1989). Different lengths of treatment with co-trimoxazole for acute uncomplicated urinary tract infections in women, *British Medical Journal* **299**, 1319–1322.
- [38] Hooton, T.M., Winter, C., Tiu, F. & Stamm, W.E. (1995). Randomised comparative trial and cost analysis of 3-day antimicrobial regimens for treatment of acute cystitis in women, *The Journal of the American Medical Association* **273**(1), 41–45.
- [39] Barry, H.C., Ebell, M.H. & Hickner, J. (1997). Evaluation of suspected urinary tract infection in ambulatory women: a cost-utility analysis of office-based strategies, *Journal of Family Practice* **44**(1), 49–60.
- [40] Doubilet, P., Begg, C.B., Weinstein, M.C., Braun, P. & McNeil, B.J. (1985). Probabilistic sensitivity analysis using Monte Carlo simulations, *Medical Decision Making* **5**, 157–177.

Related Articles

Efficacy

Randomized Controlled Trials

Risk–Benefit Analysis for Environmental Applications

ELISABETH A.L. FENWICK AND ANDREW H. BRIGGS

Counterterrorism

No one who has seen them can forget the video images of airplanes hitting the twin towers of the World Trade Center on September 11, 2001. Terrorist acts, even those directed at or taking place within the United States, had occurred previously, but worldwide a new sense of vulnerability has arisen. Arguably, since 2001 there has been little diminution of terrorist actions: examples include the 2004 train bombings in Madrid, the 2005 attacks on the London underground, and endemic violence in the Middle East.

Counterterrorism (see **Managing Infrastructure Reliability, Safety, and Security; Game Theoretic Methods; Sampling and Inspection for Monitoring Threats to Homeland Security**) is the process of forestalling or mitigating the consequences of terrorist acts by acquiring advance knowledge of them, intervening before they can take place, and minimizing human, economic, and other impacts should they occur. Because so much of the process is conducted in secret by governments, it is not possible in this encyclopedia to provide details. Instead, this article outlines some of the dimensions and complexities of countering terrorist plots, and discusses how terrorist risk relates to other risks. Counterterrorism is one of many settings in which the realities of risk and the perception of risk may be inconsistent. In the United States and elsewhere, broad-based counterterrorism strategies address both reality and perception.

What is Counterterrorism?

Defining terrorist acts (see **Human Reliability Assessment**) may be impossible, but most people would identify terrorist acts to have some or all of the following characteristics:

- They take place (spatially and temporally) outside of “ordinary” military conflicts, and may be committed for nonmilitary (for example, religious or political) reasons.
- They target civilians rather than the uniformed military.
- They are committed by small groups or individuals who are often willing to die in the process.

- Overt or implicit support by a legitimate national state is not necessary, but may be present.
- They are, by design, unpredictable.
- They are specifically meant to inspire consequences such as fear and economic or political disruption.

As compared with ordinary military conflicts, terrorist acts are complex, heterogeneous, and geographically dispersed to point that they may be seen (and may be meant to be seen) as essentially random. Pan Am Flight 103 exploded over Lockerbie, Scotland, in 1988 as the result of a (state-sponsored) terrorist act. Four days previous to that, this author was on a TWA flight from Frankfurt to the United States from which luggage had to be removed because its owners had not boarded the flight. It is impossible not to ask “What if?” As suggested in the introductory text, counterterrorism [1, 2] can be thought of as having the following five principal components:

Information

Advance knowledge is the only basis for preventing terrorist acts, at least within our current legal framework, which does not allow action without information against those who “might” commit terrorist acts. Significant parts of the process and the results are classified, but the wealth of information turned up *ex post facto* about the 9/11/01 airplane hijackers leads to tantalizing feelings that the answer is also there beforehand, if only we could find the “needle in the haystack”.

Intervention

Given sufficient, and sufficiently credible, information that a terrorist act may occur, a government can intervene to prevent it. This author does not know the specifics of intervention strategies. It appears that, depending on situational factors, a successful intervention may or may not be made public.

Deterrence

Every counterterrorism strategy (indeed, it seems any strategy to combat any unlawful behavior) must involve deterrence mechanisms. These mechanisms may operate well before the fact (“We are watching you”), at the time and place of an intended act (physical barriers, searches of air passengers’ possessions), and after the fact (“You will be identified, tracked down, and punished”).

Mitigation

It is not prudent to believe that deterrence, information, and intervention will always succeed. Counterterrorism strategies also include steps to mitigate effects should a terrorist act occur. These range from the physical (building escape routes, isolation strategies for infected animal herds) to the operational (intercommunication among first responders) and beyond.

Control of consequences

Because terrorist acts are meant to have consequences beyond the immediate, these must be addressed as well. For instance, introduction of foot-and-mouth disease into cattle in the United States might, given effective interventions, lead to the isolation or destruction of a few hundred animals. Nevertheless, because of fear (In fact, foot-and-mouth disease is quite harmless to healthy humans.), an ensuing economic dislocation on the order of \$1 billion or more could occur.

One uniquely complicated consequence – risk perception – is discussed in the section titled “Perception *versus* Reality”.

Perception *versus* Reality

A disclaimer: what follows is in no sense meant to downplay the horrific events of 9/11/01, but only to illustrate the complexity of the issues. Terrorists have succeeded in their objective to instill fear, with resultant economic and political dislocations, in significant part because of public perception of risk [3].

Approximately 3000 people died in the attacks on 9/11/01. That same month, 3555 people died in automobile crashes in the United States (and 42 196 for the entire year of 2001). In 2001, 29 573 people in the United States died from gunshot wounds. Less than 1/10th as many people died in the United States in 2001 from anthrax attacks than from being hit by lightning. Yet, reaction to perceived terrorist risks far exceeds that to far greater (and more widespread) other risks. One (seemingly easily) thwarted attempt to detonate an explosive device in a shoe led to immense cost, lost time, and inconvenience from people removing their shoes in airports. Nor is there public evidence that this precaution has prevented any further attacks.

Risk perception is an enormously complex phenomenon, of which inconsistency may be the only consistent characteristic. To illustrate, driving is more dangerous than flying by 2 orders of magnitude in terms of population-level death rates, yet many more people fear the latter. Factors that feed such feelings include control (“I am in control of my car, so I can avoid accidents”), severity/certainty of consequences (“Everyone is killed in a plane crash, but many people survive automobile accidents”), and media visibility (All airplane crashes make the television news, but only the worst automobile accidents do), as well as the practical (“I can’t not drive, so why worry, and in any case I drive carefully”). All of these, and others, notably unpredictability, apply in terrorist settings.

Because risk perception is inherently inconsistent or even palpably irrational, counterterrorism strategies cannot ignore it. Perhaps removing shoes at the airport reduces fear more than it prevents terrorist attacks, but it is still legitimately part of a counterterrorism strategy. But, perception also adds to the already incredible complexity of the problems. Returning to an earlier illustration, automobile fatalities in United States in September, 2001 increased as compared to previous years (although not enough to vitiate the point of the second paragraph of this section) because of fear of flying resulting from the 9/11 attacks (and the impossibility of flying for several days thereafter). Increased perceived risk of one mode of transportation increased the actual risk of another.

Other Risks

The political debate and public conversation about risk and counterterrorism are dominated by the risk of occurrence of terrorist acts. However, consistent with the nature of the problem, there are multiple other risks, some of which are controversial.

While they may be impossible to quantify, risks associated with *false positives* – believing and acting on the belief that a person is a terrorist who in fact is not – exist and are unavoidable. Some researchers fear that analytical processes (one example: data mining of massive, disparate databases) that are powerful enough to detect terrorist attacks beforehand will be overwhelmed by false positives. (When you do not know what a needle looks like, how do you find one in a mile-high haystack?) In the United

States, the public has tolerated the hitherto fairly small numbers of false positives. If the number of false positives increases dramatically, that tolerance will diminish.

The information assembly and analysis aspect of counterterrorism carries other risks. Many people, across a political spectrum, believe that there is real – there surely is perceived – risk from the government’s access to intimate personal details (*see* **Coding: Statistical Data Masking Techniques**) (where one travels, what one buys, with whom one associates, what books one reads) in the process of searching for terrorists. Some counterterrorist initiatives, notably the Total Information Awareness project of the Defense Advanced Research Projects Agency, have been halted because of public outcry over these risks. Not surprisingly, exactly what the risks are and how big they are is more elusive. The major perceived risks are generic loss of privacy and the possibility that information will be used for purposes other than counterterrorism. How well-founded the fears are now, let alone in the future, is impossible to say, in large part because of lack of experience and data.

Counterterrorism as Risk Analysis

Whether counterterrorism is amenable to extant or foreseeable methods for quantitative, or even qualitative, risk analysis is a difficult question.

Some factors suggest that it is not. One example was discussed in the section titled “Perception *versus* Reality” – factors associated with perception. Perceived risks, in general, defy analysis. But might the risk of terrorist acts, as opposed to the consequences of those acts, be assessed independently of perception? Issues then arise that range from the conceptual to the practical. How (that is, using what data and what models) can we assess the risk of something that has never occurred, and which we may not even be able to define? There is strong suspicion on the part of some people that while it might be possible to characterize and prevent recurrences of acts that have been committed, this is missing the point. Terrorist organizations are not bureaucratic behemoths, and if one path of attack (physical entry into aircraft cockpits) is precluded, another (subverting pilots or baggage handlers or aviation fuel suppliers) will be found. No one could reasonably have estimated the

probability of anthrax being distributed in the mail in 2001, and no one may be able to do so now. “Impossible” may be the most defensible estimate.

Simply knowing that terrorist attacks have occurred may also become an issue. Because of the “fear” objective discussed in the section titled “What is Counterterrorism?”, terrorists often proclaim publicly – and sometimes falsely – responsibility for their acts. At some point uncertainty (Was yesterday’s collapse of the Internet a hardware failure or a terrorist act?) may become the behavior of choice. Were this to happen, the already significant paucity of reliable data may become insuperable.

Indeed, data availability and quality are likely to remain impediments for the foreseeable future to research in risk analysis as it relates to counterterrorism. It is, of course, not possible to make classified information available to the public, but in the absence of data, many analyses are speculative and most lack scientific verification.

Solid evidence is lacking, but there is indication in the United States of unwillingness to balance risks and benefits in a counterterrorism setting. Public (and political) attitudes about risk and benefit already vary wildly. The risk of being one of over 40 000 automobile accident victims is balanced in Congress and elsewhere against the benefits of using automobiles, and there is understanding that the cost of some solutions (e.g., making cars into 1 mile/gallon tanks that are impervious to crashes) exceed the benefits. On the other hand, the public stands willing for drugs that may be beneficial to millions of people to be removed from market in light of a handful of adverse incidents, not all of which can be linked causally to use of the drug.

Counterterrorism currently seems to lie closer to the latter. As benefits the most noble aspirations of a society, individuals have been willing to accept (not insignificant) personal costs in return for social benefits. Whether this will change if terrorism were no longer in the public eye remains to be seen. Currently, there appears to be little public will to deal in a principled manner with false positives, but this also may change.

References

- [1] Howard, R.D. & Sawyer, R.L. (2005). *Terrorism and Counterterrorism: Understanding the New Security Environment, Readings and Interpretations*, 2nd Edition, McGraw-Hill, New York.

- [2] Wilson, A.G., Olwell, D.H. & Wilson, G. (eds) (2006). *Statistical Models in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance and Biometric Authentication*, Springer-Verlag, New York.
- [3] Slovic, P. (2000). *The Reception of Risk*, EarthScan/James & James, London.

ALAN F. KARR

Credibility Theory

Credibility theory (*see* **Insurance Pricing/Nonlife**) refers to a body of techniques actuaries use to assign premiums to individual policyholders in a heterogeneous portfolio. These techniques are applicable in fields other than insurance wherever one needs to distinguish between individual and group effects in a set of nonidentically distributed data. As such, credibility theory is sometimes defined shortly as “the mathematics of heterogeneity”.

With origins dating back to 1914, credibility theory is widely recognized as the cornerstone of casualty actuarial science (*see* **Dependent Insurance Risks; Reinsurance; Securitization/Life**). It is divided in two main approaches, each representing a different manner of incorporating individual experience in the ratemaking process (*see* **Nonlife Loss Reserving**). The oldest approach is limited fluctuations credibility, where the premium of a policyholder is based solely on its own experience, provided its experience is stable enough to be considered fully credible. The theory behind limited fluctuations credibility is rather simple and the technique has few correct applications, so it will be reviewed only briefly in this article.

The second approach concentrates on the homogeneity of the portfolio rather than the stability of the experience to determine the best premium for a policyholder. In greatest accuracy credibility, individual experience is taken into consideration only if it is significantly different from that of the portfolio. The more heterogeneous the portfolio, the more important becomes individual experience, and *vice versa*.

Experience Rating

The first concern of an insurer – whether a public or a private corporation – when building a tariff is to charge enough premiums to fulfill its obligations. For various reasons, including competitiveness, the insurer may then seek to distribute premiums fairly between policyholders. A classification structure (for example: according to age, sex, type of car, etc.) is usually the first stage in premium distribution. Experience-rating (*see* **Risk Classification in Nonlife Insurance**) systems, in general, and credibility methods, in particular, then constitute an efficient second stage.

As the name suggests, an experience-rating system takes the individual experience of a policyholder into account to determine its premium. Such systems require the accumulation of a significant volume of experience. Experience rating is thus especially used in workers compensation and automobile insurance.

On a more formal basis, Bühlmann [1] proposes the following definition:

Experience rating aims at assigning to each individual risk its own *correct premium* (rate). The *correct premium* for any period depends exclusively on the (unknown) claims distribution of the individual risk for this same period.

Example 1 Consider this simplified example (more details are presented in [2] or [3]): a portfolio is composed of ten policyholders and each policyholder can incur at most one claim of amount 1 per year. Before observing any experience for this portfolio, the insurer considers the policyholders equivalent on a risk level basis. Expecting that, on average, the portfolio will incur two claims per year, the insurer charges an identical premium of 0.20 to each policyholder. This is the collective premium.

The experience for this portfolio after 10 years is shown in Table 1. The insurer observed a total of 23 claims, for an average cost per policyholder of $23/100 = 0.23$. However, these 23 claims are not uniformly distributed among policyholders: policyholders 7, 8, and 10 incurred no claims, while policyholder 9 alone incurred seven.

Table 1 Experience of the simplified portfolio of Example 1 after 10 years

[illegible]

Therefore, the collective premium in this example is globally adequate, but clearly not fair. To avoid antiselection, the insurer should charge more to some policyholders and less to others. Experience showed that the portfolio is, to some degree, heterogeneous.

One could very well restate this example in a context other than insurance. The data of Table 1 merely establish whether an “event” occurred or not in 10 “trials” and for 10 different “subjects”. The goal is then to estimate the probability of occurrence for trial 11 for each subject, knowing that the subjects are not “equivalent”.

There exist numerous experience-rating systems, including bonus–malus and merit–demerit systems, participating policies or commissions in reinsurance [1, 4, 5], but the most widely used methods are credibility models.

Limited Fluctuations Credibility

Limited fluctuations credibility originated in the early 1900s with Mowbray’s paper “How Extensive A Payroll Exposure Is Necessary To Give A Dependable Pure Premium?” As the title suggests, Mowbray was interested in finding a level of payroll in workers compensation insurance for which the pure premium of a given employer would be considered fully dependable or, in other words, fully credible.

We may believe that this question arose from one or a few large employers requesting from their insurer to pay a premium better tailored to their own experience rather than to the experience of their group. The employers would rightfully argue that given their large size and by virtue of the law of large numbers their experience is fairly stable in time, thus allowing the insurer to compute more accurate and fair pure premiums.

Mowbray [6] defines a dependable pure premium to be “one for which the probability is high that it does not differ from the [true] pure premium by more than an arbitrary limit”. Translated in modern mathematical terms, this is essentially asking that

$$\Pr[(1 - k)E[S] \leq S \leq (1 + k)E[S]] \geq p \quad (1)$$

for a small value of k and a large probability p . Here, the random variable S represents the experience of a policyholder, in one form or another. We detail three typical examples below.

1. Mowbray defined S as the number of accidents of an employer, assuming a binomial distribution with parameters n , the number of employees, and θ , the known probability of accident. Solving equation (1) for n , using the central limit theorem yields

$$n \geq \left(\frac{\zeta_{(1-p)/2}}{k} \right)^2 \frac{1 - \theta}{\theta} \quad (2)$$

where ζ_α is the $100(1 - \alpha)$ th percentile of the standard normal distribution.

2. The size of a policyholder may be defined as the expected number of claims in a given period (typically, 1 year). For such cases, one usually defines S as the total amount of claims in the period and further assumes that the distribution of S is a compound Poisson. Solving equation (1) for the Poisson parameter λ yields

$$\lambda \geq \left(\frac{\zeta_{(1-p)/2}}{k} \right)^2 \left(1 + \frac{\text{var}[X]}{E[X]^2} \right) \quad (3)$$

where X is the random variable of claim amounts.

3. The two examples above determine a policyholder’s admissibility to full credibility one period at a time. An alternative criterion is the number of periods of experience. This requires to set S in equation (1) as the average total amount of claims after n years:

$$S = \frac{S_1 + S_2 + \cdots + S_n}{n} \quad (4)$$

where $S_1, \dots, S_n$ are independent and identically distributed random variables of the total amount of claims per period. Then the experience of a policyholder is considered fully credible after

$$n \geq \left(\frac{\zeta_{(1-p)/2}}{k} \right)^2 \frac{\text{var}[S_i]}{E[S_i]^2} \quad (5)$$

periods of experience.

The interested reader will find more examples in [7].

A policyholder’s experience is thus considered fully credible if it fluctuates moderately from one period to another. That is, the credibility criterion is stability. This stability of the experience usually increases with the volume of the policyholder, whether it is expressed in premium volume, number

of claims, number of employees, number of years of experience, or any other exposition base.

Truly appropriate applications of Mowbray's procedure are rare. In insurance, it should be reserved for applications where the stability of the experience is of foremost importance. One good example is the determination of an admissibility threshold to a retrospective insurance system, where the policyholder's premium is readjusted at the end of the year after the total claim amount is known.

To this day, limited fluctuations remains the credibility procedure most widely used by American actuaries. It has been extended in various and often *ad hoc* ways. One such extension deals with the case of a policyholder only partially reaching the full credibility criterion. As proposed originally by Whitney in his seminal paper of 1918 [8], one charges a pure premium π that is a weighted average of the individual experience S and the collective premium m :

$$\pi = zS + (1 - z)m \quad (6)$$

Here, $0 \leq z \leq 1$ is the so-called credibility factor.

Over the years, many partial credibility formulas have been proposed. Among the most widely used are the following:

$$z = \min \left\{ \sqrt{\frac{n}{n_0}}, 1 \right\} \quad (7)$$

$$z = \min \left\{ \left(\frac{n}{n_0} \right)^{2/3}, 1 \right\} \quad (8)$$

and

$$z = \frac{n}{n + K} \quad (9)$$

where n_0 is the full credibility level and K is a constant determined upon judgment, usually to limit the size of the premium change from one year to the next. The third formula is the one introduced by Whitney and the only one in which the (partial) credibility level never reaches unity.

It should be emphasized that in such a context, partial credibility does not seek to find the most accurate premium for a policyholder. The goal is rather to incorporate in the premium as much individual experience as possible while still keeping the premium sufficiently stable. When credibility is used to find the

best estimate of a policyholder's pure risk premium, one should turn toward greatest accuracy methods.

Greatest Accuracy Credibility

The greatest accuracy approach to credibility theory seeks to find the "best" (in a sense yet to be defined) premium to charge a policyholder. This is achieved by distributing the collective premium in an optimal way among the members of a group.

It is now well recognized that the first paper on greatest accuracy credibility was the aforementioned 1918 paper by Whitney [8]. Whitney developed formula (6) out of "the necessity, from the standpoint of equity to the individual risk, of striking a balance between class-experience on the one hand and risk-experience on the other." How this balance is calculated does not depend solely on the stability of the experience, like in the limited fluctuations approach, but rather on the homogeneity of the portfolio. Indeed, Whitney writes:

There would be no experience-rating problem if every risk within the class were typical of the class, for in that case the diversity in the experience would be purely adventitious.

However, Whitney was ahead of his time and his ideas were either not well received or not well understood. For example, he was criticized for using an early version of the Bayes rule; see [9]. For decades, actuaries essentially remembered and used formula (6) and the form $z = n/(n + K)$ for the credibility factor.

The greatest accuracy approach lay dormant until the publication of Arthur L. Bailey's papers [10, 11]. The second paper is an especially enlightening exposition of the state of statistical and actuarial practices at the time. In particular, with respect to credibility theory, Bailey writes:

The trained statistician cries "Absurd! Directly contrary to any of the accepted theories of statistical estimation." The actuaries themselves have to admit that they have gone beyond anything that has been proven mathematically, that all of the values involved are still selected on the basis of judgment, and that the only demonstration they can make is that, in actual practice, it works.

Bailey then poses himself as an advocate of the then controversial Bayesian philosophy (*see Natural*

Resource Management; Risk in Credit Granting and Lending Decisions; Credit Scoring; Reliability Demonstration; Cross-Species Extrapolation), arguing that the notion of prior opinion is natural in actuarial work:

At present, practically all methods of statistical estimation appearing in textbooks on statistical methods or taught in American universities are based on an equivalent to the assumption that any and all collateral information or *a priori* knowledge is worthless. . . . Philosophers have recently discussed the credibilities to be given to various elements of knowledge [12], thus undermining the accepted philosophy of the statisticians. However, it appears to be only in the actuarial field that there has been an organized revolt against discarding all prior knowledge when an estimate is to be made using newly acquired data.

Bailey then shows that the Bayesian estimator of the policyholder's true risk premium obtained by minimizing the mean square error is a credibility premium of the form of equation (6) for certain combinations of distributions. The credibility factor is still of the form $z = n/(n + K)$, where K depends on the parameters of the model.

Modeling Heterogeneity

The classical mathematical model of greatest accuracy credibility was formalized by Bühlmann [1, 5]. Consider a heterogeneous portfolio (group) of I policyholders. The risk level of policyholder $i = 1, \dots, I$ (whether it is a good or a bad driver, for example) is unknown, but past claims data $S_{i1}, \dots, S_{in}$ are available for ratemaking purposes.

We make the following assumptions:

1. Claim amounts of policyholder i are (conditionally) independent and identically distributed with cumulative distribution function (cdf) $F(x|\theta_i)$. Parameter θ_i is the realization of a random variable Θ_i .
2. Random variables $\Theta_1, \dots, \Theta_I$ are identically distributed with cdf $U(\theta)$.
3. The policyholders are independent, meaning that the claims record of one policyholder has no impact on the claims record of another.

The random variable Θ_i represents the risk level of policyholder i . This is an abstract and nonobservable random variable – otherwise, the ratemaking problem

would be easily solved. The portfolio is heterogeneous since each policyholder has its own risk level. Yet, the identical distribution assumption for the random variables $\Theta_1, \dots, \Theta_I$ means that the risk levels all come from the same process. Hence the policyholders are similar enough to justify grouping them in the same portfolio.

In a pure Bayes setting, $U(\theta)$ represents the prior distribution of the risk levels. The distribution is revised – yielding the posterior distribution – as new claims data become available. In the more practical empirical Bayes setting, $U(\theta)$ is rather seen as the structure function of the portfolio, that is the distribution of risk levels within the group of policyholders.

Prediction

The goal in greatest accuracy credibility is to compute the best prediction of the future claims $S_{i,n+1}$ for every policyholder. If the risk level of policyholder i were known, the best (in the mean square sense) prediction would be the expected value

$$\mu(\theta_i) = E[S_{it}|\Theta_i = \theta_i] = \int_0^\infty x dF(x|\theta_i) \quad (10)$$

In the actuarial literature, this function is called the *risk premium* (see **Equity-Linked Life Insurance; Inequalities in Risk Theory; Risk Attitude**). Now, the risk levels and, consequently, the risk premiums are unknown. One is thus left with the equivalent problems of predicting future claims $S_{i,n+1}$ or finding approximations of the risk premiums $\mu(\theta_i)$.

A first approximation of the risk premiums is the weighted average of all possible risk premiums:

$$m = E[\mu(\Theta)] = \int_{-\infty}^\infty \mu(\theta) dU(\theta) \quad (11)$$

This approximation will be the same for all policyholders. It is the *collective premium* (see **Premium Calculation and Insurance Pricing; Ruin Probabilities; Computational Aspects**).

As explained earlier, the collective premium, although globally adequate, fails to achieve an optimal premium distribution among policyholders. In statistical terms, this means that there exist better approximations of the risk premiums when experience is available. Indeed, the best approximation (or

estimation, or prediction) of the risk premium $\mu(\theta_i)$ is the function $g^*(S_{i1}, \dots, S_{in})$ minimizing the mean square error

$$E[(\mu(\Theta) - g(S_{i1}, \dots, S_{in}))^2] \quad (12)$$

where $g(\cdot)$ is any function. Most standard mathematical statistics text (see, e.g. [13]) show that function $g^*(S_{i1}, \dots, S_{in})$ is the so-called Bayesian premium

$$\begin{aligned} B_{i,n+1} &= E[\mu(\Theta)|S_{i1}, \dots, S_{in}] \\ &= \int_{-\infty}^{\infty} \mu(\theta) dU(\theta|S_{i1}, \dots, S_{in}) \end{aligned} \quad (13)$$

Function $U(\theta|x_1, \dots, x_n)$ is the aforementioned posterior distribution of the risk levels, or the revised structure function of the portfolio, after claims experience became available.

Let $f(x|\theta) = F'(x|\theta)$ be the conditional probability density function (pdf) or probability mass function (pmf) of the claim amounts and $u(\theta) = U'(\theta)$ be the pdf or pmf of the risk levels. Then, the posterior distribution of the risk levels is obtained from the prior distribution using the Bayes rule:

$$u(\theta_i|x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n|\theta_i)u(\theta_i)}{\int_{-\infty}^{\infty} f(x_1, \dots, x_n|\theta) dU(\theta)} \quad (14)$$

By conditional independence of claim amounts, this can be rewritten as

$$u(\theta_i|x_1, \dots, x_n) = \frac{\prod_{t=1}^n f(x_t|\theta_i)u(\theta_i)}{\int_{-\infty}^{\infty} \prod_{t=1}^n f(x_t|\theta) dU(\theta)} \quad (15)$$

The Bayesian premium $B_{i,n+1}$ is thus the best computable prediction of the future claims $S_{i,n+1}$. Akin to the collective premium, it is a weighted average of all possible risk premiums, but using the posterior rather than the prior distribution of the risk levels as weighting function. Reversing the argument, one can also see the collective premium as the Bayesian premium of the first year, when no experience is available.

Example 2 Consider policyholder 1 of the simplified portfolio of Example 1. This policyholder can incur at most one claim of amount 1 per period, but with unknown probability. If S_t is the random variable of the experience in year $t = 1, \dots, n$, then $S_t|\Theta = \theta$ has a Bernoulli distribution with parameter θ . Parameter θ is seen as an outcome of a random variable Θ having a beta prior distribution with parameters α and β . That is, we have

$$f(x_t|\theta) = \theta^x (1 - \theta)^{1-x}, \quad x = 0, 1 \quad (16)$$

and

$$u(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \quad 0 < \theta < 1 \quad (17)$$

The risk premium, or average claim amount, is $\mu(\theta) = E[S_t|\Theta = \theta] = \theta$. This information is of no real use since the true value of θ is unknown. One can however estimate the average claim amount by the collective premium

$$m = E[\mu(\Theta)] = E[\Theta] = \frac{\alpha}{\alpha + \beta} \quad (18)$$

This is the best approximation in the first year, when no other information is available. However, the results of Table 1 show that the collective premium is inappropriate for this policyholder.

The accumulation of experience allows one to better assess the true value of θ by means of equation (15). The posterior distribution of Θ after n years of experience is, up to a proportionality constant,

$$\begin{aligned} u(\theta|x_1, \dots, x_n) &\propto \theta^{\alpha-1} (1 - \theta)^{\beta-1} \\ &\times \prod_{t=1}^n \theta^{x_t} (1 - \theta)^{1-x_t} \\ &= \theta^{\alpha + \sum_{t=1}^n x_t - 1} (1 - \theta)^{\beta + n - \sum_{t=1}^n x_t - 1} \end{aligned} \quad (19)$$

That is, the distribution of $\Theta|S_1 = x_1, \dots, S_n = x_n$ is a beta distribution with updated parameters $\tilde{\alpha} = \alpha + \sum_{t=1}^n x_t$ and $\tilde{\beta} = \beta + n - \sum_{t=1}^n x_t$. Therefore, the Bayesian premium for year $n + 1$ is

$$B_{n+1} = E[\mu(\Theta)|S_1, \dots, S_n] \quad (21)$$

$$= E[\Theta | S_1, \dots, S_n] \quad (22)$$

$$= \frac{\tilde{\alpha}}{\tilde{\alpha} + \tilde{\beta}} \quad (23)$$

$$= \frac{\alpha + \sum_{t=1}^n S_t}{\alpha + \beta + n} \quad (24)$$

Setting $\alpha = 1$ and $\beta = 4$ in the above results gives a collective premium of 0.2, as in Example 1. Figure 1 shows the evolution of the probability of claim distribution from the prior to the posterior after 10 years of experience. One sees that the distribution gets increasingly concentrated around the true value of θ . Accordingly, the Bayesian premiums in Table 2 are getting closer to the true risk level of the policyholder.

Table 2 Bayesian premiums for policyholder 1 of Table 1 using a beta prior distribution with $\alpha = 1$ and $\beta = 4$

n	x_n	$\sum_{t=1}^n x_t$	$\alpha + \sum_{t=1}^n x_t$	$\beta + n - \sum_{t=1}^n x_t$	B_{n+1}
0	—	—	1	4	0.200
1	0	0	1	5	0.167
2	1	1	2	5	0.286
3	1	2	3	5	0.375
4	0	2	3	6	0.333
5	0	2	3	7	0.300
6	0	2	3	8	0.273
7	1	3	4	8	0.333
8	1	4	5	8	0.385
9	1	5	6	8	0.429
10	1	6	7	8	0.467

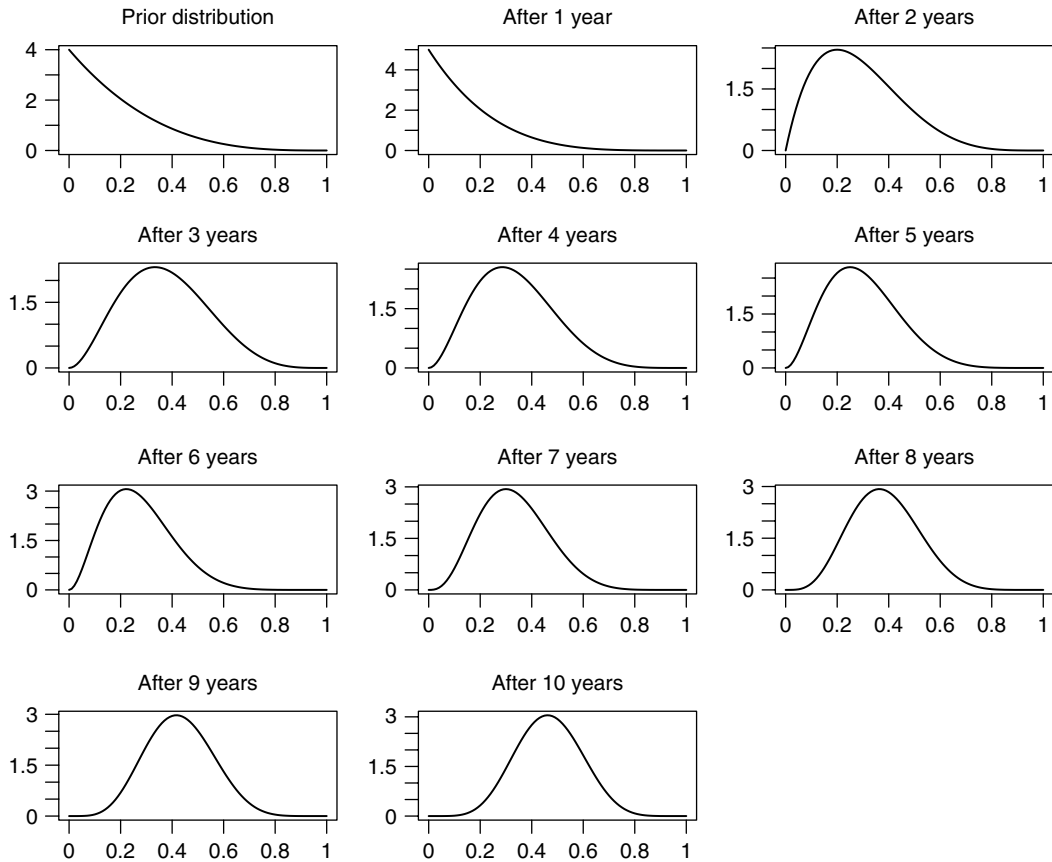


Figure 1 Posterior distributions for policyholder 1 of Table 1 using a beta prior distribution with $\alpha = 1$ and $\beta = 4$

Linear Bayes Credibility

The Bayesian premium equation (24) can be rewritten as

$$B_{n+1} = \frac{n}{n + \alpha + \beta} \bar{S} + \frac{\alpha + \beta}{n + \alpha + \beta} \frac{\alpha}{\alpha + \beta} \quad (25)$$

$$= z\bar{S} + (1 - z)m \quad (26)$$

with $z = n/(n + \alpha + \beta)$, the same form as equation (6). A premium of this form is called a *credibility premium* and z is the *credibility factor*.

Whitney [8] and Bailey [11] were the first to show that the Bayesian premium is a credibility premium for certain combinations of distributions and, furthermore, that the credibility factor is always of the form $z = n/(n + K)$, with K some constant. Bailey's partial results were extended by Mayerson [14] and later unified by Jewell [15].

The five main combinations of distributions giving a linear Bayesian premium are given in Table 3 (see [3] for a more complete table). Note that in all cases the posterior distribution of the risk parameter is of the same form as the prior, but with revised parameters.

The credibility premium is a weighted average between the collective premium and the individual average of a policyholder. The weight given to the latter increases with the number of years of experience (n). This is both simple and intuitively sound, two desirable properties of the credibility premium.

The Bühlmann and Bühlmann–Straub Models

In practice, the Bayesian premium equation (13) is of limited interest for two main reasons. First, one has to make assumptions on the distributions of $S_{it}|\Theta_i$ and Θ_i , something that can be difficult and highly subjective. Second, the Bayesian premium can be complicated to obtain – see [2] for what happens when the beta distribution in Example 2 is replaced by an apparently simple uniform distribution on (a, b) – and does not in general lie between the individual and collective premiums.

Bühlmann [5] proposed to restrict the approximation of the risk premium to linear functions of the observations. The optimal linear prediction happens to be a credibility premium

$$\pi_{i,n+1}^B = z\bar{S}_i + (1 - z)m \quad (27)$$

with

$$z = \frac{n}{n + s^2/a} \quad (28)$$

$$s^2 = E[\sigma^2(\Theta)] \quad (29)$$

$$a = \text{var}[\mu(\Theta)] \quad (30)$$

and $\sigma^2(\Theta_i) = \text{var}[S_{it}|\Theta_i]$.

The so-called structure parameters m , s^2 and a define the inner structure of the portfolio:

1. m measures the global average of the portfolio;
2. s^2 measures the variability within policyholders, that is of claims in time;

Table 3 Combinations of distributions yielding a linear Bayesian premium and main results

$f(x \theta)$	$u(\theta)$	B_{n+1}	z
Bernoulli(θ)	Beta(α, β)	$\frac{\alpha + \sum_{t=1}^n S_t}{\alpha + \beta + n}$	$\frac{n}{n + \alpha + \beta}$
Geometric(θ)	Beta(α, β)	$\frac{\beta + \sum_{t=1}^n S_t}{\alpha + n - 1}$	$\frac{n}{n + \alpha - 1}$
Poisson(θ)	Gamma(α, λ)	$\frac{\alpha + \sum_{t=1}^n S_t}{\lambda + n}$	$\frac{n}{n + \lambda}$
Exponential(θ)	Gamma(α, λ)	$\frac{\lambda + \sum_{t=1}^n S_t}{\alpha + n - 1}$	$\frac{n}{n + \alpha - 1}$
Normal(θ, σ_2^2)	Normal(μ, σ_1^2)	$\frac{\sigma_1^2 \sum_{t=1}^n S_t + \sigma_2^2 \mu}{n\sigma_1^2 + \sigma_2^2}$	$\frac{n}{n + \sigma_2^2/\sigma_1^2}$

3. a measures the variability between policyholders, that is the heterogeneity of the portfolio.

One can see that more weight will be given to individual experience in the following scenarios:

1. the number of periods of experience (n) is large. This is justified since the experience of a policyholder represents its true risk level in the long run;
2. the value of parameter s^2 is small, the claims experience is globally stable in time. The individual averages $\bar{S}_i$ are thereby indicative of the risk levels, hence reducing the need for the collective premium;
3. the value of a is large, the portfolio is heterogeneous. Therefore, the individual averages are more accurate approximations of the risk premiums than the collective premium.

Figures 2 and 3 provide a graphical interpretation of the last two points. Each curve in these figures represents the experience of one policyholder. In each case, the credibility factor is largest in the figures marked (b).

The Bühlmann–Straub model (*see Insurance Pricing/Nonlife*) [16] is a generalization of the Bühlmann model allowing for variable exposure. This is essential in lines of business in which the size of the policyholders can vary within a portfolio. Consider for example workers compensation insurance where

policyholders are employers: the exposure to risk of an employer with 1000 employees is much larger than that of an employer with only 10 employees.

In the Bühlmann–Straub model, a weight w_{it} (the number of employees or the payroll, for example) is assigned to each observation that we will now denote by X_{it} . One would expect that the experience of the larger policyholder will be more stable in time. To reflect this, the assumption of identical distribution used in Bühlmann’s model is relaxed to let the conditional variance of an observation be inversely proportional to its weight:

$$\text{var}[X_{it}|\Theta_i = \theta_i] = \frac{\sigma^2(\theta_i)}{w_{it}} \quad (31)$$

For this relation to hold, the observations X_{it} must now be ratios. Typically, the data will be defined as claim amounts divided by the exposure base:

$$X_{it} = \frac{S_{it}}{w_{it}} \quad (32)$$

The credibility premium differs from equation (27) in two respects only: the individual premium is estimated by a weighted average, and the credibility factor is a function of the total weight of a policyholder. We have

$$\pi_{i,n+1}^{\text{BS}} = z_i X_{iw} + (1 - z_i)m \quad (33)$$

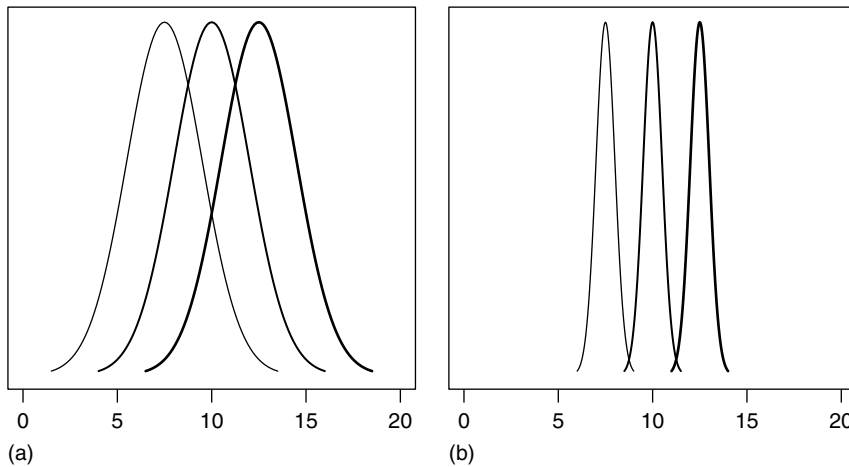


Figure 2 Effect of $s^2 = E[\sigma^2(\Theta)]$ on the credibility factor. (a) Large s^2 , experience is too volatile to be reliable. (b) Small s^2 , individual averages are more reliable. Means are the same in (a) and (b)

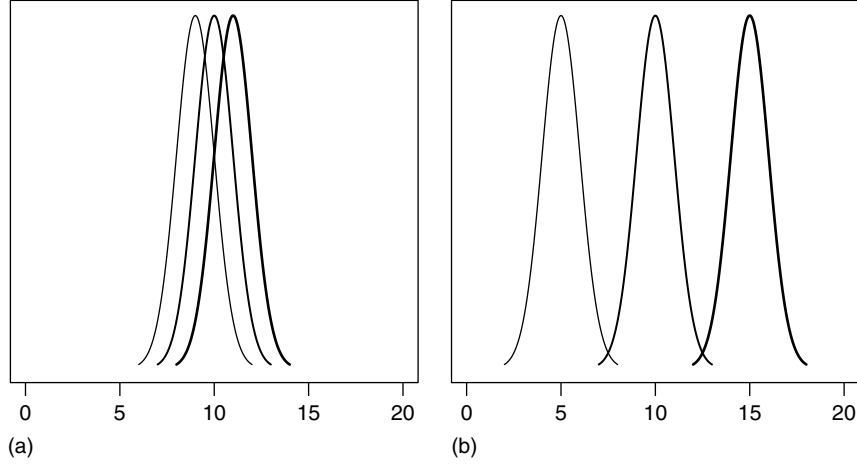


Figure 3 Effect of $a = \text{var}[\mu(\Theta)]$ on the credibility factor. (a) Small a , the portfolio is homogeneous. (b) Large a , the portfolio is heterogeneous. Variances are the same in (a) and (b)

with

$$z_i = \frac{w_{i\Sigma}}{w_{i\Sigma} + s^2/a} \quad (34)$$

$$X_{iw} = \sum_{t=1}^n \frac{w_{it}}{w_{i\Sigma}} X_{it} \quad (35)$$

and $w_{i\Sigma} = \sum_{t=1}^n w_{it}$. Parameters s^2 and a are defined as in equations (29) and (30), respectively.

In practice, the unknown structure parameters m , s^2 , and a are estimated from the available data; see [17, 18] or [3] for discussions on this topic. (Function `cm` of the R package `actuar` [19] implements all credibility calculations.)

Example 3 The Bühlmann model is appropriate to estimate the pure premiums (or probability of accident) for year 11 in Example 1. Unbiased estimators of the structure parameters are

$$\hat{m} = \bar{S} = 0.23 \quad (36)$$

$$\hat{s}^2 = \frac{1}{I(n-1)} \sum_{i=1}^I \sum_{t=1}^n (S_{it} - \bar{S}_i)^2 = 0.1367 \quad (37)$$

$$\hat{a} = \frac{1}{I-1} \sum_{i=1}^I (\bar{S}_i - \bar{S})^2 - \frac{\hat{s}^2}{n} = 0.0464 \quad (38)$$

Replacing these values in equation (27) yields the estimated credibility premiums of Table 4. One will

note that the premium for policyholder 1 obtained with this non parametric approach is different from the pure Bayes premium of Example 2.

Other Credibility Models

Random Coefficients Regression Model

Consider a data set in which a structural linear upward trend is present. Since the Bühlmann and Bühlmann–Straub models assume that the risk premium remains constant in time, these models will

Table 4 Bühlmann credibility premiums for the data of Table 1

Policy holder	Individual premium ($\bar{S}_i$)	Credibility factor ($\hat{z}$)	Credibility premium ($\hat{\pi}_{i,11}$)
1	0.6	0.772	0.516
2	0.3	0.772	0.284
3	0.2	0.772	0.207
4	0.2	0.772	0.207
5	0.2	0.772	0.207
6	0.1	0.772	0.130
7	0.0	0.772	0.052
8	0.0	0.772	0.052
9	0.7	0.772	0.593
10	0.0	0.772	0.052

systematically underestimate future claims. The random coefficients regression credibility model introduced by Hachemeister [20] remedies this situation by relaxing the constant mean assumption of the Bühlmann–Straub model in favor of

$$E[X_{it}|\Theta_i = \theta_i] = \beta_0(\theta_i) + \beta_1(\theta_i)t \quad (39)$$

Therefore, one has to estimate the intercept and slope of a regression line for each policyholder, but these depend on the unknown risk levels of the policyholders. The optimal estimation of the slope for one policyholder is a credibility weighted average between the estimation using this policyholder's experience only and the estimation using all the portfolio data. The same idea holds for the intercept.

The formulas in this model make heavy use of matrix notation and quickly become cumbersome. We refer the interested reader to [18] for an up-to-date discussion on this topic.

Of course, one can use something other than a linear trend model. In general, one would have

$$E[\mathbf{X}_i|\Theta_i = \theta_i] = \mathbf{Y}\boldsymbol{\beta}(\theta_i) \quad (40)$$

where $\mathbf{X}_i = (X_{i1}, \dots, X_{in})'$, $\mathbf{Y}$ is an $n \times (p+1)$ design matrix and $\boldsymbol{\beta}(\theta_i) = (\beta_0(\theta_i), \dots, \beta_p(\theta_i))'$ is a vector of random coefficients.

Hierarchical Credibility

It is common that the group of policyholders that we have been referring to as *the portfolio* is actually just a fraction of a much larger classification scheme. Consider, for example, a hierarchical (or treelike) structure in which policyholders are classified first in “sectors”, and then in “classes”. Applying a credibility model within one class will not make use of collateral data from other classes of the same sector, even though this data can be useful to assess the true risk level of the class and, ultimately, of the policyholders (*see Dependent Insurance Risks*).

Exploiting collateral data was the main rationale brought forth by Jewell [21] to introduce the hierarchical credibility model. The credibility premium of a policyholder remains of the same form as equation (27) or equation (33), but with the complement of credibility $(1 - z_i)$ given to the *credibility premium* of the class. The latter is, in turn, a weighted average between the experience of the class and the credibility premium of the sector, and so on, until the top level is reached.

Complete formulas for the hierarchical credibility model can be found in [3, 18, 22, 23]. Hierarchical credibility is particularly well suited for a geometrical interpretation; see [24]. Moreover, [3] has a discussion on the uses and misuses of hierarchical models.

Crossed Classification Credibility

Consider an automobile insurance portfolio in which policyholders are classified according to gender and age. Intuitively, young women share risk characteristics with young men, who themselves share risk characteristics with older men. A hierarchical classification structure is inappropriate for such a portfolio since the risk factors are not nested, but rather crossed.

Dannenburg [25] adapted the crossed classification random-effects models, well known in statistical modeling, to the credibility theory context. This results in a very general model encompassing the Bühlmann and Bühlmann–Straub models and, to some extent, the hierarchical model; see [3, 26] for details and formulas.

Concluding Remarks

Credibility theory is related – without being identical – to various forecasting techniques in statistics. There has been a trend over the last decade or so to increasingly draw links between credibility and techniques such as the Kalman filter (*see Non-life Loss Reserving*) [27], longitudinal models [28], variance components models [26], or generalized linear models [29]. See also [30] for further comments and references.

A lot of research in credibility theory has been and will likely continue to be devoted to estimation of the structure parameters. Robust estimation or generalized estimating equations come to mind as procedures making their way into credibility. Moreover, the tidal wave of dependence modeling that has swept risk theory around the start of the new millennium has just started to touch credibility theory [31]. More is expected in the near future.

References

- [1] Bühlmann, H. (1969). Experience rating and credibility, *ASTIN Bulletin* 5, 157–165.

- [2] Norberg, R. (1979). The credibility approach to ratemaking, *Scandinavian Actuarial Journal* **1979**, 181–221.
- [3] Goulet, V. (1998). Principles and application of credibility theory, *Journal of Actuarial Practice* **6**, 5–62.
- [4] Neuhaus, W. (2004). Experience rating, in *Encyclopedia of Actuarial Science*, ISBN 0-4708467-6-3, J.L. Teugels & B. Sundt, eds, John Wiley & Sons, Vol. 2, pp. 639–646.
- [5] Bühlmann, H. (1967). Experience rating and credibility, *ASTIN Bulletin* **4**, 199–207.
- [6] Mowbray, A.H. (1914). How extensive a payroll exposure is necessary to give a dependable pure premium? *Proceedings of the Casualty Actuarial Society* **1**, 25–30.
- [7] Longley-Cook, L.H. (1962). An introduction to credibility theory, *Proceedings of the Casualty Actuarial Society* **49**, 194–221.
- [8] Whitney, A.W. (1918). The theory of experience rating, *Proceedings of the Casualty Actuarial Society* **4**, 275–293.
- [9] Fisher, A. (1919). The theory of experience rating: discussion, *Proceedings of the Casualty Actuarial Society* **5**, 139–145.
- [10] Bailey, A.L. (1945). A generalized theory of credibility, *Proceedings of the Casualty Actuarial Society* **32**, 13–20.
- [11] Bailey, A.L. (1950). Credibility procedures, Laplace's generalization of Bayes' rule and the combination of collateral knowledge with observed data, *Proceedings of the Casualty Actuarial Society* **37**, 7–23.
- [12] Russell, B. (1948). *Human Knowledge: Its Scope and Limits*, George Allen & Unwin, London.
- [13] Hogg, R.V., Craig, A.T. & McKean, J.W. (2005). *Introduction to Mathematical Statistics*, ISBN 0-1300850-7-3, 6th Edition, Prentice Hall, Upper Saddle River.
- [14] Mayerson, A.L. (1964). A Bayesian view of credibility, *Proceedings of the Casualty Actuarial Society* **51**, 85–104.
- [15] Jewell, W.S. (1974). Credible means are exact Bayesian for exponential families, *ASTIN Bulletin* **8**, 77–90.
- [16] Bühlmann, H. & Straub, E. (1970). Glaubwürdigkeit für Schadensätze, *Bulletin of the Swiss Association of Actuaries* **70**, 111–133.
- [17] Dubey, A. & Gisler, A. (1981). On parameter estimation in credibility, *Bulletin of the Swiss Association of Actuaries* **81**, 187–211.
- [18] Bühlmann, H. & Gisler, A. (2005). *A Course in Credibility Theory and Its Applications*, ISBN 3-5402575-3-5, Springer.
- [19] Goulet, V. (2005). *Actuar: an R Package for Actuarial Science*, <http://www.actuar-project.org>.
- [20] Hachemeister, C.A. (1975). Credibility for regression models with application to trend, in *Credibility, Theory and Applications, Proceedings of the Berkeley Actuarial Research Conference on Credibility*, Academic Press, New York, pp. 129–163.
- [21] Jewell, W.S. (1975). The use of collateral data in credibility theory: a hierarchical model, *Giornale dell'Istituto Italiano degli Attuari* **38**, 1–16.
- [22] Goovaerts, M.J., Kaas, R., van Heerwaarden, A.E. & Bauwelinckx, T. (1990). *Effective Actuarial Methods*, ISBN 0-4448839-9-1, North-Holland, Amsterdam.
- [23] Goovaerts, M.J. & Hoogstad, W.J. (1987). *Credibility Theory, Surveys of Actuarial Studies*, Number 4, Nationale-Nederlanden, Netherlands.
- [24] Bühlmann, H. & Jewell, W.S. (1987). Hierarchical credibility revisited, *Bulletin of the Swiss Association of Actuaries* **87**, 35–54.
- [25] Dannenburg, D. (1995). Crossed classification credibility models, in *Transactions of the 25th International Congress of Actuaries*, International Actuarial Association Vol. 4, pp. 1–35.
- [26] Dannenburg, D.R., Kaas, R. & Goovaerts, M.J. (1996). *Practical Actuarial Credibility Models*, ISBN 9-0802117-3-7, Ceuterick, Leuven.
- [27] de Jong, P. & Zehnwirth, B. (1983). Credibility theory and the Kalman filter, *Insurance: Mathematics and Economics* **2**, 281–286.
- [28] Frees, E.W., Young, V. & Luo, Y. (1999). A longitudinal data analysis interpretation of credibility models, *Insurance: Mathematics and Economics* **24**, 229–247.
- [29] Kaas, R., Goovaerts, M., Dhaene, J. & Denuit, M. (2001). *Modern Actuarial Risk Theory*, ISBN 0-7923763-6-6, Kluwer Academic Publishers, Dordrecht.
- [30] Norberg, R. (2004). Credibility theory, in *Encyclopedia of Actuarial Science*, ISBN 0-4708467-6-3, J.L. Teugels & B. Sundt, eds, John Wiley & Sons, Vol. 1, pp. 398–406.
- [31] Denuit, M., Dhaene, J., Goovaerts, M. & Kaas, R. (2005). *Actuarial Theory for Dependent Risks*, ISBN 0-4700149-2-X, John Wiley & Sons, Chichester.

VINCENT GOULET

Credit Migration Matrices

Credit migration (*see* **Default Risk**) or transition matrices (*see* **Group Decision**), which characterize past changes in credit quality of obligors (typically firms), are cardinal inputs to many risk management applications, including portfolio risk assessment (*see* **Credit Scoring via Altman Z-Score; Volatility Modeling**), pricing of bonds and credit derivatives (*see* **Insurance Pricing/Nonlife; Default Risk; Mathematical Models of Credit Risk**), and assessment of risk capital. For example, standard bond pricing models such as in [1] require a ratings projection of the bond to be priced. These matrices even play a role in regulation: in the New Basel Capital Accord (*see* **Enterprise Risk Management (ERM); Credit Risk Models; Solvency; Extreme Value Theory in Finance**) [2], capital requirements are driven in part by ratings migration. This article provides a brief overview of credit migration matrix basics: how to compute them, how to make inference and compare them, and some examples of their use. We pay special attention to the last column of the matrix, namely, the migration to default. Along the way, we illustrate some of the points with data from one of the rating agencies, Standard & Poor's (S&P).

To fix ideas, suppose that there are k credit ratings for $k - 1$ nondefault states and one default state. This rating is designed to serve as a summary statistic of the credit quality of a borrower or obligor such as a firm. Ratings can be either public as provided by one of the rating agencies (*see* **Nonlife Insurance Markets; Informational Value of Corporate Issuer Credit Ratings; Credit Scoring via Altman Z-Score**) such as Fitch, Moody's, or S&P, or they can be private such as an obligor rating internal to a bank. For firms that have issued public bonds, typically at least one rating from a rating agency is available [3]. As such, the rating agencies are expected to follow the credit quality of the firm. When that changes, the agency may decide to upgrade or downgrade the credit rating of the firm. In principle, the process from initial rating to the updates is the same within a financial institution when assigning so-called internal ratings, though the monitoring may be less intensive and hence the updating may be less frequent [4]. Purely as a matter of convenience, we follow the notation used by Fitch and S&P which, from best

to worst, is AAA, AA, A, BBB, BB, B, CCC, and of course, D; so $k = 8$.

All three rating agencies actually provide rating modifiers (e.g., for Fitch and S&P, these are $\pm$, as in AA- or AA+) to arrive at a more nuanced, 17+ state rating system. But for simplicity, most of the discussion in this article is confined to whole grades.

A concrete example of a credit migration is given below in Table 1 in which we show the 1-year migration probabilities for firms, estimated using S&P ratings histories from 1981 to 2003. A given row denotes the probability of migrating from rating i at time T to any other rating j at time $T + 1$. For example, the 1-year probability that an AA rated firm is downgraded to A is 7.81%.

Several features – and these are typical – immediately stand out. First, the matrix is diagonally dominant, meaning that large values lie on the diagonal (made bold for emphasis) denoting the probability of no change or migration: for most firms ratings do not change. Such stability is by design [5]: the agencies view that investors look to them for just such stable credit rating assessments. The next largest entries are one step off the diagonal, meaning that when there are changes, they tend to be small, that is, one or perhaps two rating grades.

The final column, the migration to default, deserves special attention. Probabilities of default increase roughly exponentially as one descends the credit spectrum from best (AAA) to worst (CCC). Note that in our sample period of 1981–2003, no AAA-rated firm defaulted, so that the empirical estimate of this probability of default (PD) (PD_{AAA}) is zero. But is zero an acceptable estimate of PD_{AAA} ? We revisit this question below.

To square the matrix, the last row is the unit vector that simply states that default is an absorbing state: once a firm is in default, it stays there. By implication, it means that all firms eventually default, though it may take a (very) long time. A firm that emerges from bankruptcy (default) is typically treated as a new firm.

Estimation

Several approaches to estimating these migration matrices are presented and reviewed in [6] and compared extensively in [7]. Broadly there are two approaches, cohort and two variants of duration

Table 1 One-year credit migration matrix using S&P rating histories, 1981–2003. Estimation method is cohort. All values in percentage points

		T + 1							
		AAA	AA	A	BBB	BB	B	CCC	D
T	AAA	92.29	6.96	0.54	0.14	0.06	0.00	0.00	0.000
	AA	0.64	90.75	7.81	0.61	0.07	0.09	0.02	0.010
	A	0.05	2.09	91.38	5.77	0.45	0.17	0.03	0.051
	BBB	0.03	0.20	4.23	89.33	4.74	0.86	0.23	0.376
	BB	0.03	0.08	0.39	5.68	83.10	8.12	1.14	1.464
	B	0.00	0.08	0.26	0.36	5.44	82.33	4.87	6.663
	CCC	0.10	0.00	0.29	0.57	1.52	10.84	52.66	34.030
	D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100

(or hazard)–parametric (imposing time homogeneity) and nonparametric (relaxing time homogeneity). The assumption of time homogeneity essentially implies that the process is time invariant: the analyst can be indifferent between two equally long samples drawn from different time periods.

The straight forward cohort approach has become the industry standard. In simple terms, the cohort approach just takes the observed proportions from the beginning of the year to the end (for the case of annual migration matrices) as estimates of migration probabilities (*see Default Risk*). Suppose there are $N_i(t)$ firms in rating category i at the beginning of the year t , and $N_{ij}(t)$ migrated to grade j by year end. An estimate of the transition probability for year t is $P_{ij}(t) = \frac{N_{ij}(t)}{N_i(t)}$. For example, if two firms out of 100 migrated from grade “AA” to “A”, then $P_{AA \rightarrow A} = 2\%$. Any movements within the year are not accounted for. Typically, firms whose ratings were withdrawn or migrated to not rated (NR) status are removed from the sample. This approach effectively treats migrations to NR as being noninformative [8]. It is straightforward to extend this approach to multiple years. For instance, suppose that we have data for T years, then the estimate for all T years is:

$$P_{ij} = \frac{N_{ij}}{N_i} = \frac{\sum_{t=1}^T N_{ij}(t)}{\sum_{t=1}^T N_i(t)} \quad (1)$$

Indeed, this is the maximum likelihood estimate of the transition probability under a discrete time-homogeneous Markov chain. The matrix shown in Table 1 was estimated using the cohort approach.

Any rating change activity that occurs *within* the period is ignored, unfortunately. A strength of the alternative duration approach is that it counts *all* rating changes over the course of the year (or multiyear period) and divides by the number of firm-years spent in each state or rating to obtain a matrix of migration intensities that are assumed to be time homogenous. Under the assumption that migrations follow a Markov process, these intensities can be transformed to yield a matrix of migration probabilities.

Following [6], the $k \times k$ transition probability matrix $\mathbf{P}(t)$ can be written as

$$\mathbf{P}(t) = \exp(\mathbf{\Gamma}t) \quad t \geq 0 \quad (2)$$

where the exponential is a matrix exponential, and the entries of the generator matrix $\mathbf{\Gamma}$ satisfy

$$\begin{aligned} \gamma_{ij} &\geq 0 \text{ for } i \neq j \\ \gamma_{ii} &= - \sum_{j \neq i} \gamma_{ij} \end{aligned} \quad (3)$$

The second expression in equation (3) merely states that the diagonal elements are such to ensure that the rows sum to zero.

The maximum likelihood estimate of an entry γ_{ij} in the intensity matrix $\mathbf{\Gamma}$ is given by

$$\hat{\gamma}_{ij} = \frac{n_{ij}(T)}{\int_0^T Y_i(s) ds} \quad (4)$$

where $Y_i(s)$ is the number of firms with rating i at time s , and $n_{ij}(T)$ is the total number of transitions over the period from i to j where $i \neq j$. The denominator in equation (4) effectively is the number of “firm-years” spent in state i . Thus for a horizon

the additional year of data has only a modest impact on the estimate of the migration matrix.

One may be particularly interested in the precision of default probability estimates. The first to report confidence sets for default probability estimates were Christensen *et al.* [13], who used a parametric bootstrap. An interesting approach for the common case where no defaults have actually been observed was developed in [17] on the basis of the most prudent estimation principle, assuming that the ordinal borrower ranking is correct (i.e., monotonic). A systematic comparison of confidence intervals was provided by Hanson and Schuermann [18], using several analytical approaches, as well as finite-sample confidence intervals obtained from parametric and nonparametric bootstrapping. They find that the bootstrapped (*see Benchmark Dose Estimation; Combining Information; Nonlife Loss Reserving*) intervals for the duration-based estimates are surprisingly tight and that the less efficient cohort approach generates much wider intervals. Yet, even with the tighter bootstrapped confidence intervals for the duration-based estimates, it is impossible to statistically distinguish notch-level (grade with $\pm$ modifiers) PD s for neighboring investment grade ratings, e.g., a PD_{AA-} from a PD_{A+} or even a PD_A . However, once the speculative grade barrier (i.e., moving from $BBB-$ to $BB+$) is crossed, they are able to distinguish quite cleanly notch-level estimated default probabilities. Moreover, both [18, 19] show that PD point estimates and, unsurprisingly, their confidence intervals vary substantially over time.

An advantage of the duration over the cohort estimation approach is that it delivers nonzero default probability estimates even when no actual defaults were observed. As a result, the PD estimates can be quite different, even taking into account the issue of estimation noise raised above. This is shown in Table 3 at the more granular notch-level using S&P ratings histories for 1981–2004. Note that neither method produces monotonically increasing PD s, though in the presence of estimation noise, this nonmonotonicity need not be surprising. Put differently, even if the true but unknown PD s are monotonically increasing, because each rating's PD is estimated with error, the *estimates* need not be monotonic. Since no actual defaults have been observed for AAA-rated (nor AA+ or AA rated) firms over the course of the sample period, the cohort estimates must be identically equal to zero even as the duration

Table 3 Unconditional probability of default (PD) estimates compared using S&P rating histories, 1981–2004. All values are in basis points (bp), where 100 bp = 1%

	Cohort	Duration
AAA	0.00	0.02
AA+	0.00	0.05
AA	0.00	0.65
AA–	2.43	0.26
A+	5.42	0.31
A	4.00	0.88
A–	3.97	0.68
BBB+	22.99	3.22
BBB	30.42	8.48
BBB–	41.76	12.03
BB+	57.24	20.61
BB	104.63	41.41
BB–	197.34	84.32
B+	336.84	172.33
B	942.53	707.07
B–	1384.62	1247.41
CCC	3253.55	4320.35

(*see Asset–Liability Management for Life Insurers; Securitization/Life*) approach generates a very small, but nonzero, estimate of 0.02 basis points (bp) (or 0.0002%).

In looking at the difference for PD_{CCC} between the two methods, Lando and Skodeberg [6] observe that most firms default after only a brief stop in the CCC rating state. By contrast, the intermediate grades generate duration PD estimates that are below the cohort estimates. As argued in [18], if ratings exhibit downward persistence (firms that enter a state through a downgrade are more likely to be downgraded than other firms in the state), as shown among others in [6, 11, 12], one would expect PD s from the duration-based approach, which assumes that the migration process is Markov, to be downward biased. Such a bias would arise because the duration estimator ignores downward ratings momentum and consequently underestimates the probability of a chain of successive downgrades ending in default.

The New Basel Accord sets a lower bound of 0.03% on the PD estimate that may be used to compute regulatory capital for the internal ratings based (IRB) approach [2, §285]. Table 3 suggests that the

top two ratings, AAA and AA, would both fall under that limit and would thus be indistinguishable from a regulatory capital perspective. Indeed Hanson and Schuermann [18] report that once 95% confidence intervals are taken into account, the top three ratings, AAA through A, are indistinguishable from 0.03%.

Applications

The applications and uses of credit migration matrices are myriad, from asset pricing to portfolio choice, and risk management to bank regulation. Two examples are presented here; the pricing of a yield spread option and the computation of risk capital for a credit portfolio using CreditMetrics.

Yield Spread Option

A yield spread option enables the buyer and seller to speculate on the evolution of the yield spread. The yield spread is defined as the difference between the continuously compounded yield of a risky and a riskless zero-coupon bond with the same maturity. Depending on the option specifications, the relevant spread is either an individual forward spread in case of European options, or a bundle of forward spreads in case of American options. Call (put) option buyers expect a decreasing (increasing) credit spread.

Yield spread options are priced using Markov chain models such as the one presented in [20]. To price such an option, the following are needed: the yield curve of default-free zero-coupon bonds, the term structure of forward credit spreads, both the option and yield spread maturity (of the underlying bond), an estimate of the recovery rate in the event of default, the current rating of the bond, and of course the migration vector of the same maturity as the option corresponding to that rating. This is illustrated in Figure 1 below, for a yield spread option on a BBB-rated bond.

Risk Capital for a Credit Portfolio

The purpose of capital is to provide a cushion against losses for a financial institution. The amount of required economic capital is commensurate with the risk appetite of the financial institution. This boils down to choosing a confidence level in the loss (or value change) distribution of the institution

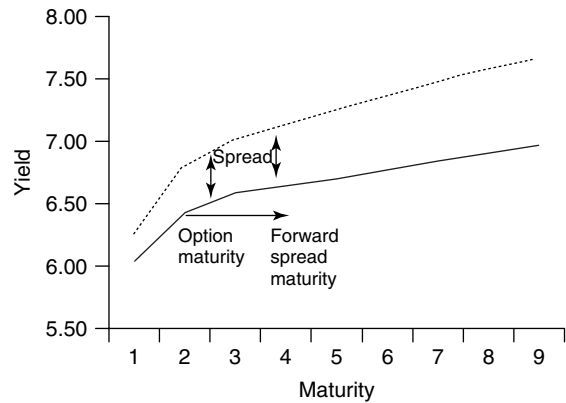


Figure 1 Illustration of a yield spread option on a BBB-rated bond

with which senior management is comfortable. For instance, if the bank wishes to have an annual survival probability of 99%, this will require less capital than a survival probability of 99.9%, where the latter is the confidence level to which the New Basel Capital Accord is calibrated [2], and is typical for a regional bank (commensurate with a rating of about A-/BBB+, judging from Table 3). The loss (or value change) distribution is arrived at through internal credit portfolio models.

One such credit portfolio model (*see Credit Risk Models; Credit Value at Risk*) is CreditMetrics [21], a portfolio application of the options-based model of firm default due to Merton [22]. Given the credit rating distribution of exposures today, and inputs similar to the pricing of the yield spread option discussed in the section titled “Yield Spread Option”, namely, the yield curve of default-free zero-coupon bonds, the term structure of forward credit spreads, an estimate of the recovery rate in the event of default for each rating, and of course the credit migration matrix, the model generates a value distribution of the credit portfolio through stochastic simulation. An illustration is given in Figure 2 above. For instance, the 99% value at risk (VaR) (*see Credit Scoring via Altman Z-Score; Risk Measures and Economic Capital for (Re)insurers; Default Risk*) is -14.24%, and the 99.9% VaR is -19.03%. Thus only for 1 year out of 1000 would the portfolio manager expect his or her portfolio to lose more than 19.03% of its value.

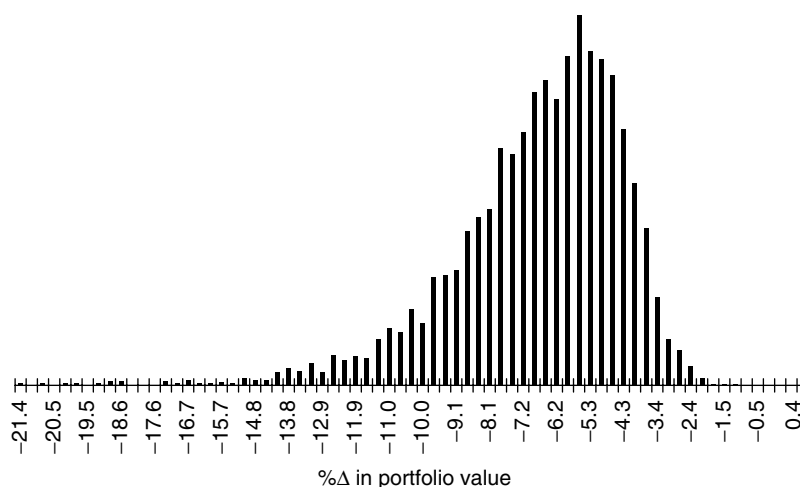


Figure 2 Illustrative example of a credit portfolio value change distribution using CreditMetrics

Summary

This article provided a brief overview of credit migration or transition matrices, which characterize past changes in credit quality of obligors (typically firms). They are cardinal inputs to many risk management applications, including portfolio risk assessment, the pricing of bonds and credit derivatives, and the assessment of regulatory capital as is the case for the New Basel Capital Accord. We addressed the question of how to estimate these matrices, how to make inference and compare them, and provided two examples of their use: the pricing of a derivative called a yield spread option, and the calculation of the value distribution for a portfolio of credit assets. The latter is especially useful for risk management of credit portfolios.

Acknowledgment

I would like to thank Benjamin Iverson for excellent research assistance, and Sam Hanson and an anonymous referee for helpful comments and suggestions. Any views expressed represent those of the author only and not necessarily those of the Federal Reserve Bank of New York or the Federal Reserve System.

References

- [1] Jarrow, R.A., Lando, D. & Turnbull, S.M. (1997). A Markov model for the term structure of credit risk spreads, *Review of Financial Studies* **10**, 481–523.
- [2] Bank for International Settlements (2005). *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*, <http://www.bis.org/publ/bcb118.htm>.
- [3] Cantor, R. & Packer, F. (1995). The credit rating industry, *Journal of Fixed Income* **11**(1), 10–34.
- [4] Mählmann, T. (2005). Biases in estimating bank loan default probabilities, *Journal of Risk* **7**, 75–102.
- [5] Altman, E.I. & Rijken, H.A. (2004). How rating agencies achieve rating stability, *Journal of Banking and Finance* **28**, 2679–2714.
- [6] Lando, D. & Skødeberg, T. (2002). Analyzing ratings transitions and rating drift with continuous observations, *Journal of Banking & Finance* **26**, 423–444.
- [7] Jafry, Y. & Schuermann, T. (2004). Measurement, estimation and comparison of credit migration matrices, *Journal of Banking & Finance* **28**, 2603–2639.
- [8] Carty, L.V. (1997). *Moody's Rating Migration and Credit Quality Correlation, 1920–1996*, Moody's Special Comment, Moody's Investor Service, New York.
- [9] Altman, E.I. & Kao, D.L. (1992). Rating drift of high yield bonds, *Journal of Fixed Income* **13**, 15–20.
- [10] Carty, L.V. & Fons, J.S. (1993). *Measuring Changes in Corporate Credit Quality*, Moody's Special Report, Moody's Investors Service, New York.
- [11] Nickell, P., Perraudin, W. & Varotto, S. (2000). Stability of rating transitions, *Journal of Banking & Finance* **24**, 203–227.
- [12] Bangia, A., Diebold, F.X., Kronimus, A., Schagen, C. & Schuermann, T. (2002). Ratings migration and the business cycle, with applications to credit portfolio stress testing, *Journal of Banking & Finance* **26**, 445–474.
- [13] Christensen, J.E.H. & Lando, D. (2004). Confidence Sets for continuous-time rating transition probabilities, *Journal of Banking & Finance* **28**, 2575–2602.

-
- [14] Giampieri, G., Davis, M. & Crowder, M. (2005). A hidden Markov model of default interaction, *Quantitative Finance* **5**, 27–34.
- [15] Gagliardini, P. & Gourioux, C. (2005). Stochastic migration models with application to corporate risk, *Journal of Financial Econometrics* **3**, 188–226.
- [16] Frydman, H. (2005). Estimation in the mixture of Markov chains moving with different speeds, *Journal of the American Statistical Association* **100**, 1046–1053.
- [17] Pluto, K. & Tasche, D. (2005). Estimating probabilities of default for low default portfolios, *Risk* **18**, 76–82.
- [18] Hanson, S.G. & Schuermann, T. (2006). Confidence intervals for probabilities of default, *Journal of Banking & Finance* **30**, 2281–2301.
- [19] Trück, S. & Rachev, S.T. (2005). Credit portfolio risk and probability of default confidence sets through the business cycle, *Journal of Credit Risk* **1**, 61–88.
- [20] Kijima, M. & Komoribayashi, K. (1998). A Markov chain model for valuing credit risk derivatives, *Journal of Derivatives* **6**, 97–108.
- [21] Gupton, G.M., Finger, C.C. & Bhatia, M. (1997). *CreditMetricsTM – Technical Document*, JP Morgan, New York.
- [22] Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.

TIL SCHUERMANN

Credit Risk Models

The computation of the distribution of aggregate losses in credit portfolios has become especially important for risk management (*see Asset–Liability Management for Life Insurers; Multistate Models for Life Insurance Mathematics; Mathematics of Risk and Reliability: A Select History*) and securitization (*see Securitization/Life*) purposes. Banks and financial institutions need to assess the risks within their credit portfolios both for regulatory requirements and for internal risk management: for instance, they may compute risk measures such as the value at risk or the expected shortfall associated with their credit risk exposure. Credit risk has been transferred from banks to other investors such as insurance companies or hedge funds that act with respect to commercial banks as reinsurance companies act with respect to insurance firms. This securitization process involves some “tranching” of credit risk similar to stop-loss reinsurance. Specific products such as credit default swaps and single tranche collateralized debt obligations (CDOs) (*see Tavakoli [1] for a description of the CDO market*) have been designed for such risk transfer. The computation of the CDO tranche premiums also involves the aggregate credit loss distributions over different time horizons. The purpose of this article is to describe the most commonly used approach to compute such loss distributions. Before this, we briefly review the risks involved in a credit portfolio.

An Overview of Risks Involved

Default Risk. Default risk is related to the inability of a borrower to reimburse a loan or a bond. More precisely, one can think of different credit events such as the following:

- the failure to meet payment obligations when due;
- bankruptcy (for nonsovereign entities) or Moratorium (for sovereign entities only);
- repudiation;
- material adverse restructuring of debt;
- obligation acceleration or obligation default.

The definition of a credit event hinges on the relevant bankruptcy rules (*see Default Risk*) that themselves depend upon geographical regions, quality of

the borrower, and seniority of the loan. For example, personal bankruptcy could be accessed more easily in the United States than in the United Kingdom. The same comparison also holds for corporates. This will result in differences in frequency and claim severity. Let us additionally remark that as far as retail credit is concerned, the banks need to decide at which stage a loan is actually in default. For instance, a single outstanding payment on a mortgage is not usually considered as a default event. We refer to de Malherbe [2] for a detailed quantitative assessment of the credit default swaps covenants.

Change in Credit Quality: Credit Migrations, Change in Credit Spreads. In addition to default risk, credit migrations are associated with changes in credit quality. For example, in financial markets, even if default-free interest rates remain constant, defaultable bond prices change prior to default. This is also known as *credit spread risk* (*see Extreme Value Theory in Finance*). For example, an increase in default probability (*see Credit Migration Matrices; Default Correlation; Copulas and Other Measures of Dependency*) following a credit rating downgrading, adverse earnings announcements should lead to a drop in the company’s bond prices. This is not specific to corporates; default probabilities for individuals usually increase when the borrower becomes unemployed.

Let us emphasize that credit contracts may have much longer maturities than insurance contracts. For example, in a typical 30-year mortgage, the annuities remain fixed even if the credit quality of the borrower deteriorates; the lender cannot raise payments yearly as in the insurance markets.

Recovery Risk. In case of default, the lender only gets a fraction of the promised payments. In the most severe cases, no further cash flows are being paid by the borrower. In case of bankruptcy, the assets of the defaulted firm are being sold to investors and the proceeds are used to reimburse the lenders. After liquidation, the proceeds may be in excess of the commitments to the lenders, which means that the loss severity can be equal to zero. For simplicity, we will consider that the loss given default (*LGD*) is equal to the difference between the face value of the loan or the bond and the recovered value. As a consequence, *LGD* are bounded.

Dependence Between Defaults. Since common macroeconomic factors, such as business cycles, level of unemployment, and shifts in monetary policy drive default both frequency and credit loss severity, we cannot rely on the standard insurance framework and we need to cope with dependence between default events. Macroeconomic shocks can be considered as exogenous to defaults and usually lead to positive dependence. Dependence can also occur at a local level owing to interactions between firms. For example, default of a bank can be seen as a bad signal with respect to the assets of its competitors in an incomplete information framework. Conversely, default of a firm could raise the market power of competitors, leading to extra profits and an increase in credit quality. As can be seen from these examples, defaults can be informative with respect to the credit quality of survivors; this is known as *contagion effects* or *infectious defaults*.

We will further detail the most commonly used approaches to the dependence between defaults as far as the computation of loss distributions is involved. Dependence usually results in much fatter tails and an increase in senior CDO tranche premiums, which correspond to stop-loss reinsurance premiums in insurance markets.

Computation of Loss Distributions

Marginal Default Probabilities. The individual default probabilities of the obligors within a credit portfolio are a key input in the computation of aggregate loss distributions (see **Premium Calculation and Insurance Pricing; Risk Measures and Economic Capital for (Re)insurers**).

As far as retail portfolios are involved, banks could either use internal data or, if available, public information about defaults and obligors' individual characteristics. Usually, obligors are grouped in homogeneous clusters and thus marginal default probabilities depend upon static observables, such as age, marital status, and so on. Credit scoring techniques based upon logistic regression or classification techniques are routinely used. For instance in the United States, FICO scores have almost become a market standard.

When it comes to corporate obligors, the number of obligors is usually much smaller and especially for investment grade bonds. Thus the number of observed

defaults is much smaller, which leads to some statistical difficulties. To assess the magnitude of default probabilities, creditors need to rely on the analysis of financial statements. This can be done at the lender's level, usually within the credit department of lending banks or owing to the rating agencies. While the financial analysis is forward looking and rates obligors along a discrete grid, further data analysis is required to translate credit ratings into default probability as illustrated by Moody's historical default rates (see **Default Risk**) of bond issuers.

Another approach that can be thought of, especially for publicly traded companies, is the use of the stock market. In the structural model of Merton (see **Credit Scoring via Altman Z-Score; Default Correlation**), default occurs if the market value of the company's assets falls below a prespecified threshold. The lower the value of the stock, the smaller the distance to the default barrier and the more likely default will occur. This is the route followed by Kealhofer [3].

Historical and Market Implied Default Probabilities

In frictionless and arbitrage-free financial markets (see **Equity-Linked Life Insurance**), prices of financial securities are computed as the expectation of the discounted risky cash flows. Let us consider a simplified defaultable bond with a single promised cash flow of 1 \$ to be paid at time t . In case of default, we assume that no payment will be made (recovery rate is equal to zero). Let us denote by $\bar{B}$ the market price for such a bond and by r , the (continuously compounded) discount rate for maturity t . $\bar{B} = e^{-rt} Q(\tau > t)$, where τ denotes the default date and $Q(\tau > t)$ is the survival probability. If the previous defaultable discount bond is actually traded, the survival probability can be readily computed as $Q(\tau > t) = \bar{B} \times e^{rt}$. It appears that the default probabilities extracted from bond prices are higher than the default probabilities computed from historical data on defaults. Investors require a risk premium to hold defaultable bonds, reflecting imperfect diversification of default risk: if default events of obligors were independent and the number of obligors very large, default risk could be perfectly diversified owing to the law of large numbers. This is the classical framework in insurance theory and in such a competitive market, the computation of bond prices should

follow the “pure premium actuarial rule”, meaning that default probabilities implied from financial markets should equal the “historical default probabilities”. For high quality names, say AAA bonds, defaults are rare events, the number of traded bonds small, and the dependence between default events is larger than for high yield names since extreme macroeconomic factors drive defaults of such firms. Not surprisingly, the ratio of market implied and historical default probabilities is larger for these high quality names compared with speculative bonds (*see Credit Migration Matrices*).

Credit risk measurement usually involves historical default probabilities, while securitization and risk transfer (typically the pricing of CDO tranches) requires the use of market implied probabilities. We will denote by $F_1, \dots, F_n$ the marginal distribution functions associated with the names in the credit portfolio, that can be either historical or market implied depending on the context.

Individual and Collective Models. Let us denote by $\tau_1, \dots, \tau_n$ the default dates of n obligors, by $N_1(t) = 1_{\{\tau_1 \leq t\}}, \dots, N_n(t) = 1_{\{\tau_n \leq t\}}$ the corresponding default indicators for some given time horizon t and by LGD_i the LGD on name i . We assume that the maximum loss is normalized to unity, the aggregate loss on the credit portfolio for time horizon t is then given by:

$$L(t) = \frac{1}{n} \sum_{i=1}^n LGD_i \times N_i(t) \quad (1)$$

Let us remark that we only cope with defaults (or “realized losses”) and not losses due to changes in credit quality of nondefaulted bonds. As can be seen from equation (1), the standard credit risk model is an individual model.

Structural Models and Gaussian Copulas

Inspired by the structural approach of Merton, defaults occur whenever assets fall below a prespecified threshold (see Gupton *et al.* [4], Finger [5], and Kealhofer [3]). In the multivariate case, dependence between default dates ensues dependence between asset price processes. On the other hand, the most commonly used approach states that default dates are associated with a Gaussian copula (Li [6]) (*see Copulas and Other Measures of Dependency*). Thus,

default indicators follow a multivariate probit model. For simplicity, the Gaussian copula model can be viewed as a one period structural model. Consequently, Hull *et al.* [7] show that from a practical point of view, the copula and structural approaches lead to similar loss distributions. Burtshell *et al.* [8] point out that in many cases, the computation of the loss distribution is rather robust with respect to the choice of copula.

For simplicity, we will assume that LGD_i is non-stochastic. We will thus concentrate upon the modeling of dependence between default dates rather than upon the recovery rates. For large credit portfolios, taking into account stochastic LGD has only a small impact on the loss distribution provided that LGD are independent of default indicators. Altman *et al.* [9] analyze the association between default and recovery rates on corporate bonds over the period 1982–2002; they show negative dependence, i.e., defaults are more severe and frequent during recession periods. An analysis of the changes in the loss distributions due to recovery rates possibly correlated with default dates is provided in Frye [10] or in Chabaane *et al.* [11]; this usually results in fatter tails of the loss distributions.

The assumption that default indicators are independent given a low dimensional factor is another key ingredient in credit risk models (see Wilson [12, 13], Gordy [14], Crouhy *et al.* [15], Pykhtin and Dev [16], and Frey and McNeil [17]). This dramatically reduces the numerical complexities when computing loss distributions. While commercial packages usually involve several factors, we will further restrict to the case of a single factor that eases the exposition. This is also the idea behind the regulatory Basel II framework (see Gordy [18]). Thus, in the Gaussian copula (or multivariate probit) (*see Dose–Response Analysis; Potency Estimation; Credit Scoring via Altman Z-Score*) approach, the latent variables associated with default indicators $N_i(t) = 1_{\{\tau_i \leq t\}}, i = 1, \dots, n$ can be written as: $V_i = \rho_i V + \sqrt{1 - \rho_i^2} \bar{V}_i, i = 1, \dots, n$ where $V, \bar{V}_1, \dots, \bar{V}_n$ are independent standard Gaussian variables. The default times are then expressed as: $\tau_i = F_i^{-1}(\Phi(V_i))$ where Φ denotes the Gaussian cdf. In other words default of name i occurs before t if, and only if, $V_i \leq \Phi^{-1}(F_i(t))$. It can be easily checked that default dates are independent given the factor V and that the conditional default probabilities can be written as

$P(\tau_i \leq t|V) = \Phi\left(\frac{\Phi^{-1}(F_i(t) - \rho_i V)}{\sqrt{1 - \rho_i^2}}\right)$, $i = 1, \dots, n$. Let us remark that due to the theory of stochastic orders, increasing any of the correlation parameters $\rho_1, \dots, \rho_n$ leads to an increase in the dependence of the default times $\tau_1, \dots, \tau_n$ with respect to the supermodular order. The corresponding copula of default times is known as the *one-factor Gaussian copula*.

Clearly, determining the correlation parameters is not an easy task especially when the number of names involved in the credit portfolio is large. Extra simplicity consists in grouping names in homogeneous portfolios with respect to sector or geographical region. We refer to Gregory and Laurent [19] for a discussion of this approach. The easier to handle approach consists in assuming some kind of homogeneity at the portfolio level. For instance, we can assume that the correlation parameter is name independent. This is known as the *flat correlation* approach. This both underlies the computations of risk measures in the Basel II agreement framework (see **From Basel II to Solvency II – Risk Management in the Insurance Sector; Credit Migration Matrices**) and of CDO tranche premiums.

Let us also remark that when the marginal default probabilities are equal, i.e., $F_1(t) = \dots = F_n(t)$, then the default indicators $N_1(t), \dots, N_n(t)$ are exchangeable (see **Imprecise Reliability; Bayesian Statistics in Quantitative Risk Assessment**). Conversely, when the default indicators are exchangeable, one can think of using de Finetti's theorem (see **Subjective Probability**), which states the existence of univariate factor such that the default indicators are conditionally independent given that factor. In other words, for “homogeneous portfolios”, the assumption of a one-dimensional factor is not restrictive. The only assumption to be made is upon the distribution of conditional default probabilities. We refer to Burtschell *et al.* [8, 20] for some discussion of different mixing distributions.

Computation of Loss Distributions

Let us now discuss the computation of aggregate loss distribution in the previous framework. The simplest case corresponds to the previous homogeneous case. We then denote by $\tilde{p}_t = P(\tau_i \leq t|V)$ the unique conditional default probability for time horizon t . According to the homogeneity assumption, the probability of k defaults within the portfolio

($k = 0, 1, \dots, n$) or equivalently the probability that the aggregate loss $L(t)$ equals $\frac{k}{n}$ can be written as $\binom{n}{k} \int \tilde{p}^k (1 - \tilde{p})^{n-k} \nu_t(d\tilde{p})$ where ν_t is the distribution of $\tilde{p}_t$. In other words, the loss distribution is a binomial mixture. Let us denote by φ the density function of a standard Gaussian variable. We can equivalently write $P(L(t) = \frac{k}{n}) = \int_{\mathbb{R}} P(\tau_i \leq t|v)^k (1 - P(\tau_i \leq t|v))^{n-k} \varphi(v) dv$, which can be computed numerically according to a Gaussian quadrature.

An interesting feature of the above approach is the simplicity of distributions for large homogeneous portfolios. According to de Finetti's theorem, the aggregate loss $L(t)$ converges almost surely and in mean to $\tilde{p}_t = \Phi\left(\frac{\Phi^{-1}(F(t) - \rho V)}{\sqrt{1 - \rho^2}}\right)$ as the number of names tends to infinity. In the credit risk context this idea was first put in practice by Vasicek [21]. It is known as the *large portfolio approximation*. Further asymptotic developments such as the saddlepoint expansion techniques have been used, starting from Martin *et al.* [22].

Let us now consider the computation of the aggregate loss distribution for a given time horizon within the Gaussian copula framework without any homogeneity or asymptotic approximation. This is based on the computation of the characteristic function of the aggregate loss. We further denote by $\varphi_{L(t)}(u) = E[\exp(iuL(t))]$. According to the conditional independence upon the factor V , we can write $\varphi_{L(t)}(u) = \int_{\mathbb{R}} \left(\prod_{j=1}^n \left(1 + \Phi\left(\frac{\Phi^{-1}(F_j(t) - \rho_j + nu)}{\sqrt{1 - \rho_j^2}}\right) \left(e^{iu \frac{LGD_j}{n}} - 1 \right) \right) \right) \varphi(v) dv$. The previous integral can be easily computed by using a Gaussian quadrature. Let us also remark that the computation of the characteristic function of the loss can be adapted without extra complication when the losses given default LGD_i are stochastic but (jointly) independent together with the latent variables $V, V_1, \dots, V_n$. The computation of the loss distribution can then be accomplished owing to the inversion formula and some fast Fourier transform algorithm (see Laurent and Gregory [23]). A slightly different approach, based on recursions is discussed in Andersen *et al.* [24]. The previous approach is routinely used for portfolios of approximately 100 names.

Let us remark that since the standard assumption states that the copula of default times is Gaussian, we are able to derive aggregate loss for different time

horizons. This is of great practical importance for the computation of CDO tranche premiums, which actually involves loss distributions over different time horizons.

Conclusion

The computation of the loss distributions is of great importance for credit risk assessment and the pricing of credit risk insurance.

Standard risk measures involved in the credit field, such as the value at risk and the expected shortfall, can be easily derived from the loss distribution. At this stage, let us remark that the risk aggregation of different portfolios is not straightforward when these do not share the same factor. The Basel II framework makes the strong assumption that the same factor applies to all credit portfolios. Moreover, due to some large portfolio approximation, the aggregate losses are comonotonic; they only depend upon the unique factor. In that simplified framework, value at risk and expected shortfall of the wholly aggregate portfolio are obtained by summation owing to the comonotonic additive property of the previous risk measures. This departs from the Solvency II framework in insurance.

As far as the pricing of credit insurance and more precisely of CDO tranches is involved, one needs to compute stop-loss premiums $E[(L(t) - k)^+]$ for different time horizons t and “attachment points” k . According to the semianalytical techniques detailed above, these computations can be carried out very quickly and have now become the standard framework used by market participants.

References

- [1] Tavakoli, J. (2003). *Collateralized Debt Obligations & Structured Finance: New Developments in Cash and Synthetic Securitization*, John Wiley & Sons.
- [2] De Malherbe, E. (2006). A simple probabilistic approach to the pricing of default swap covenants, *Journal of Risk* **8**(3), 1–29.
- [3] Kealhofer, S. (2003). Quantifying credit risk I: default prediction, *Financial Analysts Journal* **59**(1), 30–44.
- [4] Gupton, G., Finger, C. & Bahia, M. (1997). *CreditMetrics*, Technical Document, J.P. Morgan.
- [5] Finger, C.C. (2001). The one-factor CreditMetrics model in the new Basel capital accord, *RiskMetrics Journal* **2**(1), 9–18.

- [6] Li, D. (2000). On default correlation: a copula function approach, *Journal of Fixed Income* **9**(4), 43–54.
- [7] Hull, J., Pedrescu, M. & White, A. (2006). *The Valuation of Correlation-Dependent Credit Derivatives Using a Structural Model*, Working Paper, University of Toronto.
- [8] Burtshell, X., Gregory, J. & Laurent, J.-P. (2005). *A Comparative Analysis of CDO Pricing Models*, Working Paper, ISFA Actuarial School, Université Lyon 1 & BNP Paribas.
- [9] Altman, E.I., Brooks, B., Resti, A. & Sironi, A. (2004). The link between default and recovery rates: theory, empirical evidence, and implications, *Journal of Business* **78**, 2203–2228.
- [10] Frye, J. (2000). Collateral damage, *Risk* **13**(4), 91–94.
- [11] Chabaane, A., Laurent, J.-P. & Salomon, J. (2005). Credit risk assessment and stochastic LGDs, in *Recovery Risk: The Next Challenge in Credit Risk Management*. E. Altman, A. Resti & A. Sironi, eds, Risk Publications.
- [12] Wilson, T. (1997). Portfolio credit risk I, *Risk* **10**(9) 111–117.
- [13] Wilson, T. (1997). Portfolio credit risk II, *Risk* **10**(10) 56–61.
- [14] Gordy, M. (2000). A comparative anatomy of credit risk models, *Journal of Banking and Finance* **24**, 119–149.
- [15] Crouhy, M., Galai, D. & Mark, R. (2000). A comparative analysis of current credit risk models, *Journal of Banking and Finance* **24**, 59–117.
- [16] Pykhtin, M. & Dev, A. (2002). Analytical approach to credit risk modelling, *Risk* **15**, S26–S32.
- [17] Frey, R. & McNeil, A. (2003). Dependent defaults in models of portfolio credit risk, *Journal of Risk* **6**(1), 59–92.
- [18] Gordy, M. (2003). A risk-factor model foundation for ratings-based bank capital rules, *Journal of Financial Intermediation* **12**(3), 199–232.
- [19] Gregory, J. & Laurent, J.-P. (2004). In the core of correlation, *Risk* **17**(10), 87–91.
- [20] Burtshell, X., Gregory, J. & Laurent, J.-P. (2005). *Beyond the Gaussian Copula: Stochastic and Local Correlation*, Working Paper, ISFA Actuarial School, Université Lyon 1 & BNP Paribas.
- [21] Vasicek, O. (2002). Loan portfolio value, *Risk* **15**(7), 160–162.
- [22] Martin, R., Thompson, K. & Browne, C. (2001). Taking to the saddle, *Risk* **14**(6), 91–94.
- [23] Laurent, J.-P. & Gregory, J. (2005). Basket default swaps, CDOs and factor copulas, *Journal of Risk* **7**(4), 103–112.
- [24] Andersen, L., Sidenius, J. & Basu, S. (2003). All your hedges in one basket, *Risk Magazine* **16**(11), 67–72.

Further Reading

Frye, J. (2003). A false sense of security, *Risk* **16**(8), 63–67.

JEAN-PAUL LAURENT

Credit Scoring *via* Altman Z-Score

Altman Z-Score Model

The recent interest of estimating probability of default for credit products is mainly caused by the Basel II Accord; see Basel Commission on Banking Supervision [1–3]. Such estimation is heavily related to the technique of credit scoring which has a long history in finance. Beaver [5, 6], one of the pioneers in the field, proposes the seminal idea of using the financial ratios for prediction. Beaver is able to discriminate failed firms from non-failed firms 5 years ahead. Altman [7] improves the idea by combining several relevant ratios into a vector of predictors and is the first researcher applying the method of multiple discriminant analysis (MDA). His method becomes the renowned Altman Z-Score, which comprises the main focus in this chapter.

MDA was first introduced by Fisher [8] in statistics to separate multivariate normal populations. Since the separation criterion is a linear function of the multivariate characteristics, MDA is also called the *linear discriminant analysis* (LDA). Applying MDA to predict bankruptcy is equivalent to modeling the joint relationship between the event of bankruptcy of a firm at a specific period and the corresponding firm characteristics (financial ratios) collected immediately before that period.

Specifically, let I be the Bernoulli random variable with parameter π that is the probability of bankruptcy next year. π is called the *prior probability* of bankruptcy because that is simply a rough guess without using any information. Let $X = (X_1, \dots, X_m)$ be the financial characteristic of the firm this year. The model of MDA assumes $(X|I = 1) \sim N(\mu_B, \Sigma)$ and $(X|I = 0) \sim N(\mu_S, \Sigma)$ where Σ is the common variance–covariance matrix, μ_B and μ_S are the mean vectors for bankruptcy and survival, respectively. Such model gives the posterior probability of bankruptcy in the following form:

$$\Pr\{I = 1|X = x\} = \frac{f_B(x)\pi}{(1 - \pi)f_S(x) + \pi f_B(x)} \quad (1)$$

where $f_B(x)$ and $f_S(x)$ are the probability density functions of $N(\mu_B, \Sigma)$ and $N(\mu_S, \Sigma)$. The posterior

probability can be further simplified to

$$\Pr\{I = 1|X\} = g(-[a_1X_1 + \dots + a_mX_m + \gamma]) \quad (2)$$

where

$$g(t) = \frac{1}{1 + \exp\{-t\}} \quad \text{for any real number } t \quad (3)$$

$$(a_1, \dots, a_m) = (\mu_S - \mu_B)^T \Sigma^{-1}$$

and

$$\begin{aligned} \gamma = \log(1 - \pi) - \log(\pi) + \frac{1}{2}\mu_B^T \Sigma^{-1} \mu_B \\ - \frac{1}{2}\mu_S^T \Sigma^{-1} \mu_S \end{aligned} \quad (4)$$

Note that $g(t)$ is increasing in t . Therefore, for fixed, $\Pr\{I = 1|X\}$ is a decreasing function of the linear combination $a_1X_1 + \dots + a_mX_m$. Such linear combination is called the *Z-Score* by Altman [7]. That is, the higher the Z-Score, the less likely the firm will default.

In Altman [7], π is set to be 0.5 because there is no extra prior information of bankruptcy other than X . The other parameters of the model, Σ , μ_B , and μ_S , are estimated by the method of maximum likelihood from a data set, which is a matched sample of bankrupt and nonbankrupt firms. In fact, the bankrupt group was all manufacturers that filed bankruptcy petitions under Chapter 10 from 1946 through 1965. It is clear that a 20-year sample period is not the best choice since the financial ratios X_j do shift over time. A more appropriate sample should consist of the firm characteristics in the same time period t and the bankruptcy/survival indicators in the following period $(t + 1)$. However, such an ideal sample was not available because of data limitations at that time. To remedy the inhomogeneity of the bankrupt group owing to industry and size differences, Altman [7] carefully selects the nonbankrupt control firms such that the firms in the overall data set were stratified by industry and by size, with asset size range from \$1 to \$25 million (see Altman [7, pp. 593–594] for details). Altman [7] begins the MDA with $m = 22$ potential variables and finds that the best bankruptcy prediction is given by the Z-Score from the following five variables:

X_1 : Working capital $\div$ total assets

X_2 : Retained earnings $\div$ total assets

X_3 : Earnings before interest and taxes $\div$ total assets

X_4 : Market value of equity $\div$ book value of total liabilities

X_5 : Sales $\div$ total assets

Here, “total assets” is referred to the sum of all tangible assets (excluding the intangibles) of the firm while the working capital is the difference between current assets and current liabilities. Therefore, X_1 measures the net liquid assets of the firm relative to the total capitalization.

For X_2 , “retained earnings” is the total amount of reinvested earnings and/or losses of a firm over its entire life. It is also known as the *earned surplus*. Note that adjustments should be applied to handle the case of substantial reorganization and/or stock dividend. It is easy to see that X_2 measures the cumulative profitability over the life of the company. Besides, X_2 can also be viewed as a measure of the leverage of a firm. Firms with high X_2 levels financed their assets through retention of profits; they have not utilized much debt. Therefore, X_2 highlights the use of either internally generated funds for growth (low-risk capital) or other people’s money (OPM) – high-risk capital.

X_3 is a measure of the productivity of the firm’s assets, independent of any tax or leverage factors. Since a firm’s ultimate existence is based on the earning power of its assets, this ratio appears to be particularly appropriate for studies dealing with credit risk.

It should be noted that the market value of equity in X_4 is the total market value of all shares of preferred stock and common stock while liabilities include both current and long-term obligations. The measure X_4 shows how much the firm’s assets can decline in value (measured by market value of equity plus debt) before the liabilities exceed the assets and the firm becomes insolvent. Last but not least, X_5 is the capital turnover ratio and it quantifies the sales-generating ability of the firm’s assets.

From the selected data set and the chosen variables, Altman [7] estimates the Z-Score in the form of:

$$Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5 \quad (5)$$

Although γ in equation (2) could be determined by the method of maximum likelihood and the

corresponding posterior probability could be used to predict bankruptcy, Altman [7] investigates the range of Z and realizes that all firms with $Z < 1.81$ in the study defaulted and all firms with $Z > 2.675$ in the study survived. Two decision rules are derived from such an observation:

1. Predict that the firm is going to bankrupt if $Z < 2.675$. Otherwise, the firm is going to survive.
2. Same as rule (i) but using 1.81 as the cutoff.

Note that although the Z-Score above is computed using Fisher’s approach of LDA and the method of maximum-likelihood estimation, the above decision rules and cutoffs are different from the original Fisher’s proposal that solely depends on the conditional probability $\Pr\{I = 1|X\}$. Moreover, the performance of the above Z-Score is analyzed by Altman and Hotchkiss [4, p. 244] for some data observed after 1968. They examine 86 distressed companies from 1969 to 1975, 110 bankrupts from 1976 to 1995, and 120 bankrupts from 1997 to 1999. Using a cut-off score of 2.675, the successful rates of correct type I bankruptcy prediction (those firms classified as bankrupt will default in the next period of time) are 82, 85, and 94%, respectively. However, the rates of type II error (classifying the firm as distressed when it does not go bankrupt or default on its obligations) could be as large as 25%. Using rule 2 could reduce the rates of type II error although the overall error resulted from rule 1 is smaller.

To reduce both type I and type II errors, Altman and Rijken [9] suggest applying log transformation on the retained earnings/total assets and equity/debt ratios. They argue that the high error rates are caused by the fact that US firms are far more risky today than in the past. This higher risk is manifested in the deterioration of a number of financial indicators in the Z-Score model, particularly the retained earnings/total assets and equity/debt ratios. The log function scaled these two variables to be more comparable to the other variables.

Private Firms and Emergent Markets

For private firms, Altman [10] suggests replacing X_4 by the ratio of book value of equity to the total equity and he obtains the following Z-Score:

$$Z' = 0.00717X_1 + 0.00847X_2 + 0.03107X_3 + 0.0042X_4 + 0.998X_5 \quad (6)$$

with the lower and upper thresholds being 1.23 and 2.90.

Since X_5 (the capital turnover ratio) is more industry sensitive than the others, it may be more appropriate to estimate the Z-Score model without the corresponding term under some situations. For example, Altman *et al.* [11] evaluate the Mexican firms that have issued Eurobonds denominated in US dollars by using such a model. They also set X_4 as the book value of equity in this case. See Altman and Hotchkiss [4, Chapter 12] for more details about the application to the emergent markets.

Concluding Remarks

In summary, while MDA is a model for the joint relationship between I (the indicator of default) and X (the firm characteristic), the key prediction equation (2) is phrased in terms of a conditional distribution, which can be estimated by using the methodology of logistic regression (*see Risk in Credit Granting and Lending Decisions: Credit Scoring; Logistic Regression*). The advantage of applying the technique of logistic regression is that the empirical results are theoretically sound even if X is not jointly normally distributed, given I . Although Efron [12] states that statistically MDA should outperform the logistic regression if the normality of X holds, one can easily check that the explanatory variables in Altman [7] and other Z-Score applications are rarely normal.

The methodology of logistic regression can be phrased in terms of Z-Score in the following form, given $X = (X_1, \dots, X_m)$,

$$Z = \gamma + a_1 X_1 + \dots + a_m X_m + \varepsilon \quad (7)$$

where ε is a random variable whose cumulative distribution function is given by the logistic function g . I is defined as $I\{Z < 0\}$ such that

$$\begin{aligned} \Pr\{I = 1|X\} &= E[I|X] \\ &= \Pr\{\gamma + a_1 X_1 + \dots + a_m X_m + \varepsilon < 0\} \\ &= \Pr\{\varepsilon < -\gamma - a_1 X_1 - \dots - a_m X_m\} \\ &= g(-[\gamma + a_1 X_1 + \dots + a_m X_m]) \quad (8) \end{aligned}$$

The maximum-likelihood estimate of $(\gamma, a_1, \dots, a_m)$ can be efficiently computed via Fisher scoring method. Moreover, the asymptotic inference could be

implemented easily. Similar computations could also be performed for the case of ε being standard normal, which corresponds to probit regression. See McCullagh and Nelder [13] for computational and statistical details of both models. It should be noted that the assumption of probit regression is substantially different from that of MDA even though they both involve normality. The MDA assumes joint normality of the $(X_1, \dots, X_m)$ while the probit regression assumes " $(X_1, \dots, X_m)$ " is conditionally normal. There is no obvious equivalence between these two models.

The logistic regression model has been applied to bankruptcy prediction by Ohlson [14]. Other statistical techniques include quadratic discriminant analysis as in Altman *et al.* [15], recursive partitioning in Frydman *et al.* [16], neural networks in Altman *et al.* [17], and many others in various chapters of Trippi and Turban [18].

In addition to the statistical approach, the Moody's KMV also developed a way of estimating probability of default (PD) known as EDF^{TM} (expected default frequency), which is closely related to the financial option pricing theory. See Altman and Hotchkiss [4, p. 252] for more discussion on EDF.

References

- [1] Basel Commission on Banking Supervision (1999). *Credit Risk Modeling: Current Practices and Applications*, Bank for International Settlements, Basel.
- [2] Basel Commission on Banking Supervision (2001). *The Basel Capital Accord*, Bank for International Settlements, Basel.
- [3] Basel Commission on Banking Supervision (2004). *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*, Bank for International Settlements, Basel.
- [4] Altman, E. & Hotchkiss, E. (2006). *Corporate Financial Distress and Bankruptcy*, 3rd Edition, John Wiley & Sons, New York.
- [5] Beaver, W. (1966). "Financial ratios as predictors of failures", in empirical research in accounting: selected studies, *Journal of Accounting Research* **4**, 71–111.
- [6] Beaver, W. (1968). Alternative accounting measures as predictors of failure, *Accounting Review* **43**, 113–122.
- [7] Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance* **23**, 589–609.
- [8] Fisher, R.A. (1938). The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**, 179–188.

- [9] Altman, E. & Rijken, H. (2004). How rating agencies achieve rating stability, *Journal of Banking and Finance* **28**, 2679–2714.
- [10] Altman, E. (1993). *Corporate Financial Distress and Bankruptcy*, 2nd Edition, John Wiley & Sons, New York.
- [11] Altman, E., Hartzell, J. & Peck, M. (1995). *Emerging Market Corporate Bonds – A Scoring System*, Salomon Brothers Emerging Market Bond Research. Reprinted in *The Future of Emerging Market Flows*, E. Altman, R. Levich & J.P. Mei, eds, Kluwer Publishing, Holland, 1997.
- [12] Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis, *Journal of American Statistical Association* **70**, 892–898.
- [13] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition, Chapman & Hall, London.
- [14] Ohlson, J.A. (1980). Financial ratios and the probabilistic prediction of bankruptcy, *Journal of Accounting Research* **18**, 109–131.
- [15] Altman, E., Haldeman, R. & Narayanan, P. (1977). ZETA analysis, a new model for bankruptcy classification, *Journal of Banking and Finance* **1**, 29–54.
- [16] Frydman, H., Altman, E. & Kao, D.L. (1985). Introducing recursive partitioning analysis for financial analysis: the case of financial distress, *Journal of Finance* **40**, 269–291.
- [17] Altman, E., Marco, G. & Varetto, F. (1994). Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks, *Journal of Banking and Finance* **3**, 505–529.
- [18] Trippi, R.R. & Turban, E. (1993). *Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real-world Performance*, Probus Publishers, Chicago.

EDWARD I. ALTMAN, NGAI H. CHAN AND
SAMUEL PO-SHING WONG

Credit Value at Risk

Definition and Computation

According to Jorion [1], banks allocate roughly 60% of their regulatory capital to credit risks, 15% to market risks, and 25% to operational risks. Thus, credit value at risk deserves much attention because it determines the adequate protection the banks need to prepare for possible adverse conditions.

Consider a credit portfolio (*see Comonotonicity; Credit Scoring via Altman Z-Score; Statistical Arbitrage*) that consists of default-sensitive instruments such as lines of credit, corporate bonds, and government bonds. The corresponding credit value-at-risk (VaR) (*see Credit Risk Models; Default Risk; Extreme Value Theory in Finance*), is the minimum loss of next year if the worst 0.03% event happens. In another words, 99.97% of the time the loss will not be greater than VaR. Mathematically, if L is the loss of the portfolio caused by default events next year, then

$$\Pr\{L > \text{VaR}\} = 0.03\% \text{ or } \Pr\{L \leq \text{VaR}\} = 99.97\% \quad (1)$$

VaR is equivalent to the 99.97 th percentile of L . Note that the credit VaR is measured at the time span of 1 year and is different from the 10-day convention adopted by the market VaR. Moreover, instead of using the worst 1% event as the benchmark of the market VaR, 0.03% is chosen because it is a rating agency standard of granting an AA credit rating (see Saunders and Allen [2, p. 207]).

Single Instrument

To study the computation of VaR for a credit portfolio, let us start with the case of a single instrument. The loss of a single instrument can be decomposed into three components: the default probability (*see Options and Guarantees in Life Insurance; Risk Measures and Economic Capital for (Re-)insurers; Informational Value of Corporate Issuer Credit Ratings; Copulas and Other Measures of Dependency*) of the obligor, a loss fraction, which is called the *loss given default*, and the exposure at

default (*see Default Risk*), which are denoted by PD, LGD, and EAD, respectively. That is,

$$\begin{aligned} L &= X \times \text{LGD} \times \text{EAD} \text{ and } PD = \Pr\{X = 1\} \\ &= 1 - \Pr\{X = 0\} \end{aligned} \quad (2)$$

where L is the default loss and X is the Bernoulli random variable with $\{X = 1\}$ being the event of default. EAD is nonrandom if the instrument is a bond or a line of credit. In particular, for the latter case, EAD would be the maximum amount that can be drawn because the borrowers are most likely to exhaust all possible resources before default. However, there are exceptions such as the counterparty default risk on credit derivative for which EAD is the replacement value of the derivative at the time of default (Bluhm *et al.* [3, Section 12.4]). For the sake of simplicity, EAD is assumed to be nonrandom in the subsequent discussion.

LGD is the portion of EAD that gives real negative impact in case of default. LGD is usually less than 1 because many default obligors are originally backed up by some mortgages or securities and a proportion of loss could be recovered by selling the guarantees. The recovered portion of default loss is commonly expressed as a percentage of EAD. Such a percentage is called the *recovery rate* and is related to LGD by

$$\text{Recovery rate} = 1 - \text{LGD} \quad (3)$$

The magnitude of the recovery rate is tied to the value of the collateral properties at or after the time of default. Also, the recovery rate depends on the nature of the instrument: only the loss on the principal can be claimed, but not the loss on coupon interest, when the debtor of a corporate coupon bond bankrupts (see Allen [4, p. 329]). Moreover, the seniority of the debt holder in claiming loss significantly affects the recovery rate. In fact, Moody's published figures for 1982–2003 show that senior secured debt holders had an average recovery rate of 51.6% while junior subordinated debt holders had an average recovery rate (*see Compensation for Loss of Life and Limb; Credit Migration Matrices*) of only 24.5% (Hull [5, p. 483]).

An important empirical fact about PD and LGD: they are positively correlated, that is, PD and the recovery rate are negatively correlated. Hamilton *et al.* [6] fit the following regression to the annual

default data from 1982 to 2003 and obtain $R^2 = 0.6$.

$$\begin{aligned} \text{Average recovery rate} &= 50.3\% - 6.3 \\ &\times \text{average default rate (4)} \end{aligned}$$

Since the average recovery rate can be predicted from the PD by using the above regression, the task of evaluating VaR is reduced to the determination of PD, which can be accomplished by the following two approaches:

1. Use Merton's firm-value model (*see Default Correlation*) to back out the risk-neutral PD from the observed equity price. Then, convert the risk-neutral PD to the physical PD by technology similar to the *Expected Default Frequency*TM (or simply EDF) developed by KMV Corporation. See Crosbie [7] or Hull [5, pp. 489–491].
2. Use the credit ratings of external agencies and set the PD for the company of interest to the PD of the corresponding credit class.

Note that although the risk-neutral (*see Equity-Linked Life Insurance; Optimal Risk-Sharing and Deductibles in Insurance; Options and Guarantees in Life Insurance; Mathematical Models of Credit Risk*) PD of a corporate obligor could also be obtained by using the difference between its bond yield and a "risk-free" bond yield (Hull [5, p. 484]), there is no easy way to convert such implied PD into physical PD.

Portfolio

For a credit portfolio of more than one instrument, the default loss is the sum of losses caused by the default events of the various obligors involved. The main issue in computing VaR for a credit portfolio is that the joint default probability for two obligors in general does not follow the law of independence. In fact, companies in the same industry or country tend to default together within a relatively short period of time. This tendency is also known as the *credit concentration, default concentration, or default correlation*.

Moody's KMV and CreditMetrics both try to estimate the joint default probability by using the technology of the copula. To illustrate the idea, let us consider a portfolio of two instruments whose Bernoulli random variables for default are denoted

Table 1 Annual probability transition matrix of credit ratings

Current rating	Next rating		
	A	B	Default
A	0.95	0.025	0.025
B	0.05	0.9	0.05
Default	0	0	1

by X_1 and X_2 . Suppose the first and the second instruments are rated as A and B, respectively, and the corresponding yearly transition matrix in Table 1.

From a recent sample of the equity returns of the corresponding obligors, we compute the means (m_1 and m_2), the standard deviations (s_1 and s_2), and the correlation (ρ). We then simulate $(\tilde{R}_1, \tilde{R}_2)$ from a bivariate normal distribution whose mean vector is (m_1, m_2) , the variances are s_1^2 and s_2^2 , and the covariance is $\rho \times s_1 \times s_2$. $\tilde{X}_1$ is set to be 1 if $\tilde{R}_1 < m_1 - 1.96 \times s_1$. Otherwise, $\tilde{X}_1$ is set to be 0. Note that $N(-1.96) = 0.025$ and is equal to the probability of default of class A to which the first obligor belongs. (N is the standard normal distribution function.) Similarly, since $N(-1.645) = 0.05$ and is equal to the probability of default for class B, $\tilde{X}_2$ is set to be 1 if $\tilde{R}_2 < m_2 - 1.645 \times s_2$. Otherwise, $\tilde{X}_2$ is set to be 0. The simulation is repeated many times and each simulated default loss $\tilde{L} = \tilde{X}_1 \times LGD_1 \times EAD_1 + \tilde{X}_2 \times LGD_2 \times EAD_2$ is stored where LGD_i and EAD_i are the loss given default and exposure at default of the i th instrument. VaR is computed as the 99.97th percentile of $\tilde{L}$.

Simulating $(\tilde{R}_1, \tilde{R}_2)$ is called the *method of Gaussian Copula*. As noted before, this method is adopted by both KMV and CreditMetrics with some variations (Bluhm *et al.* [3, Section 12.3]). Please also see Cherubini *et al.* [8] for other copula approaches.

If the number of instruments in the portfolio is large, the number of correlations that needs to be estimated is huge. For example, if there are 100 instruments involved, we have to estimate 100 means, 100 standard deviations, but $100 \times 99/2 = 4950$ correlations. To keep a small estimation error, an extensive history of equity returns is required. However, it is quite unlikely that the properties of the obligors are invariant over a long time. An effective way to handle such overparametrization issue is to employ a factor model. Suppose $R_1, \dots, R_J$ are the equities return for a credit portfolio of J instruments where J is large. Let $F_1, \dots, F_M$ be the factors where

M is relatively small. The following regression is fitted for each R_j :

$$R_j = a_j + b_{j1}F_1 + \cdots + b_{jM}F_M + e_j \quad (5)$$

e_j is a normal random variable with zero mean and standard deviation σ_j . Also, e_j 's are assumed to be independent of each other. $\{F_i\}_{i=1}^M$ could be the industry factor, country factor, regional factor, and any other relevant exogenous factor. The above formulation reduces the number of parameters to be estimated to $100 \times 5 = 500$ for the case of $J = 100$ and $M = 3$. Therefore, the number of observed equity returns required for good estimation is greatly reduced.

After all parameters are estimated, one can simulate $\{\tilde{R}_j\}_{j=1}^J$ by using the factor model, that is, sampling $\tilde{e}_j$ from independent normal distribution whose mean is 0 and standard deviation is σ_j and computing $\tilde{R}_j$ by

$$\tilde{R}_j = a_j + b_{j1}F_1 + \cdots + b_{jM}F_M + \tilde{e}_j \quad (6)$$

Similar to the previous case, $\tilde{X}_j$ is set to be 0 or 1 by matching the observed transition probability to default. VaR is computed as the corresponding quantile of the simulated portfolio default loss $\tilde{L} = \tilde{X}_1 \times LGD_1 \times EAD_1 + \cdots + \tilde{X}_J \times LGD_J \times EAD_J$.

Stress Testing

Stress testing (*see Operational Risk Modeling; Reliability of Consumer Goods with "Fast Turn Around"*) is the procedure of checking the robustness of VaR under different hypothetical changes to the current situation. Typical examples include perturbation of the model parameters, economic downturn of the country or region, deterioration of the business environment of industry, or the downgrade of specific obligors' credit profiles. One can check the changes in VaR under different scenarios. Another equivalent way is to fix VaR and observe how the tail area of L is affected. A systematic account can be found in the Bank of International Settlement document of "Stress testing at major financial institutions: survey results and practice", which was prepared by their Committee on the Global Financial System in 2005.

Allen [4, p. 350] highlights the country risk factor and suggests this factor has to be treated with special care. Such risks can manifest in many forms: they

can range from war and revolution to imposing defensive exchange controls by the government. They cause all firms, individuals, and government bodies in that country to be rated as poor in credit or even as defaulters. Scenario analysis of VaR can be performed by subjective assessment and using data from similar past political disruption in other countries. See Calverley [9] for details.

Implementation Details

The above presentation is very brief because we only want to show the basic idea of VaR without going into the details. In this last section, we aim at providing a hint of complexity in the reality:

- Many commercial banks are holding credit portfolios that consist of illiquid traditional loans. Estimation of the marginal PDs and the default correlations rely on internal rating systems that are very different from the approach we presented above. See Allen *et al.* [10, Chapter 4].
- Although many banks have a strong desire to apply credit VaR to both trading and loan books in an integrated manner, some of the cumbersome barriers are the differences in accounting treatment, variation of technology platforms, and the illiquid factor noted above. The banks may involve fundamental changes in organizational structure in order to implement consistent integrated risk-management systems.
- The factor model in the section titled "Portfolio" assumes that all equity returns and factors can be collected and organized in nice rectangular forms that are ready for regression routines. However, in reality, there are values missing everywhere. One of the reasons is that the sampling frequencies of the factors and those of equity returns are in general different. Another issue is that some of the companies involved in the portfolio could have become public very recently and the equity return may not be available before initial public offering (IPO). Statistical techniques, such as EM algorithm of Dempster *et al.* [11] and data augmentation algorithm of Tanner and Wong [12], can be employed to impute the missing values and estimate the model parameters.
- Similar to market VaR, backtesting is one of the goals to be achieved in addition to stress testing. However, the time span of credit VaR is typically

1 year and it is hopeless to collect enough historic credit loss data for validation purpose. Lopez and Saidenberg [13] suggest backtesting by cross-sectional simulation, which is essentially a variation of bootstrap, i.e., evaluation based upon resampled data. See Efron and Tibshirani [14].

References

- [1] Jorion, P. (2000). *Value at Risk: The New Benchmark for Managing Financial Risk*, 2nd Edition, McGraw-Hill.
- [2] Saunders, A. & Allen, L. (2002). *Credit Risk Measurement*, John Wiley & Sons.
- [3] Bluhm, C., Overbeck, L. & Wagner, C. (2003). *An Introduction to Credit Risk Modeling*, Chapman & Hall/CRC.
- [4] Allen, S. (2004). *Financial Risk Management: A Practitioner's Guide to Managing Market and Credit Risk*, John Wiley & Sons.
- [5] Hull, J.C. (2006). *Options, Futures and Other Derivatives*, 6th Edition, Pearson.
- [6] Hamilton, D.T., Varma, P., Ou, S. & Cantor, R. (2004). *Default and Recovery Rates of Corporate Bond Issuers*, Moody's Investor's Services.
- [7] Crosbie, P. (1999). *Modeling Default Risk*, KMV Corporation.
- [8] Cherubini, U., Luciano, E. & Vecchiato, W. (2004). *Copula Methods in Finance*, John Wiley & Sons.
- [9] Calverley, J. (1985). *Country Risk Analysis*, Butterworths.
- [10] Allen, L., Boudoukh, J. & Saunders, A. (2004). *Understanding Market, Credit and Operational Risk: The Value-at-Risk Approach*, Blackwell Publishing.
- [11] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
- [12] Tanner, M.A. & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association* **82**, 528–40.
- [13] Lopez, J.A. & Saidenberg, M.R. (2000). Evaluating credit risk models, *Journal of Banking and Finance* **24**, 151–165.
- [14] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.

Related Articles

Default Correlation

SAMUEL PO-SHING WONG

Cross-Species Extrapolation

Overview and Conceptualization

Extrapolation may be thought of as an inferential process by which unknown information is estimated from known data. Originally, mathematical usage involved any method of finding values outside the range of known terms or series by calculation [1–3] and was extended to a nonmathematical sense [4]. Earlier the usage of the term had a negative connotation, implying unwarranted speculation, but the term came to be used widely in scientific reasoning [5].

Cross-species extrapolation (CSE) has a rich history in biology and its validity underlies the majority of current research in human health in which nonhumans or cell systems are used to model disease or injury. In toxicology, CSE is used in the effort to estimate the effects of toxicants in one species (the target species), usually human, from knowledge of the effects of the toxicant in another species (the source species) (*see Cancer Risk Evaluation from Animal Studies*).

CSE may be conceptualized as three components: (a) *dose estimation*: given exposure to a toxicant, estimate the magnitude of the internal dose in the target species from the source species by comparative pharmacokinetics; (b) *effects prediction*: given internal dose in the target species, estimate the magnitude of effect from the source species by comparative biology; and (c) *quantification*: make the estimates numerically, bounded by confidence limits.

Dose Estimation

Principles

With few exceptions, effects of a toxicant are produced by some function of the dose of the toxicant and/or one or more of its metabolites at the organ being impacted [6, 7]. This “internal dose” is the result of the absorption, distribution, biotransformation, and elimination of the dose applied to the subject. These are determined by physical processes and described by pharmacokinetic equations. These processes can be described by a physiologically based pharmacokinetic (PBPK) model with

appropriate parameters and used to estimate the amounts of toxicants delivered to tissues, their biotransformation, and elimination [8].

Given the knowledge of applied and internal dose for some toxicant in the source species, the goal of PBPK modeling is to estimate internal dose in the target species where it may not be measured. Once a PBPK model is developed and evaluated against data in the source species, then it is presumed, in the simplest case, to be a matter of rescaling the parameters of the model to appropriate values for the target species. It is frequently assumed that the best estimate of parameters is directly proportional to body surface area [9]. Surface area may be approximated by taking body weight to the 0.75 power [10]. Such “allometric scaling” methods have been generalized for many physiological parameters [10, 11].

If the physiology or biotransformation processes differ in principle between the two species, however, then the structure of the PBPK model must be altered in the target species instead of simply rescaling. Assuming that PBPK models of appropriate form with accurate parameters are available, it is possible to estimate internal doses for various organs in the target species, ideally with the help of data from the source species. By this means, one can determine whether a particular effect in the two species from some fixed applied dose differs in magnitude because the internal doses differ or because the affected organs in the two species differ in sensitivity.

Practice

Remarkable accuracy of internal dose estimation can be achieved by PBPK models if applications are restricted to simple cases. One common restriction in practice is to parameterize the model for a static baseline case, i.e., where the organism is in some unchanging physiological state. Parameters, however, vary across individuals and thus it may be desirable to measure important parameters on individual subjects to evaluate and utilize a model (e.g., Benignus *et al.* [12] for carboxyhemoglobin or Pierce *et al.* [13] for toluene). Temporal variation of parameters as the organism changes its activity, enters new environments, ages, etc., can also be described to increase precision of estimation [14, 15] in such cases.

The correctness of the model form depends on knowledge of the relevant physiology and its completeness depends on knowledge and upon choice of detail. See Clark *et al.* [16] for evaluation criteria. Even if the form is correct and detailed enough for the purpose at hand, it must be recognized that parameter values are typically means of observations, which vary over individual subjects, time, and measurement error. Thus it is not usually sufficient to predict only the mean internal dose. The confidence limits in the predicted internal dose must be given.

One of the proposed methods of estimating variance and hence confidence limits is to determine, either analytically [17] or iteratively (by Monte Carlo simulation) [18] the variation in predicted internal dose by combining the known variance/covariances of all of the model parameters. The variances of parameters are sometimes known from experiments, but covariances are usually not known. Assuming zero covariances results in overestimation of the variance in the estimated internal dose. A Bayesian approach (see **Bayesian Statistics in Quantitative Risk Assessment**) to simulation is sometimes taken in which variances and covariances are given *a priori* values by expert consensus and then reestimated from data, presumably improving the final estimate [13].

A difficulty, both conceptual and practical, in estimating variance in the predicted mean internal dose is the complexity of a PBPK model, which, in turn, leads to large, complex simulations or analyses. This difficulty can be mitigated by first performing sensitivity analyses on the PBPK model to determine the most sensitive parameters [19]. Only the most influential parameters and those that influence them then need be included in the Monte Carlo simulation. This procedure does involve arbitrary judgment as to what magnitude of influence is important enough to be included.

Even when the above assessments of a PBPK model's correctness are satisfied, the ultimate criterion of quality is a demonstration that the model predicts correctly in a known case. Criteria for the "correctness" of a model's estimation are frequently implicit, but it is sometimes inferred that the model is verified if estimates from the model fall within the confidence limits of the measured internal dose. It is rare that model predictions are compared to empirical data by appropriate repeated-measures, mixed-model statistics [15].

Predicting Effects of the Internal Dose

Principles

Effect of toxicants has been studied at the whole organism level, at the organ level, or at many reductionistic levels, including cellular, molecular, and genetic. Studies at whole organism and organ levels tend to be the most useful for regulatory purposes such as identifying health effects, determining dose–effect relationships, and conducting benefit–cost analyses. Reductionistic data tend to be useful for constructing theory and explaining effects. They can be used to defend interpolation between observations or predict other effects not previously suspected from more macroscopic approaches.

Mathematical models have been constructed to describe the normal physiology of the body. Such models may be fragmentary, to describe the function of various organs or systems of cells [20], or they may be more complete, approaching models of the entire body physiology, e.g., The Modeling Workshop (<http://physiology.umc.edu/themodelingworkshop/>). More ambitious work is in progress to include reductionistic work from the genome to whole body function [21].

When dose–effect models are constructed from experiments with a source species, then presumably it becomes possible to model observed effects in mechanistic terms. If sufficient physiological information exists about the same organ systems in the target species, then the model may be scaled and appropriately changed in form to predict the effect of the toxicant on target organs. Such rescaling frequently involves estimation of relevant organ volumes and process rates (see discussion, above, of rescaling in PBPK models). Mechanistic approaches to effect estimation have been called *biologically based dose–response (BBDR) models* [22]. This approach would be the usual practice of comparative physiology, but mathematically expressed to give numerical predictions.

Practice

The process of constructing a BBDR model is similar to constructing a PBPK model, but usually more complex. Not only must the normal functioning of a physiological subsystem be known in sufficient detail to quantitatively model it, but the mechanism

by which toxic effects are produced must also be known. It is also necessary to apply the same criteria of quality to the modeling process [19]. For well-studied systems, e.g., certain kinds of cancer, both the physiology and cellular process are comparatively well known [23]. For example, cellular processes have been combined with general physiological and dosimetric information to construct a model for estimating the dose–effect curve of nasal cancer induced by formaldehyde exposure [24, 25]. In addition to good prediction of observed data, the model provided defensible interpolations between baseline (no effects) and the lowest-exposure empirical data and thus refined the probable form of the dose–effect relationship.

Other cases of useful BBDR models might be cited, usually in cancer effects, but there are few such cases. On a less microscopic level, physiological models have been constructed for estimating effects of simultaneous exposure to multiple toxic gases [14, 26] and simultaneous variation in environmental conditions, e.g., temperature, altitude, and physiological gases (<http://physiology.umc.edu/themodelingworkshop/>). The lack of more such models can be attributed in part to insufficient information about (a) normal physiological functioning in many organ systems, e.g., the central nervous system and (b) mechanisms of the toxic effects for many substances.

Quantitative Extrapolation

Mechanistic Approach

In a systematic sense, PBPK and BBDR models are used to adjust the magnitude and form of data from the source species so that they may be directly applied to the target species. These steps assume that physiological knowledge is adequate to make such adjustments. Unfortunately, such knowledge is not available for some organs, e.g., the central nervous system. Even if such limits of knowledge did not exist, the numerical CSE process remains an inferential leap across species that cannot be verified in cases where such an estimate would be most useful, i.e., the case where toxicant effects in humans cannot be measured. Next in importance to the numerical inferences are the confidence limits about the inferences. From the above discussions, it is evident that confidence limits based on the physiological models are

often difficult to estimate. It seems, because of these limitations, that for many physiological systems and for many important toxicants, the PBPK and BBDR approaches to CSE will not prove to be an immediately adequate approach. Therefore, an alternative practice empirically based on CSE, may be useful.

Empirical Approaches

Traditional. The traditional empirical approach can be characterized as (a) experimentally finding an effect of some suspected toxicant in a source species, (b) assuming that the same effect would occur in the target species, and (c) estimating the magnitude of the effect or the probability of its occurrence in the target species. Experiments to assess effects of a toxicant in the source species may range from binary decisions of hazard to full empirical dose–effect studies. The kind of data available from the source species affect the extrapolation strategy.

Decisions about regulation of a toxicant are frequently made from some scalar measurement of effect that is intended to characterize the highest permissible exposure or the lowest exposure to produce detectable effects. Among these, the benchmark dose [27] (*see Benchmark Dose Estimation*) is the more statistically informed. Making the inferential leap to the target species is usually accompanied by assignment of so-called between species “uncertainty factors” [28] that are intended to adjust the magnitude of effect observed in the source species to protect the target species. Uncertainty factors can take on various magnitudes, often depending upon expert opinion that yield estimates that are unsatisfyingly uncertain.

Formalization of Inference. The inferential leap to the target species can, however, be performed with more quantitative methods if appropriate data are available. Here, a more general approach to CSE will be followed. The method [29] involves extrapolating dose–effect curves with confidence limits from source to target species and requires the availability of dose–effect data in both species for some toxicants of a particular class. The phrase “class of toxicant” can be given logical *a priori* qualitative meaning, e.g., toxicants that have highly similar modes of action or chemical structure, etc.

The steps for empirical CSE for all members of a class of toxicants for which adequate data exist in both species are as follows. First, for a

given effect, construct the dose–effect function in the source and target species along with the empirical confidence limits. Then, using the two dose–effect equations and their confidence limits, compute a dose-equivalence equation (DEE), which gives a dose in the target species that is known to produce an effect of equal magnitude as any given dose in the source species. The confidence limits around the DEE can be computed [29]. Thirdly, given that the DEEs are available for several members of a toxicant class, the parameters of the DEEs can be used to form a (possibly multivariate) distribution. A final, inferential step may now be made for some member of the toxicant class for which data in the target species are absent.

The inferential leap in empirical CSE depends upon the plausibility of the hypothesis that DEEs are numerically the same for all members of a toxicant class, regardless of the relative potency of each of the members. By this hypothesis, even if the potency varies across the toxicants of a class, the relative potency across species is invariant. If this hypothesis is plausible, then the dose–effect curve in the source species can be used to give the most informed estimate of the dose–effect curve in the target species for any toxicant of the class, by application of a DEE constructed from the mean parameter values of the previously defined distribution. The confidence limits about such an extrapolation may also be computed.

The utility of this method for empirically based CSE depends (a) upon the existence of sufficient data to form a distribution of DEE parameters for a particular toxicant class and (b) the plausibility of the construct of “class of toxicants”. A first-step definition of a class of toxicants can be constructed from whatever *a priori* information is available. After obtaining the distribution of the DEE parameters, the plausibility of the class definition can be evaluated by demonstrating that the DEE parameters do not statistically differ among members of the class. The mean parameters of the DEE from the chemical class can then be assumed to represent all members of that class, including those that have not been tested. In this manner, experimental data, gathered from the source species, on an unknown chemical of that class can be used to predict effects in the target species. It is possible to consider a number of variants on this procedure. If, e.g., sufficient dose–effect data did not exist for construction of DEEs, the process could be

reduced to include whatever data were available, e.g., benchmark doses.

Advantages of the Empirical Approach. In addition to the possibility that sufficient knowledge exists for some empirical CSE, some of the problems that are inherent in PBPK and BBDR modeling are not as intractable, e.g., it may be argued that empirical confidence limits for dose–effect curves are the end result of all of the sources of variation that enter into the PBPK and BBDR models. Thus, if the confidence limits are estimated empirically, the need to know extensive model parametric data and variance/covariances are greatly reduced. On the other hand, the possible biases (systematic errors) in estimates of internal dose or effects cannot be ignored in empirical CSE. The effect of biases may be studied *via* a comprehensive sensitivity analysis [30].

Similar to mechanistic approaches, in empirical approaches too adequate knowledge is scarce. Data for particular toxicants in both species typically exist (if at all) as individual publications and can sometimes be combined *via* meta-analyses (*see Meta-Analysis in Nonclinical Risk Assessment*). Only one example of empirical DEE [31] has been published and there is no case of dose–effect CSE in literature. Perhaps, knowledge gaps in mechanistic methods can be solved by empirical methods and *vice versa* because of a complementary state of knowledge. Clearly, however, both methods suffer the same kinds of limitations.

Evaluation and Recommendation

From the above discussion, it is apparent that the area of quantitative CSE holds much promise but, compared to the promise, has managed to deliver little. The reasons for this are not simple to categorize, but can be broadly stated as “lack of sufficient knowledge”. Fortunately, one of the most important functions of explicating the goals and methods of CSE is didactic. Constructing appropriate methods quickly leads to the discovery of specific, well-defined needs.

Mechanistic Approaches

The outstanding need with regard to dosimetry is for PBPK models, which will be usable in applied cases, e.g., (a) simultaneous exposure to multiple

toxicants; (b) environments in which multiple important parameters vary as a function of time; (c) applications in which gender, age, health, species, etc., would affect the parameters; and (d) more comprehensive representation of the metabolic processes for many toxicants.

BBDR models fare less well in successful implementation and utility. It is tempting to claim that such modeling is newer, but in a general physiological overview, this is not the case. It is their application to quantitative description of toxicant effects that is more recent. Some impediments to the development of BBDR models have been ameliorated lately, e.g., larger, faster computers and measurement devices permitting the quantification of more detailed and multivariate measurements in physiological systems. A major problem in construction of BBDR models is, however, the consequence of including such increasingly detailed information. In order to synthesize all of the new knowledge into a whole system, theorists who are also well versed in mathematics and computation are required, in addition to rigorous experimentalists.

In all cases where BBDR models have been successful in toxicological CSE, the models have been specific to a particular toxicant and a particular organ. If the information about health effects of toxicants is to be useful in evaluating the overall consequences of various exposures, such parochialism will not suffice. Toxicants have multiple effects, all of them contributing to the evaluation of the cost of health effects. To evaluate the impact of toxicants on public health, BBDR models for CSE must eventually approach whole body models with sufficient detail to simulate multiple toxicants, each with multiple effects. Of course, such grandiose modeling schemes are far in the future, but it is the philosophy of the approach that is important to guide progress.

Empirical Methods

Systematic empirical methods for CSE can be considered as stopgaps for CSE until mechanistic modeling becomes available, but it is important to note that until it becomes possible to estimate confidence limits around the means of health effects estimated by mechanistic means, empirical CSE methods will be essential. Of the approaches to quantitative CSE, empirical methods that employ dose–effect information have been least employed. This is, in

part, due to a reliance in the regulatory community on point estimation of lowest safe exposures to toxicants. These methods were developed to use currently available information and technology.

Sufficient data exist in the literature to construct DEEs for some members of limited classes of toxicants. No formal effort has been made to do so. The empirical data for a broad range of toxicants is lacking. Certainly, if multiple toxicant exposure is considered, the data base is inadequate.

Conclusions

1. Approaches to CSE are needed to adjust data from the source species by dosimetric and mechanistic methods so as to translate the data into the domain of the target species.
2. Methods exist for estimating the internal dose of a target species on the basis of data from a source species (PBPK models).
3. Methods for estimating the target species effects and confidence limits have been proposed. They can be broadly categorized into mechanistic from physiological theory and empirical/statistical from existing dose–effect data.
4. For the most part, the data for numerical extrapolation do not exist, either in theoretical models or in empirical methods.
5. The methods and schemes proposed for formal extrapolation form a framework to guide the development of new methods and new research to solve the problems.

References

- [1] Watts, J.M. (1872). *Index of Spectra*, A. Weywood, Manchester, pp. ix.
- [2] Jevons, W.S. (1874). *Principles of Science, II*, MacMillan, New York, pp. 120.
- [3] Simpson, J.A. & Weiner, E.S.C. (eds) (1989). *The Oxford English Dictionary*, 2nd Edition, Clarendon Press, Oxford.
- [4] James, W. (1905). *The Meaning of Truth*, Harvard University Press, Cambridge, pp. 129.
- [5] Russel, B. (1927). *Philosophy*, W.W. Norton, New York.
- [6] Anderson, M.E., Clewell III, H. & Krishnan, K. (1995). Tissue dosimetry, pharmacokinetic modeling and interspecies scaling factors, *Risk Analysis* **15**, 533–537.
- [7] Boyes, W.K., Bercegeay, M., Krantz, T., Evans, M., Benignus, V. & Simmons, J.E. (2005). Momentary brain

- concentration of trichloroethylene predicts the effects on rat visual function, *Toxicological Sciences* **87**, 187–196.
- [8] Ramsey, J.C. & Anderson, M.E. (1984). A physiologically based description of the inhalation pharmacokinetics of styrene in rats and humans, *Toxicology and Applied Pharmacology* **73**, 159–175.
- [9] Rubner, M. (1883). Ueber den Einfluss der Körpergrösse auf Stoff- und Kraftwechsel, *Zeitschrift für Biologie* **19**, 535–562.
- [10] Kleiber, M. (1947). Body size and metabolic rate, *Physiological Review* **27**, 511–541.
- [11] Brown, R.P., Delp, M.D., Lindstedt, S.L., Rhomberg, L.R. & Beliles, R.P. (1997). Physiological parameter values for physiologically based pharmacokinetic models, *Toxicology and Industrial Health* **13**, 407–484.
- [12] Benignus, V.A., Hazucha, M.J., Smith, M.V. & Bromberg, P.A. (1994). Prediction of carboxyhemoglobin formation due to transient exposure to carbon monoxide, *Journal of Applied Physiology* **76**, 1739–1745.
- [13] Vicini, P., Pierce, C.H., Dills, R.L., Morgan, M.S. & Kalman, D.A. (1999). Individual prior information in a physiological model of 2H8-toluene kinetics: an empirical Bayes estimation strategy, *Risk Analysis* **19**, 1127–1134.
- [14] Benignus, V.A. (1995). A model to predict carboxyhemoglobin and pulmonary parameters after exposure to O₂, CO₂ and CO, *Journal of Aviation, Space and Environmental Medicine* **66**, 369–374.
- [15] Benignus, V.A., Coleman, T., Eklund, C.R. & Kenyon, E.M. (2006). A general physiological and toxicokinetic (GPAT) model for simulating complex toluene exposure scenarios in humans, *Toxicology Mechanisms and Methods* **16**, 27–36.
- [16] Clark, L.H., Setzer, R.W. & Barton, H.A. (2004). Framework for evaluation of physiologically-based pharmacokinetic models for use in safety or risk assessment, *Risk Analysis* **24**, 1697–1717.
- [17] Price, P.S., Conolly, R.B., Chaisson, C.F., Gross, E.A., Young, J.S., Mathis, E.T. & Tedder, D.R. (2003). Modeling interindividual variation in physiological factors used in PBPK models of humans, *Critical Reviews in Toxicology* **33**, 469–503.
- [18] Cronin, W.J., Oswald, E.J., Shelly, M.L., Fisher, J.W. & Flemming, C.D. (1995). A trichloroethylene risk assessment using a Monte Carlo analysis of parameter uncertainty in conjunction with physiologically-based pharmacokinetic modeling, *Risk Analysis* **15**, 555–565.
- [19] Evans, M.V., Crank, W.D., Yang, H.-M. & Simmons, J.E. (1994). Applications of sensitivity analysis to a physiologically based pharmacokinetic model of carbon tetrachloride in rats, *Toxicology and Applied Pharmacology* **128**, 36–44.
- [20] Guyton, A.C. & Coleman, T.G. (1967). Long term regulation of circulation: interrelations with body fluid volumes, in *Physical Basis of Circulatory Transport: Regulation and Exchange*, E.B. Reeve & A.C. Guyton, eds, WB Saunders, Philadelphia, pp. 179–201.
- [21] Bassingthwaigthe, J.B. (2000). Strategies for the physiome project, *Annals of Biomedical Engineering* **28**, 836–848.
- [22] Lau, C. & Setzer, R.W. (2000). Biologically based risk assessment models for developmental toxicity, *Methods in Molecular Biology* **136**, 271–281.
- [23] Moolgavkar, S., Krewski, D. & Schwartz, M. (1999). Mechanisms of carcinogens and biologically based models for estimation and prediction of risk, *IARC Scientific Publications* **131**, 179–237.
- [24] Conolly, R.B., Kimbell, J.S., Janszen, D., Schlosser, P.M., Kalisak, D., Preston, J. & Miller, F.J. (2003). Biologically motivated computational modeling of formaldehyde carcinogenicity in the F344 rat, *Toxicological Sciences* **75**, 432–447.
- [25] Conolly, R.B., Kimbell, J.S., Janszen, D., Schlosser, P.M., Kalisak, D., Preston, J. & Miller, F.J. (2004). Human respiratory tract cancer risks of inhaled formaldehyde: dose-response predictions derived from biologically-motivated computational modeling of a combined rodent and human data set, *Toxicological Sciences* **82**, 279–296.
- [26] Stuhmiller, J.H., Long, D.W. & Stuhmiller, L.M. (2006). An internal dose model of incapacitation and lethality risk from inhalation of fire gases, *Inhalation Toxicology* **21**, 347–364.
- [27] Crump, K.S. (1984). A new method for determining allowable daily intakes, *Fundamental and Applied Toxicology* **4**, 854–871.
- [28] MacPhail, R.C. & Glowa, J.R. (1999). Quantitative risk assessment in neurotoxicology: past, present and future, in *Neurotoxicology*, H.A. Tilson & G.J. Harry, eds, Taylor & Francis, New York, pp. 367–382.
- [29] Benignus, V.A. (2001). Quantitative cross-species extrapolation in noncancer risk assessment, *Regulatory Toxicology and Pharmacology* **34**, 62–68.
- [30] Saltellio, A., Tarantola, S. & Campolongo, F. (2000). Sensitivity analysis as an ingredient of modeling, *Statistical Sciences* **15**, 377–395.
- [31] Benignus, V.A., Boyes, W.K., Hudnell, H.K., Frey, C.M. & Svendsgaard, D.J. (2291). Quantitative methods for cross-species mapping (CSM), *Neuroscience and Biobehavioral reviews* **15**, 165–171.

VERNON A. BENIGNUS, WILLIAM K. BOYES
AND PHILIP J. BUSHNELL

Cumulative Risk Assessment for Environmental Hazards

As part of their existence, people and ecosystems are commonly exposed to a diverse and dynamic mixture of environmental agents, including biological (e.g., *Mycobacterium tuberculosis*), chemical (e.g., organochlorine pesticides), physical (e.g., UV radiation), and psychosocial (e.g., overcrowding) stressors. *Cumulative risk* refers to the combined probability of harm from exposure to all of these different kinds of stressors across all relevant exposure pathways and routes. In the case of humans, there is solid scientific evidence, for example, documenting increased cancer risks from concurrent exposure to tobacco smoke (see **Environmental Tobacco Smoke**) and asbestos (see **Asbestos**), to tobacco smoke and radon (see **Radon**), and to aflatoxin and hepatitis B infection. While researchers and risk assessors have known for decades about the possibility of enhanced cumulative risk from exposure to multiple environmental stressors, progress in understanding responsible biological mechanisms and developing suitable quantitative methods for risk assessment have been slow because of insufficient regulatory attention, technological limitations, and scarce funding. Consequently, realistic assessment of cumulative risks from exposure to real-world mixtures of environmental stressors is limited in most cases by a paucity of relevant data and a lack of scientific understanding [1, 2].

Cumulative risk assessment involves the analysis, characterization, and possible quantification of the combined risks to human health [3, 4] or environmental systems [5] from exposure to multiple environmental stressors (see **Environmental Health Risk; Environmental Hazard**). It is a tool for organizing and evaluating risk-related information so that risk assessors and risk managers can make realistic determinations of the scope and magnitude of amalgamated adverse effects from environmental mixtures [1, 2]. Compared to traditional human risk assessment, cumulative risk assessment is different in four ways: it may be qualitative (rather than quantitative) depending on the circumstances; the combined effects of more than one agent are evaluated; the emphasis is on actual (real) at-risk populations

rather than hypothetical individuals or scenarios; and assessments can involve more than just chemicals.

The framework for assessment of cumulative risks, as shown in Figure 1, encompasses three interrelated and generally sequential phases: (a) planning, scoping, and problem formulation; (b) data analysis; and (c) interpretation and risk characterization [1, 2]. A team of researchers, risk assessors, and stakeholders work together in the first phase to determine the goals, breadth, depth, and focus of the assessment. Outcomes of this step include a conceptual model that identifies the stressors to be evaluated, the adverse endpoints to be assessed, and the nature of relationships among exposures and effects, as well as an analysis plan that specifies the data needed, the approach to be taken, and the types of results expected in subsequent phases. In the second phase, exposure profiles are developed, risks are estimated, and variability and uncertainty are discussed. During this step it is imperative to describe the interactions among stressors and their effects on mixture toxicity, and identify potentially vulnerable populations. The products of the analysis phase are either a quantitative or qualitative estimate of cumulative risk along with a description and estimation of related uncertainty and variability. In the final phase, the risk estimates are explained, overall confidence in the estimates is defined, effects of key assumptions on results are described, and a determination is made as to whether the assessment met the goals and objectives set forth in phase one. Although cumulative risk assessment is still in its infancy, the literature contains numerous examples of diverse methods used to evaluate combined effects on human health [6–10] and ecological systems [11–14] (see **Ecological Risk Assessment**).

There are several different theoretical approaches for estimating risk from exposure to multiple stressors [1, 2]. For example, by assuming toxicologic independence or similarity it is possible to approximate the joint exposure–response relationship for a mixture of environmental stressors using only information on individual stressors. Assuming toxicologic independence means that toxicity modes of action are biologically independent and there are no pretoxicity interactions (e.g., metabolic inhibition). By assuming toxicologic similarity, stressors can be grouped according to a common mode of action and “dose addition” (e.g., relative potency factors, toxic equivalency factors) used to estimate cumulative risk [1, 2].

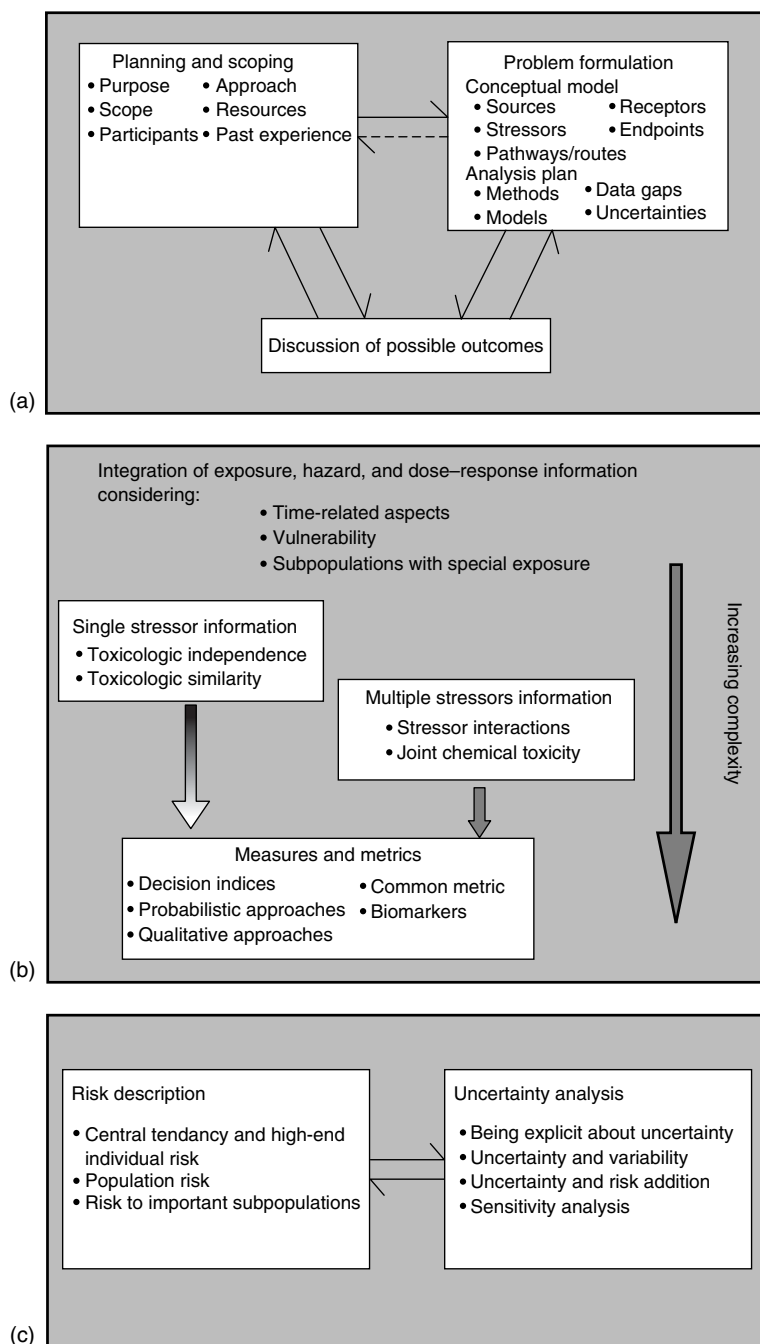


Figure 1 The three interrelated and generally sequential phases of cumulative risk assessment. (a) Planning, scoping, and problem formulation phase; (b) analysis phase; and (c) interpretation and risk characterization phase [Reproduced from [2]. National Institute of Environmental Health Sciences, 2007.]

“Response addition”, like dose addition, assumes there are no interactions among mixture constituents because they either act independently or have different target organs so that the presence of one stressor does not affect the toxicity of another. The response addition method has historically been used primarily for chemical carcinogens [1, 2].

Dose–response evaluation (*see* **Dose–Response Analysis**) can occur separately from the exposure assessment step if certain simplifying assumptions are made, such as all exposures occur continuously, exposure sequence is unimportant, and mixture composition is constant. It is then possible to set health-protective action levels using estimates of upper bounds on toxic potency caused by exposure, and lower bounds on the acceptable (safe) exposure level [1, 2].

In cases where there are divergent data from different sources on exposure (e.g., time-weighted *versus* continuous) and toxicity (e.g., quantitative *versus* qualitative), indices can be used to convert dissimilar multivariate data into a single number. The hazard index (HI) for noncarcinogens (systemic toxicants) is the most common example of this approach for cumulative risk assessment. Each HI is calculated for a specific target organ (e.g., respiratory, neurological, and cardiopulmonary) by summing the hazard quotients of each stressor (i.e., the exposure concentration divided by the relevant reference dose or concentration) in the environmental mixture of interest. A mixture $HI \leq 1$ is generally interpreted to mean that the mixture is unlikely to be harmful, while a $HI > 1$ indicates the need for further evaluation [1, 2].

Selecting an approach for conducting a cumulative risk assessment depends on a variety of factors, including whether there are interactions among mixture components; the types of interactions (e.g., toxicokinetic and toxicodynamic) that occur; the quality and quantity of available data; and the objectives of the analysis. Application of more than one method may be justified, depending on the circumstances. Moreover, a simple, less data-intensive method may be appropriate for initial screening, while a more intricate and data-intensive method may be appropriate for follow-up analysis [1, 2]. In conclusion, risk assessment in regulatory agencies, particularly the US Environmental Protection Agency (EPA), is currently undergoing a transition. Risk assessors and risk managers are beginning to move beyond their historical focus on command-and-

control strategies, end-of-pipe controls, narrow media-based enforcement, one-size-fits-all regulations, rigid and prescriptive rules, and process-based technology standards. The challenges of the twenty-first century are pushing them to focus more on cooperative and voluntary control strategies, pollution prevention, holistic multimedia approaches, place-based environmental decisions, flexible, easy-to-adjust rules, and outcome-based standards. These changes in regulatory philosophy require complementary changes in risk assessment, and emphasis is shifting away from a narrow focus on single stressors, endpoints, sources, pathways, and environmental media to a more inclusive approach that comprises multiple stressors, endpoints, sources, pathways, and environmental media [1, 2].

Cumulative risk assessment is an essential and inevitable element in this ongoing transition. As more data and better scientific understanding become available about exposures to environmental mixtures, interactions among mixture constituents, and combined effects on human populations and environmental resources, cumulative risk assessment will emerge as a critical tool for risk-based decision making.

References

- [1] U.S. Environmental Protection Agency (2003). *Framework for Cumulative Risk Assessment*, Risk Assessment Forum, Office of Research and Development, EPA/630/P-02/001A, Washington, DC.
- [2] Callahan, M.A. & Sexton, K. (2007). If ‘cumulative risk assessment’ is the answer, what is the question? *Environmental Health Perspectives* **115**(5), 799–806.
- [3] Sexton, K. & Hattis, D. (2007). Assessing cumulative health risks from exposure to environmental mixtures – three fundamental questions, *Environmental Health Perspectives* **115**(5), 825–832.
- [4] de Fur, P.L., Evans, G.W., Cohen-Hubal, E.A., Kyle, A.D., Morello-Frosch, R.A. & Williams, D.R. (2007). Vulnerability as a function of individual and group resources in cumulative risk assessment, *Environmental Health Perspectives* **115**(5), 817–824.
- [5] Menzie, C.A., MacDonnell, M. & Mumtaz, M. (2007). A phased approach for assessing combined effects from multiple stressors, *Environmental Health Perspectives* **115**(5), 807–816.
- [6] U.S. Environmental Protection Agency (2002). *Revised Organophosphate Pesticide Cumulative Risk Assessment*, Office of Pesticide Programs, Office of Prevention, Pesticides, and Toxic Substances, EPA/630/R-00/002, Washington, DC.

- [7] U.S. Environmental Protection Agency (2005). *Estimation of Cumulative Risk from N-Methyl Carbamate Pesticides: A Preliminary Assessment*, Office of Pesticides Programs, Office of Prevention, Pesticides, and Toxic Substances, Washington, DC.
- [8] Caldwell, J., Woodruff, T., Morello-Frosch, R. & Axelrad, D. (1998). Applications of health information to hazardous air pollutants modeled in EPA's cumulative risk project, *Toxicology and Industrial Health* **14**, 429–454.
- [9] van den Berg, M., Birnbaum, L., Bosveid, A.T.C., Brunstrom, B., Cook, P., Feeley, M., Fielder, H., Hakansson, H., Hanber, A., Haws, L., Rose, M., Safe, S., Schrenk, D., Tohyawa, C., Tritscher, A., Tuomista, J., Tysklind, M., Walker, N. & Peterson, R.E. (1998). Toxic equivalency factors (TEFs) for PCBs, PCDDs, PCDFs for humans and wildlife, *Environmental Health Perspectives* **106**, 775–792.
- [10] Walker, N.J., Crockett, P.W., Nyska, A., Brix, A.E., Jokinen, M.P., Sells, D.M., Hailey, J.R., Easterling, J.K., Haseman, M., Yin, M.F., Wyde, J.R., Bucher, J.R. & Portier, C.J. (2005). Dose-additive carcinogenicity of a defined mixture of dioxin-like compounds, *Environmental Health Perspectives* **113**, 43–48.
- [11] U.S. Environmental Protection Agency (2002). *Clinch and Powell Vally Watershed Ecological Risk Assessment*, National Center for Environmental Assessment, Office of Research and Development, EPA/600/R-01/050, Washington, DC.
- [12] U.S. Environmental Protection Agency (2002). *Waquoit Bay Watershed Ecological Risk Assessment: The Effect of Land-Derived Nitrogen Loads on Estuarine Eutrophication*, National Center for Environmental Assessment, Office of Research and Development, EPA/600/R-02/079, Washington, DC.
- [13] Gentile, J.H., Harwell, M.A., Cropper, W., Harwell, C.C., DeAngelis, D., Davis, S., Ogden, J.C. & Lirman, D. (2001). Ecological conceptual models: a framework and case study on ecosystem management for South Florida sustainability, *Science of the Total Environment* **274**, 231–253.
- [14] Obey, A.M. & Landis, W.G. (2002). A regional multiple stressor risk assessment of the Cordous Creek watershed applying the relative risk model, *Human and Ecological Risk* **8**, 405–428.

KEN SEXTON

Cyber Risk Management

The cyber security landscape is made up of the asset owners, their critical missions, their supporting information systems, and process control systems (collectively called the *user model*), the set of adversaries, their skills, motivations, and resources (collectively called the *adversary model*), the space of all cyber-related attacks (see **Digital Governance, Hotspot Detection, and Homeland Security**), and the space of all cyber attack countermeasures. These different models and spaces are interrelated. The asset owners use information computer networks, information and process control systems to more efficiently achieve their operational missions. The adversaries have motivation, cyber attack skills, and resources to attack the asset owner's critical information and process control systems by exploiting vulnerabilities in the computer technology or operational procedures. Finally, the asset owners can choose to deploy security countermeasures to protect against the cyber attacks; however, each security countermeasure comes at a cost in terms of money, functionality, interoperability, ease of use, etc. So, the asset owner must deploy security countermeasures only when needed. A structured cyber risk-management process helps the asset owner analyze each model and space to determine the appropriate set of cyber security countermeasures. The cyber risk-management process will assess the vulnerabilities to the asset owner's systems, determine the impact to the mission if these vulnerabilities were to be exploited, assess the asset owner's adversaries and their motivation, skills, and resources to mount the cyber attacks that exploit the system vulnerabilities, and estimate the probabilities and expected loss of cyber attacks. Finally, the cyber risk-management process will compare the costs and benefits of adding, deleting, or modifying the cyber security countermeasures.

Definitions

Risk management is the process of measuring or assessing risk, and then determining the appropriate strategy to deal with the risk. Risk is a commodity and can be bought and sold and moved from one location to another. Possible risk-management strategies include combinations of transferring the

risk to another party (e.g., an insurer), decreasing the risk, accepting the risk, or increasing the risk.

Risk assessment is the critical step in risk management of measuring the risk to an organization. Risk assessment typically includes two metrics, the probability of a harmful event, P_e , and the consequences of a harmful event, C_e . Thus, the risk of a particular event is the product of these two components:

$$R_e = P_e \times C_e \quad (1)$$

Cyber risk assessment is the process of measuring the asset owner's risk introduced by the use of computer networks, information systems, and process control systems for mission-critical functions.

Cyber risk is the risk to an asset owner's critical missions introduced by the reliance on computer networks. Cyber risk has two components; nonmalicious events caused by human error, pranksters, or failures in the computer systems, and malicious attacks intentionally caused by intelligent adversaries.

Cyber attacks are a series of intentional, malicious activities executed by an intelligent adversary that exploit vulnerabilities in an asset owner's computer systems with the intent to degrade their critical missions. Cyber attacks typically target either the availability, integrity, or confidentiality of mission-critical information.

Active versus passive: An "active attack" attempts to alter the computer system resources or affect their operation. A "passive attack" attempts to learn or make use of information from the computer system but does not affect system resources (e.g., wiretapping).

Insider versus outsider: An "inside attack" is an attack initiated by an entity inside the security perimeter (an "insider"), i.e., an entity that is authorized to access the computer system resources but uses them in a way not approved by those who granted the authorization. An "outside attack" is initiated from outside the perimeter, by an unauthorized or illegitimate user of the system (an "outsider"). In the Internet, potential outside attackers range from amateur pranksters to organized criminals, international terrorists, and hostile governments.

Attack characteristics are those aspects of an attack against a computer system that determine the level of risk introduced by that attack. The attack

characteristics include the attack objective (the attack usually targets a mission-critical function), the cost to the adversary of attempting the attack, the level of difficulty of the attack, and the level of detection of the attack.

Costing attacks is the risk assessment phase of assigning quantitative values to the attack characteristics.

Threat is the combination of an adversary with an intention to cause harm or damage, the corresponding attack objective, attack techniques, and mission impact.

Cyber adversaries are those groups or individuals that have the motivation and ability to conduct cyber attacks against the asset owner's mission-critical computer systems.

Cyber vulnerability is a flaw or weakness in a computer system's design, implementation, or operation and management that could be exploited to violate the computer system's security policy. Most computer systems have vulnerabilities of some sort, but this does not mean that all the computer systems are too flawed to use. Not every threat results in an attack, and not every attack succeeds. Success depends on the susceptibility of the computer system, the degree of difficulty of exploiting the vulnerability, the skill level of the attacker, and the effectiveness of any countermeasures that may mitigate the vulnerability. If the attacks needed to exploit vulnerability are very difficult to carry out, then the vulnerability may be tolerable. If the perceived benefit to an attacker or impact to the defender is small, then even an easily exploited vulnerability may be tolerable. However, if the attacks are well understood and easily made, and if the vulnerable system is used by a wide range of users, then it is likely that there will be enough benefit for someone to make an attack.

Cyber Risk-Management Workflow

The following list describes a work flow for a generic cyber risk-management process [1]:

1. Characterize mission-critical functions and assets

Identify and quantify the mission criticality of each cyber function and asset. The value of cyber assets to the asset owner far exceeds the replacement cost of the hardware and software. Oftentimes, the true value of cyber assets and functions are best

measured in terms of impact to business continuity. For example, the availability and integrity of a company's accounting software is more valuable to the company than the purchase price of the software package. For each critical cyber asset and function, estimate the loss to the organization if the availability, integrity, or confidentiality of that cyber asset or function is lost.

2. Identify adversaries

Today's critical computer systems must operate in a hostile environment, protecting themselves against a variety of threats. We define *threat* as the combination of an adversary, attack objective, attack technique, and mission impact. *Adversaries* are those groups or individuals that have the motivation and ability to conduct cyber attacks against the asset owner's mission-critical computer systems. Adversaries include hostile insiders or outsiders. The insider may be co-opted by another adversary or be working on his own behalf. Some examples of hostile outsiders include foreign nation states, terrorist organizations, organized crime, hackers, and economic competitors. Modeling the adversary by characterizing the adversary's desire and ability to conduct attacks helps predict adversary behavior. The adversary characteristics include the following:

- (a) Attack objectives – ways of degrading the customer's critical missions.
- (b) Available resources – money, manpower, and time.
- (c) Sophistication – technology and capabilities for conducting difficult cyber attacks.
- (d) Risk tolerance – willingness to take chances and risk failure and detection.

3. Characterize computer system

Defining the system to be analyzed is a crucial step in the cyber risk assessment process. Having a detailed, well-defined system description helps scope the analysis. The system description starts with a mission statement, a concept of operations, and a high-level functional description of the system. More details will be needed for security-critical components and functions. Finally, information flow diagrams and use cases map the system description to the critical missions and help with the risk assessment.

4. Identify attack objectives

Combine the adversary's high-level attack objectives with the system description to derive system-specific

attack objectives (*see Use of Decision Support Techniques for Information System Risk Management*). Because the computer system provides mission-critical functionality to the asset owner, the cyber adversary will harm the organization by attacking the computer system. Example attack objectives include “steal the organization’s proprietary product designs” or “destroy the organization’s engineering database”.

5. Generate attack list

The objective of identifying attacks is not to simply identify some vulnerabilities, but to discover all the effective ways of defeating the security of the computer system. Doing so will identify the true weak points of the computer system. To help identify a comprehensive list of system attacks, the cyber risk analyst should consider the following:

- (a) Start by identifying all the critical missions and corresponding high-level attack objectives, and keep these in mind during the entire process.
- (b) Think about all the ways the adversary could accomplish the attack objective. Think broadly at first, and then fill in the details later.
- (c) Consider using attack trees to structure the thinking [2–5].
- (d) Use comprehensive attack taxonomies that include all of the adversary’s attack capabilities. A good attack taxonomy will include attack capabilities against which the computer system may not have a proper defense. For example, the adversary may introduce an unknown vulnerability into the design of a commercial component using a life-cycle attack [6].
- (e) Combine attack steps from different attack disciplines. For example, the adversary could co-opt an insider, and then use the insider access to launch sophisticated computer network attacks (CNAs).
- (f) Consider the adversary’s attack strategy over time. The adversary may gain access to the system earlier in the life cycle, but may not pull the trigger on the final attack step until much later when the impact will be most devastating to the customer’s mission.

Thinking about all the ways the adversary could break the rules will help the defender identify all

the most serious attack scenarios (*see Multistate Systems*).

6. Quantify risk of each attack

The next step in the cyber risk assessment process is to quantify the risks by scoring the attack steps according to factors that influence the adversary’s attack strategy: attack cost and difficulty, risk of detection, consequences of detection, and the impact of a successful attack. For each identified attack, the cyber risk analyst scores the attack by estimating the resources required to carry out the attack, the capabilities needed to succeed, the degree of difficulty of the attack, and the probability that the attack will be detected.

7. Manage the risk

The final step in the risk-management process (*see History and Examples of Environmental Justice; Syndromic Surveillance*) is to compare alternatives and perform cost/benefit analysis to make trade-offs between adding, modifying, or deleting security countermeasures to most effectively provide security, functionality, and usability within the cost and time constraints of the asset owner. Because of the various costs incurred, cyber security countermeasures should be applied to directly prevent, detect, or recover from the most serious cyber attacks identified during the structured cyber risk assessment. The overall goal is to provide the right balance of cyber security to optimize mission effectiveness for the asset owner.

Cyber Adversary Space

The cyber adversary space includes different classes of adversaries that use a combination of cyber attack techniques to cause harm and inflict damage on the defender’s mission or assets. The following sections describe the cyber adversaries in terms of their observed behavior, their motivations to conduct cyber attacks, their resources available to conduct cyber attacks, their cyber attack skills, and their tolerance for being detected.

Nation States

This class of adversary is a conglomerate of all state-sponsored foreign intelligence services (FISs) and special military units (e.g., an Information Warfare division) that have a potential to attack critical communications and computer systems. Foreign

nation states, whether they be enemy or ally, continuously conduct intelligence operations to defeat the confidentiality of sensitive information. In a crisis or war scenario, the foreign nation state may also conduct special information operations to attack the confidentiality, integrity, and especially the availability of mission-critical infrastructure systems.

Note that it is impossible to characterize the entire spectrum of foreign nation states with a single adversary model. Different nations have vastly different characteristics. For example, some nations are quite belligerent, but do not have sophisticated information warfare capabilities. Other nations might be technically sophisticated, but may be very peaceful. This adversary model attempts to characterize the most significant foreign threats from modern nations, realizing that not all nation states fall into this characterization.

Because nation states behave so differently in times of peace, crisis, and war, this adversary model separates this class of adversary into two subclasses; the peacetime adversary is described by the FIS, and the wartime adversary is described as the Info Warrior.

Foreign Intelligence Service (FIS)

The FIS represents those nations that conduct active espionage against other nations.

- *FIS motivation* – The primary attack objective of the FIS is intelligence gathering. The FIS will target computer systems that store or process information with high intelligence value.
- *FIS resources* – The FIS adversary has an extremely high level of resources available for cyber attacks, including money, man power, time, and high technology. Some nation states spend a significant level of their annual gross national product on intelligence gathering.
- *FIS skills* – The FIS is highly sophisticated in human intelligence (HUMINT), signals intelligence (SIGINT), CNA, SpecOps, and electronic warfare (EW) attack techniques. For those nations that have an indigenous high technology industry, the FIS can conduct life-cycle attacks through their own industry.
- *FIS risk tolerance* – This adversary is risk averse, in that espionage is most effective when not detected. Some FIS are not above taking physical

risk by breaking and entering to steal information or leave a listening device. In addition to these traditional techniques, the modern FIS has adopted new CNA techniques to steal sensitive information from critical computer networks. However, given their preference to conceal their advanced capabilities, an FIS with sophisticated network attack techniques may not display them until absolutely necessary, instead relying on adequate, yet less sophisticated, network attack tools such as publicly available hacking and system penetration tools.

Info Warriors

The info warriors are a team of highly trained, highly motivated military professionals from a nation state's Information Warfare division, and act on behalf of the nation state in preparation or execution of a war with other nations.

- *Info warrior motivation* – The info warrior seeks to destroy the opposing critical infrastructure to reduce the effectiveness to wage war and to erode public support for the war.
- *Info warrior resources* – Since the info warrior is part of the nation state's arsenal of highly sophisticated weapons, the budget and resources for the info warrior are extremely high. During times of war, the nation state's military budget, and hence the info warrior's budget, will increase and become a national priority.
- *Info warrior skills* – Information operations conducted by nation states at war include manipulation, degradation, or destruction of data or supporting infrastructures within the targeted network. Information operations during times of crisis and war are generally conducted with full-time intelligence support from the FIS that aids in identifying network vulnerabilities, structure, and the mapping of mission-critical information about and within the targeted network. Once the information operation begins, the info warriors will use any and all means to achieve their objectives.
- *Info warrior risk tolerance* – Once in a state of war, the info warrior is willing to take extreme risks. For these reasons, the info warrior is a very formidable adversary.

Terrorist (Nonstate Actors)

This class of adversary includes organizations that engage in terrorist activities to further their political or religious goals. Terrorist organizations may cooperate with a foreign nation state, or they may be seeking visibility and publicity for specific issues or to influence social, political, or military behavior among a targeted audience.

- *Terrorists' motivation* – Terrorists seek publicity and political change by wreaking havoc, motivated by their ideology or cause. Terrorists typically resort to high-visibility destructive attacks. Currently, terrorists have not posed a significant cyber threat to critical communications and information systems. While current expert opinion believes that cyber attacks against information systems will not supplant more violent, physical (kinetic) terrorist actions, information operations attacks may be used as a force multiplier to maximize the impact of violent operations by complicating emergency responses and coordinated reactions to physical attacks. In addition, attacks on networks may disable certain physical security measures that may also increase the impact or probability of success of violent attacks. After the physical attacks occur, terrorists may target information systems to gain intelligence about planned responses or retaliation operations, and potentially attempt to disrupt the US responses to their acts of aggression. But this is speculation and currently we do not have evidence to support this.
- *Terrorists' resources* – Terrorist groups' sizes range from a few isolated and uncontrolled amateur individuals to military-structured and trained organizations. In some cases, foreign nation states sponsor terrorist groups by providing training, equipment, and political asylum.
- *Terrorists' skills* – Traditional terrorist expertise lies in making bombs as opposed to being IT or even INFOSEC specific. However, some groups have allegedly raised funds through software piracy and TV descrambler trafficking, which implies some basic hacking skills. Like other organized criminals, terrorists are likely to have basic communications interception capabilities. Some terrorist groups have embraced cellular and encryption technology to frustrate law enforcement and national intelligence gathering

efforts, and they have used the Internet to distribute propaganda regarding their organizations and objectives.

- *Terrorists' risk tolerance* – This adversary may lack high-tech capability, but they make up for it by taking extreme risks, i.e., they are willing to die to make their cause known.

Hackers

This class of adversary does not always fit into the "opposition" classification because of the diversity of hackers. A typical hacker is the recreational hacker. The danger posed by recreational hackers is that they may harm a critical computer system unintentionally. This is particularly true for the less sophisticated hackers, often referred to as *script kiddies*, who use readily available tools developed by others. Recreational hackers often engage in a form of random selection for attacking a system, choosing '.mil' and '.gov' URLs as attractive targets. Another method is to "surf" the Internet until they find an interesting location based on attention-getting characteristics such as an interesting web page found through a keyword search engine. At the other end of the spectrum is the "cracker", who is a person with the capability to penetrate an information system for malicious purposes.

- *Hackers' motivation* – Hackers conduct attacks on systems for a variety of reasons. Notoriety, thrills, and the occasional theft of service motivate the hacker. Some hackers attack computer systems simply for the challenge of penetrating an important system based on the value of the information within the network. These hackers want to match wits with the best network security experts to see how they fare. Other hackers attack information systems for the purpose of revealing vulnerabilities and weaknesses in popular software applications and hardware. Finally, crackers may attack computer systems for financial or personal gain, effectively using their skills to steal money, services, or otherwise valuable data, which they can either sell or use to blackmail their victims. The recreational hacker visits a system just because it is there or because it poses a particular challenge to the hacker's knowledge and skills.

Both hackers and crackers could be working, wittingly or unwittingly, for another class of

adversary such as the foreign nation state or terrorist organization. Some recent foreign hacker activities suggest that many hackers/crackers have some level of operational plan and may operate in concert with other hackers in mounting coordinated attacks. The 1998 attacks on the Pentagon systems known as *Solar Sunrise* is a well-known example of such an effort (http://www.sans.org/resources/idfaq/solar_sunrise.php). During periods of crisis and war, hacker activity targeting military systems tend to increase. This is probably because of the increase in awareness among the hacker community of the military mission and its presence on the Internet. During crisis and war, hackers may be using their skills to voice their opposition to the military actions by making a nuisance of themselves. But the increase in detected hacker activity during these periods is typically limited to port scans and pings, and rarely results in a security incident or causes any noticeable degradation in system functionality. Some hackers conduct useful evaluations of commercial security products and are actually beneficial to the security community. These hackers regularly test popular products looking for vulnerabilities. By publicly exposing these vulnerabilities, as opposed to secretly exploiting them on operation systems, these hackers effectively convince the product vendors to correct the vulnerabilities and improve their products.

- *Hackers' resources* – The hacker may not have much money, but will leverage technical knowledge into computing resources. Hackers often organize in underground groups, clubs, or communities, and share and exchange vulnerability data, experiences, techniques, and tools. Hackers brag about successful attacks against prestigious networks or web sites. They share detailed instructions on how to defraud service providers such as PTOs and ISPs. Hackers routinely publicize their findings through chat rooms, web sites, trade magazines, and yearly conferences.
- *Hacker's skills* – Hackers are masters of network-based attacks that have a relatively low cost of entry. Typically all they need is a personal computer and Internet access to launch their attacks. They use social engineering to steal passwords and insider information about their intended target. Although they have worldwide electronic access through the Internet, hackers

typically will not conduct attacks that require any physical access such as breaking into an office space or digging trenches to wiretap cables.

Not all hackers are as brilliant as reporters like to depict, but some of them are experts in their field and have sufficient knowledge to develop highly sophisticated tools. Once these sophisticated tools enter the public domain, less skilled hackers, so-called script kiddies, will use them indiscriminately. The alarming trend is that as the tools and attack scripts available to hackers become more sophisticated and powerful, the level of sophistication and expertise needed to run the associated attacks is dropping.

- *Hackers' risk tolerance* – Although hackers like notoriety, they do not want to get caught or detected by the law enforcement because recent offenses have resulted in severe sentences. Hackers are also very unlikely to engage in a physical attack prior to perpetrating a cyber attack: the physical risk of being discovered or killed is usually too high. Hacker attacks against classified military systems are rare because of the high physical risk involved. So far, hackers have only attempted attacks against the integrity or availability of unclassified Internet sites when the hacker believes the attack is anonymous.

Organized Crime Groups

This class of adversary includes individuals or organized criminal groups that may be either foreign or domestic. Organized crime is composed of the various national and international crime syndicates, drug cartels, and well-organized gang activities.

- *Organized crime's motivation* – The primary objective of organized criminals is to realize financial profits and establish control over markets for their exploitation. They regularly use the international financial infrastructure to launder money. Other criminals simply steal information or goods for profit or counterfeit high technology products to sell on the black market.
- *Organized crime's resources* – Organized criminals have a wide range of resources. Some criminal groups are strictly small-time with little or no budget for cyber attacks. Other groups, such as large drug cartels, are well funded and are quite experienced using computer networks to their advantage.

- *Organized crime's skills* – Criminal groups may have in-house hacking expertise, but are more likely to recruit appropriate experts in the hacking community or from among former nation state intelligence experts. Organized criminal groups use bribery, blackmail, or torture to convince cyber attack experts to work on behalf of the criminal group. As cyber attacks become more profitable, criminal groups will become more technically sophisticated.
- *Organized crime's risk tolerance* – The primary operational concerns for organized crime are to avoid interdiction, detection, and apprehension. The backbone of successful criminal operations is intelligence collection and exploitation against their intended targets and against law enforcement. Criminals generally have a low tolerance to the risk of discovery because the success of their operations depends on their ability to avoid detection and apprehension. If the risk of detection and apprehension for a particular mission is high, organized criminal groups will often use expendable low-level foot soldiers to take the risk, and hence the fall.

Economic Competitors

Economic competitors are a class of adversary that uses cyber attacks to gain an economic competitive edge in the market. Economic competitors include rival companies and even foreign governments that use national assets to foster their internal industries.

- *Economic competitor's motivation* – The primary motivation for economic competitors is to seek an economic competitive advantage through cyber attacks. This economic advantage could come from learning about a rival company's advanced research and product development before that information becomes public, or it could come from insider information about the rival company's finances or bargaining positions.
- *Economic competitor's resources* – Rival companies typically have only moderate monetary resources and they need a positive expected return on cyber attack investment. If the economic competitor has nation state sponsorship, then this class of adversary has significant resources, comparable to that nation's FIS.
- *Economic competitor's skills* – Economic competitors are fairly sophisticated in HUMINT

(recruiting insiders from their competitors), CNA, and sometimes in SpecOps through physical break-ins to steal information or technology.

- *Economic competitor's risk tolerance* – As conducting cyber attacks against rival companies could hurt the economic competitor's reputation and could perhaps lead to fines and even jail time, economic competitors typically have low risk tolerance for detection and attribution.

Malicious Insiders (Angry or Unethical Employees, Contractors, Consultants, or Vendors)

Historically, this adversary class poses the greatest threat to most critical computer systems because malicious insiders are authorized users who perform unauthorized activities with the intent to damage or steal critical resources. Insiders are particularly difficult to prepare for and defend against. Insiders include employees, contractors, consultants, product vendors, and ancillary staff such as service providers' employees and maintenance personnel. Malicious insiders may be acting on their own behalf, or they may be co-opted by another adversary class such as an FIS.

- *Malicious insider's motivation* – Motivation comes from personal gain, revenge, or ideology. Malicious insiders may have been recruited by an FIS, and thus financially rewarded by the FIS for their crimes. Others may be disgruntled employees who seek revenge against their organization for some perceived slight. Still others may have some ideological differences with the organization and seek to undermine the organization's critical missions.

In addition to malicious insiders, nonmalicious insiders also pose a threat to critical computer systems. "Attacks" from nonmalicious insiders may be the results of misuse, accidents, or other unintentional sources of harm such as the accidental insertion of malicious code from a downloaded or corrupted file. Therefore, the nonmalicious insider could be the unwitting vehicle for other classes of adversaries.

- *Malicious insider's resources* – The insider does not have or need significant resources to cause great harm to their organization. Even the most compartmented organizations must extend enough trust to employees and personnel to perform the duties for which they are assigned; therefore,

some degree of access is usually available for exploitation by an insider. Furthermore, insiders are in a position to defeat all three security services (confidentiality, integrity, and availability), and would most likely place themselves in a position where the attack is not easily attributable to themselves. While the expected number of individuals within an organization that might wish to cause harm may be low, the potential impact caused by an insider attack could be massive.

- *Malicious insider's skills* – Unlike outsiders, insiders do not need to penetrate a computer system by defeating its security mechanisms, but rather, are given access to the internal system or portions therein. For this reason, insiders do not require sophisticated attack skills. However, the most dangerous insider is the system administrator with detailed knowledge of the computer system and its security mechanisms.
- *Malicious insider's risk tolerance* – Just like any other criminal, insiders do not want to get caught and prosecuted. They will take the necessary precautions to avoid detection and attribution. However, because of their status as a trusted employee, the insider is capable of conducting attacks that are deemed too risky by other external threats. This makes insiders a valuable commodity for other adversaries such as the FIS.

Cyber Attack Space

The cyber attack space is divided into two dimensions, the time-phased attack script and the attack classes.

Time-phased attack script

A generic time-phased attack script can best be described in three main phases: (a) gain access to the system, (b) escalate privileges, and (c) execute the attack by defeating a security service and obtaining the attack objective. Each phase is comprised of one or more attack steps. Other optional attack phases include diagnosis and reconnaissance of the system and avoiding detection by covering the tracks of the attack. To protect the system, the defender needs to mitigate only one of the steps in any one of the three generic attack phases. The goal of the system designer is to find the most cost-effective way to mitigate any potential attack. The system

designer must estimate the adversary's risk tolerance and willingness to expend resources to attack the system. Of course, mitigating multiple attack steps, or mitigating a single attack step in multiple ways, provides defense in depth and more assurance against implementation flaws in the countermeasures.

Table 1 shows an example of three-phase attack that uses CNA techniques to gain access and escalate privileges, and then defeats the integrity security service by publishing fraudulent data.

Cyber attack classes

Most cyber attacks involve, at a minimum, three generic attack steps: get access to the computer system, elevate privileges, and then defeat some desired security service (i.e., confidentiality, integrity, or availability). Cyber attacks are not limited to only CNAs, but may also include attack steps from other attack classes. The following attack classes represent the tools of the trade of the cyber adversary. The attack classes form the basis of the threats that the defender must counter.

Computer Network Attacks (CNAs)

CNAs are techniques that use the computer network against itself; the network itself becomes the avenue of attack. CNA techniques include stealing, corrupting, or destroying information resident in computers and computer networks or disrupting or denying access to the computers and networks themselves. CNAs defeat vulnerabilities in applications, operating systems (O/S), protocols, and networks. Sometimes known as *hacking*, when in the hands of a well-resourced, sophisticated adversary like a nation state, CNAs are lethal weapons against modern information systems.

CNA is useful in all three of the generic attack steps. An adversary can use CNAs to gain access to

Table 1 Three-phase attack example

Three-step attack script	Attack step
Get access	Hack through firewall from Internet Send malware <i>via</i> web page
Get privilege	Gain write privileges on engineering database
Defeat security service	Alter engineering data

the information by penetrating the network perimeter defenses such as firewalls and guards. Stealthy CNAs can avoid detection by defeating or disabling the intrusion detection systems (IDS). By targeting access control and privileges management mechanisms, CNA techniques can escalate the attacker's privileges from untrusted outsider to trusted insider. Finally, CNAs can disrupt network services by flooding critical network resources, and they can defeat confidentiality and integrity by exploiting vulnerabilities in applications and operating systems.

Some examples of CNA techniques include the following:

- **Malware** – is malicious software designed to exploit a vulnerability on a computer to infiltrate the security perimeter or steal or damage the computer system resources. Examples of malware include computer viruses, worms, Trojan horses, spyware, and other malicious and unwanted software. Malware should not be confused with defective software, that is, software that has a legitimate purpose but contains harmful bugs as a result of unintentional programming errors.
- **Botnet** – is a term for a collection of software robots, or bots, which run remotely across a computer network. The bots propagate from one infected computer to another by using malware (e.g., viruses, worms, and Trojan horses) to exploit a common vulnerability on the computers. The infected computers, or bots, are remotely controlled by the bot originator, and is usually used for nefarious purposes, such as wide-scale distributed denial of service (DDoS) attacks in which all the infected computers are commanded to send oversized Internet protocol (IP) packets to a single web server.
- **Flooding** – is a specific type of denial of service (DoS) attack techniques in which the attacker demands more computer resources than are available. The flooding could occur on the network by sending too many packets or by opening too many services, or at a device level by filling up the memory or tying up the central processing unit (CPU) with computationally intensive operations.
- **Eavesdropping** – is the act of passively snooping on a computer network and collecting private or sensitive information, such as user names and passwords. Eavesdropping requires access to the computer network, which is usually obtained by

installing a sniffer on the network or by collecting the transmissions from a wireless access point.

- **Malware propagation method** – malware spreads itself, or propagates from one infected host to another, in two basic ways: a computer virus infects a piece of executable code and replicates itself onto another piece of executable code, whereas a worm is self-replicating by exploiting a common vulnerability across a computer network.
- **Attack signatures** – are typical patterns exhibited by malicious code that are used for computer IDS. A virus scanner that searches all incoming e-mail and attachments for known viruses is an example of a signature-based intrusion detection system that uses a database of known virus signatures for its pattern-matching search.
- **Polymorphic code** – is malicious code that attempts to hide from signature-based IDS by continually mutating the code while keeping the original malicious algorithm intact.
- **Stealthy attacks** – are malicious codes that attempt to hide from the defender's intrusion detection system. The most effective stealthy cyber attacks are zero-day attacks (i.e., their attack signature is not in the defender's signature database), use polymorphic code, and do not exhibit anomalous user or network behavior (i.e., does not trigger an alarm in an anomaly-based intrusion detection system).

Special Operations (SPECOPS)

SPECOPS cyber attacks are physical attacks that include “black-bag” operations, midnight breaking and entering, and physical overruns to get access to the computer systems. These attacks usually involve highly trained covert operatives and counter-intelligence agents. Some typical “black-bag” operations might include breaking into a secure area to plant a bug in the room or to install an exploit in the computer or network device. SPECOPS attacks are useful for penetrating the physical perimeter to steal information or for destroying or altering the security-critical components. Only specially trained military units (e.g., information warriors) or organized crime would carry out such high-risk operations.

Human Intelligence (HUMINT)

HUMINT is the cyber attack discipline that uses insider access to steal information or gain privileged

access to critical system resources. The world's second oldest profession, HUMINT, comprises the time-honored practice of espionage (i.e., stealing or buying information from insiders). HUMINT includes bribery, blackmail, coercion, and social engineering to co-opt an insider to do the bidding of the sponsoring adversary (e.g., nation state, terrorist, organized crime, or economic competitor). Hackers have shown a penchant for social engineering, which is fooling an unwitting insider to unintentionally violate some security policy, such as divulging his user name and password. HUMINT is a useful technique for violating the physical and cyber defensive perimeter, as well as for gaining the necessary privileges to defeat the confidentiality, integrity, and availability of the critical information residing on the computer system.

Signals Intelligence (SIGINT)

SIGINT is the cyber attack discipline that captures network communications with the objective of stealing information from the target. SIGINT is the catchall phrase that includes listening to a target's communications by either intercepting wireless transmissions or by tapping a wire or cable. SIGINT attacks will target the information system's encrypted (black) network traffic. SIGINT also includes collecting compromising electromagnetic emanations (e.g., TEMPEST emanations), performing cryptanalysis on encrypted data, and using covert channels to exfiltrate sensitive information from the unencrypted (red) side of the cryptography.

Even if the SIGINT adversary cannot break the encryption, the SIGINT adversary can still perform traffic analysis on the encrypted signal. By analyzing the source, destination, frequency, and duration of each communications session, the SIGINT adversary can derive some intelligence about the nature of the communications and its participants.

By defeating the authentication system, the SIGINT adversary can perform masquerade attacks, or man-in-the-middle attacks, in which the SIGINT adversary appears to be a legitimate user of the network by stealing or forging the necessary cryptographic authentication credentials.

Electronic Warfare (EW)

EW includes any military action involving the use of electromagnetic and directed energy to control the

electromagnetic spectrum or to destroy an enemy's sensitive electronic systems. Directed high-energy attacks can jam radio frequency (RF) links or destroy fragile electronic components. Examples of EW include high-energy radio frequency (HERF), laser, and electromagnetic pulse (EMP). These high-energy attacks target the availability of critical communications components like satellite links.

Life Cycle

Life-cycle attacks introduce exploitable vulnerabilities to a security-critical component during its design, production, distribution, or maintenance. Some critical components, like the cryptography, require a trusted development and production team to counter this life-cycle threat. Other components, such as commercial routers and firewalls, have unknown pedigrees and are, therefore, more susceptible to life-cycle attacks. Life-cycle attacks are useful techniques to introduce vulnerabilities in the network perimeter defenses and access control mechanisms, as well as to introduce a DoS Trojan horse in the critical network components (e.g., backbone routers).

Management and Control

Since most modern networks span large geographical areas, the network operators and system administrators must rely on a remote management and control (M&C) subsystem to configure and maintain the correct network functionality. Unfortunately, this distributed, remote M&C capability provides a lucrative target for the cyber adversary. The M&C subsystem is usually an afterthought for system security engineers; consequently, the M&C subsystem is often protected using inadequate defensive mechanisms. M&C attacks target vulnerabilities in this remote subsystem, resulting in the adversary having the ability to reconfigure the critical network components. M&C attacks are useful for gaining access to the system by opening up holes in the boundary protection. They are useful for granting additional privileges by reconfiguring the access control and privilege management servers. Finally, M&C attacks are useful for defeating security services like availability by erasing the network routing tables.

Table 2 shows which attack classes typically accomplish each attack phase.

Table 2 Attack class pairing with attack script phases

Attack Script Attack Class	Get Access	Elevate Privileges	Defeat Security Service
CAN			
SIGINT			
Lifecycle			
HUMINT			
Special Ops			

Cyber Risk-Management Challenges

Cyber risk management faces a few unresolved key issues, including the following:

- Independent validation of cyber risk assessment results such as adversary behavior and probabilities of successful cyber attacks.
- The need for absolute metrics for risk as opposed to the relative metrics that current cyber risk assessment processes use.
- The drive toward accurate measurements of attack probabilities derived from authentic attack data collected from operational networks.
- The need to factor low-probability/high-impact cyber attacks, for which there is no historical evidence, into the risk assessment calculations.

References

- [1] Evans, S. & Wallner, J. (2005). Risk-based security engineering through the eyes of the adversary, *IEEE IA Workshop Proceedings*, IEE CS Press.
- [2] Salter, C., Saydjari, S., Schneier, B. & Wallner, J. (1998). Toward a secure system engineering methodology, *Proceedings of New Security Paradigm Workshop*, Charlottesville.
- [3] Schneier, B. (1999). Attack trees: modeling security threats, *Dr. Dobbs' Journal* **24**(12), 21–29.
- [4] Schneier, B. (2000). *Secrets and Lies: Digital Security in a Networked World*, John Wiley & Sons, Indianapolis.
- [5] Schneier, B. (2003). *Beyond Fear: Thinking Sensibly About Security in an Uncertain World*, Copernicus Books, New York.
- [6] Evans, S., Heinbuch, D., Kyle, E., Piorkowski, J. & Wallner, J. (2004). Risk-based systems security engineering: stopping attacks with intention, *IEEE Security and Privacy* **2**, 59–62.

Further Reading

- Amenaza (2002). *Predicting Adversarial Behavior with SecurITree*, at http://www.amenaza.com/downloads/docs/PABSecurITree_WP.pdf.
- Bahill, A.T. & Dean, F.F. (2003). *What is Systems Engineering? A consensus of Senior Systems Engineers*.
- Brooke, J. & Paige, R.F. (2003). Fault trees for security systems analysis and design, *Journal of Computer Security* **22**, 256–264.
- Buckshaw, D., Parnell, G.S., Unkenholz, W.L., Parks, D.L., Wallner, J.M. & Saydjari, O.S. (2004). Mission oriented risk and design analysis for critical information systems, Technical Report 2004-02, *Innovative Decisions*.
- Buckshaw, D. (2003). Information system risk assessment and countermeasure allocation strategy (presentation), *MORS Decision Aids/Support to Joint Operations Planning*, Innovative Decision, at http://www.mors.org/meetings/decision_aids/da_pres/Buckshaw.pdf.
- Futoransky, A., Notarfrancesco, L., Richarte, G. & Sarraute, C. (2003). *Building Computer Network Attacks*, at http://www.coresecurity.com/corelabs/projects/attack_planning/Futoransky_Notarfrancesco_Richarte_Sarraute_NetworkAttacks_2003.pdf.
- Kerzner, H. (1995). *Project Management: A systems Approach to Planning, Scheduling, and Controlling*, Reinhold, New York.
- Tidwell, T., Larson, R., Fitch, K. & Hale, J. (2001). Modeling Internet attacks, *Proceeding of the 2001 IEEE Workshop on Information Assurance and Security*, United States Military Academy, West Point.

JAMES WALLNER

Data Fusion

Loosely speaking, data fusion is combining information from different sources. This is done routinely in all areas of science and engineering, though often in an *ad hoc* manner informed only by constraints and objectives of a particular analysis. The topic is ever more important as new technologies lead to new ways of collecting data, and the pervasiveness of grid computing continues to increase our ability to form virtual data sets.

Consider the following motivating example. NASA's Earth Observing System (EOS) is a constellation of Earth-orbiting satellites each of which carries a number of remote sensing instruments. These instruments collect data on the Earth's atmosphere, oceans, and land surfaces, and are intended to be used in a synergistic manner to provide a complete picture of the health of Earth's most important systems. Each instrument produces data for a multitude of variables using its own sampling and transformation algorithms to convert raw measurements into geophysical variables. Data are stored in different physical locations, in different formats, and are broken up into files according to different rules.

The logistics of finding and obtaining EOS data have led to increased interest in web services at NASA and in the science community it serves. Web services allow users to pass arguments over the Internet and invoke programs on remote computers. In particular, where large volumes of data are concerned, it often makes more sense to push analysis algorithms to the computers where data reside, than to pull data to local computers on which analysis programs are developed. A natural extension of this idea is to install components of data fusion algorithms at data repositories in such a way that a driver program could assemble and fuse data from different locations and instruments "on the fly".

Such a vision requires a unified and principled theory of data fusion specifically designed to account for and make maximum use of statistical properties of the component data sets. This article does not pretend to develop such a theory, but rather to point out that a significant portion of it already exists in several domains: survey sampling, spatial statistics (see **Global Warming; Environmental Remediation; Geographic Disease Risk**), and the study of

copulas (see **Enterprise Risk Management (ERM); Credit Risk Models; Copulas and Other Measures of Dependency**). We also discuss some of the challenges posed by data fusion for understanding and quantifying risk, and end with a short summary.

What Is Data Fusion?

The term *data fusion* means different things to different people. Related terms that may or may not mean the same thing include information fusion, sensor fusion, and image fusion, depending on the context. In all cases the terms relate to the act of combining information or data from heterogeneous sources with the intent of extracting information that would not be available from any single source in isolation. Data and sensor fusion tend to connote military applications, such as combining different forms of intelligence or different satellite images to facilitate identification and tracking of objects or entities. Image fusion has a more engineering flavor, such as reconstruction of three-dimensional objects from sets of two-dimensional views. For example, in medical imaging interest may lie in the three-dimensional morphology of tumors reconstructed from a set of two-dimensional magnetic resonance imaging (MRI) scans. In this article we focus on a more specifically statistical definition: data fusion is the act of inferring a joint distribution when one only has information about the marginal distributions.

Statistical Theory and Development

Two main lines of research lead to a statistical theory of data fusion. One emerges from the area of statistical matching: business and marketing applications in which incomplete and incongruent sample survey information is to be combined. The second arises from spatial statistical applications in which incongruent spatial sampling and different spatial support necessitate a theory for making optimal inferences from such incompatible data sources. A third area, the study of copulas, is also relevant to data fusion, and we close this section with a very brief mention of that topic.

Statistical Matching

Statistical matching [1] addresses the following problem. One has access to two independently acquired

sample survey data sets and would like to combine the information therein. Few, if any, individuals are part of both surveys, but there are some variables (survey questions) in common. The goal is to create a single data set that can be regarded as a sample from the joint distribution of all variables of interest (see Figure 1).

Statistical matching suggests the need to match individual records from the two samples, but in fact this problem is more commonly known as *record linkage* (see **Cohort Studies**). In record linkage one presumes that more or less the same individuals, or units, are present in both samples, but the mapping from one survey sample to the other is uncertain owing to transcription, recording, or other errors [1]. Not surprisingly, research and methodologies for matching and linkage have common roots and the line between them continues to be fuzzy even today [2].

Sample 1		Sample 2	
X	Z	Y	Z
x_{11}	z_{11}	y_{21}	z_{21}
x_{12}	z_{12}	y_{22}	z_{22}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1N}	z_{1N}	y_{2M}	z_{2M}

Combined sample		
X	Z	Y
x_{11}	z_{11}	
	z_{21}	y_{21}
x_{12}	z_{12}	
	z_{22}	y_{22}
$\vdots$	$\vdots$	$\vdots$
x_{1N}	z_{1N}	
	z_{2M}	y_{2M}

Figure 1 Schematic view of the statistical data fusion problem in the context of statistical matching. Samples 1 and 2 record measurements on two sets of units which have few if any members in common. The combined sample results from concatenating samples 1 and 2, but results in missing values. There are a number of ways to impute the missing values with varying degrees of theoretical justification

Early statistical matching literature focuses on situations where one sample is identified as the “donor” and the other as the “recipient”. Typically, the donor is larger and more complete, but the recipient sample is of greater interest. The objective is to complete the recipient sample by inferring values on a variable that exists only in the donor, for all units in the recipient. For instance, this might be done by partitioning both samples into homogenous strata, and imputing the recipient’s missing values using regressions determined from suitable strata of the donor file. For the interested reader [1], provides a thorough history of the development and application of statistical matching theory and methods.

Regardless of what it is called, the fusion problem is also closely related to the missing data paradigm of Little and Rubin [3] (see **Repeated Measures Analyses**). Two survey samples, say with N and M units (records) respectively, can be concatenated to form a single sample of length $N + M$ in which the first N units have missing values for the variables unique to the second survey and the second M units have missing values for the variables unique to the first survey. Data fusion attempts to impute the missing data for purposes of making inferences about the underlying common population.

The framework of Little and Rubin explicitly models missingness as a random variable, which may be related to the survey variables themselves. The forms of those relationships are expressed, as with all relationships among variables, by the joint and conditional distributions of quantities of interest. Likelihood-based inference for parameters of the underlying joint distribution proceeds as usual by examining the probability of obtaining the observed sample as a function of the unknown parameters. If prior distributions are attached to the parameters, the approach is Bayesian. Missing data values are inferred from their posterior distribution, given the data. In this setting, emphasis is often placed on parameter estimation in presence of missing data rather than on data fusion *per se*.

If the primary objective is creation of a fused data set, the expectation-maximization (EM) algorithm [4] is a natural candidate for producing the desired results. EM produces maximum-likelihood estimates of unknown parameters by iteratively and alternately filling in missing data with their expected values given the observed data and the current best estimate of the parameter, and then updating

the parameter estimate to maximize the likelihood given the observed and (estimated) missing data. The EM algorithm does not necessarily converge to the global maximum of its optimization criterion, and as observed by Ressler [1], if there are no actual realizations that provide information about two variables to be fused, EM solutions may not be unique. Additional information or assumptions are required.

One way of accomplishing this is to impose prior distributions on the unknown parameters. Unfortunately, however, in the presence of missing data the computations become even messier than usual and in fact are often unsolvable analytically [1]. Markov chain Monte Carlo (MCMC) (*see Non-life Loss Reserving; Reliability Demonstration; Bayesian Statistics in Quantitative Risk Assessment*) techniques, already brought to bear for other computationally difficult Bayesian problems [5], can be used here as well. Once the posterior distributions of the missing data (given the observed data) and the unknown parameter (given the observed data) are estimated, random draws from the former can be used to fill in missing values. In fact, the principle of multiple imputation [6] uses multiple random draws to create multiple synthetic data sets in order to capture uncertainty inherent in filling in missing values. One might also simply carry the variance of the posterior distribution of the unobserved data the same way one carries and uses the variance of the presumed distribution of the observed data in further calculations on the fused data set.

To clarify and organize the data fusion framework, it is useful to introduce the following notation. Let X and Z be two random variables encoding responses from survey 1. Let Y and Z be two random variables encoding responses from survey 2. Let S_X and S_Z be indicators of nonmissingness for X and Z in the first survey, and let R_Y and R_Z be indicators of nonmissingness for Y and Z in the second survey. Finally, let $\{(X_n, Z_{1n})\}_{n=1}^N$ and $\{(Y_m, Z_{2m})\}_{m=1}^M$ be the observed samples from the two surveys respectively. Z_{1n} and Z_{2m} have the same distribution – that of Z . That is, $\{(X_n, Z_{1n})\}_{n=1}^N$ is a sample from the joint distribution of X and Z conditional on the fact that at least one of them is observed in the first survey and $\{(Y_m, Z_{2m})\}_{m=1}^M$ is a sample from the joint distribution of Y and Z conditional on the fact that at least one of them is observed in the second survey. We use the notation $[U]$ to indicate the distribution of a random variable U , and $[U|V]$ to

denote the conditional distribution of U given V . The first survey provides a sample from $[X, Z|S_X = 1 \text{ or } S_Z = 1]$. The second survey provides a sample from $[Y, Z|R_Y = 1 \text{ or } R_Z = 1]$.

If it were known, the entire joint distribution $[X, Y, Z, S_X, S_Z, R_Y, R_Z]$ could be used to obtain $[X, Y]$ by integrating out the other quantities. We only have samples corresponding to $[X, Z|S_X = 1 \text{ or } S_Z = 1]$ and $[Y, Z|R_Y = 1 \text{ or } R_Z = 1]$, so other assumptions must be made. Conditional distributions depending on $[X, S_X]$ define the bias, if any, of sample 1 relative to the unconditional distribution of X (both observed and unobserved). The same is true of $[Y, R_Y]$ for sample 2 and its relationship to Y . It is probably reasonable in most cases to assume that S_X and R_Y are independent, but it may not be reasonable to assume that S_Z and R_Z are independent depending on how the two samples were collected.

Exact and exhaustive specification of all relationships embodied in $[X, Y, Z, S_X, S_Z, R_Y, R_Z]$ depends on specific circumstances, and to the extent that future uses of the fused data set are known. However, more than one author has noted that conditional independence of X and Y on Z is required to avoid identifiability problems. Gilua *et al.* [7] provides a simple but thorough and clear discussion of this issue.

Spatial Applications

Data fusion in the context of spatial data sets has much in common with the statistical matching and missing data approaches. Both seek to impute missing values using assumptions about relationships among units and among variables. In statistical matching, the units are typically individuals – respondents in sample surveys. In spatial problems, the units might be measurement locations or individual pixels in a remote sensing image. See Figure 2.

Spatial problems highlight the notion of aggregation hierarchies: data to be fused may exist at different levels of aggregation. For instance, one may wish to fuse a data set with pixels that are 10 km^2 with one having pixels that are 17 km^2 . Moreover, the pixels may not be mutually exclusive: there may be spatial overlap and alignment issues. In other words, the units to be fused may not be congruent in the sense that discrete units (such as people) are.

These issues are known in the spatial statistics literature as change of support problems [8], but aggregation issues also exist in areas traditionally

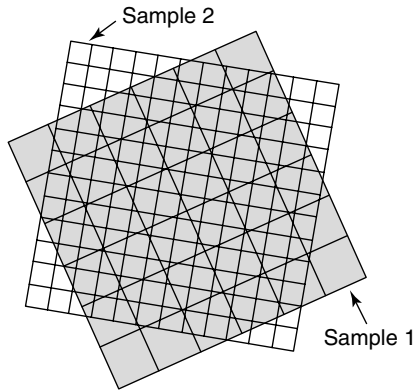


Figure 2 General problem of data fusion in the spatial context. The units of sample 1 are the pixels in the gray grid. The units of sample 2 are pixels of the white grid. The units overlap but are of different spatial resolutions, are misaligned, and are oriented differently

under the purview of statistical matching [9]. In fact the ecological fallacy that occurs when inferences are made about individuals based on statistical aggregates is known in both spatial [10] and nonspatial [11] applications.

Kriging [12] (*see Environmental Remediation*) makes explicit use of covariances among units to impute values for unobserved ones. This is often done entirely within the context of a single data set. However, if a second data set is available that measures the same quantity, one would ideally like to fuse the two to fill in gaps and to obtain better estimates where data from more than one source are available. Or, perhaps two data sets measure different things, and interest lies in understanding their joint relationship. That is, if the two variables measure the same underlying quantity, they can be averaged in some way, and if they measure different quantities, they can be concatenated. One way around incongruent units is to krig each data set separately to a common set of locations and then fuse at each location in the same way one would had real measurements been available. Note however that the kriged field may exhibit excessive spatial dependence since the same set of observed data gives rise to kriged estimates at multiple locations. Further, optimal statistical fusion of kriged data sets may be inferior to optimal statistical fusion of the original data sets (Noel Cressie, 2007, personal communication).

Kriging deals with the change of support problem by exploiting the continuity of the underlying field and the linearity of the covariance operator. The usual spatial statistical model ([12] for example) assumes that the quantity of interest varies continuously over its domain; that is, it is a “point level” phenomenon, which varies from geographic point to point. Block kriging [12] uses the statistical characterization of point level relationships (covariances) to derive the covariance structure between “blocks” of aggregated measurements. For example, remote sensing instruments observe pixels that are spatial averages of the measured quantity over the footprint of the pixel. Since averages are linear functions, the covariances at the point level can be used to obtain covariances at the block (pixel) level.

It is interesting to consider applying the same strategy in the context of statistical matching. One might postulate a covariance function for the units based on demographic or other auxiliary variables that play the role of spatial location. If all the units represent discrete objects (e.g. people), the situation is analogous to kriging point data: predict missing response values as linear combinations of existing response values in the same data set. Then, given two imputed data sets, combine records from units that are similar with respect to the auxiliary variables by forming linear combinations of the donors. This is in fact a strategy similar to that described in [1, Chapter 2], and the same caveats expressed above for spatial data fusion apply.

The idea of common variables in statistical matching is close to that of auxiliary variables in co-kriging, followed by fusion of some sort at common locations. Missing data indicators could also be incorporated into the co-kriging framework as additional covariates to account for sampling biases. However, this is more often accomplished in spatial situations using Bayesian hierarchical models [13] that relate measurements to the true quantity of interest through an error term. The difference is that in statistical matching attention is focused primarily on the unknown joint distribution of two variables from two different samples, while in spatial applications attention is focused more on filling in missing information. If the further step of combining information from two separately kriged samples is taken, the fused data can be used to estimate characteristics of the joint distribution. Note that the latter may not be statistically

optimal because it does not bring to bear all the information available from both data sources at the initial stage.

A recent article in the *Journal of Marketing Research* by Gilula *et al.* [7] takes a significant step toward unifying the matching and spatial approaches. They use a Bayesian hierarchical model to (*see Geographic Disease Risk; Meta-Analysis in Nonclinical Risk Assessment*) "... estimate the joint distribution of the target variables directly rather than to use a matching or concatenation approach. The joint distribution can then be used to solve the inference problems It is important to emphasize that our methods are 'automatic' in the sense that they do not require modeling or in-depth analysis of the nature of the common demographics variables." Gilula *et al.* do not address the problem of incongruent units that arises in spatial applications. Their approach could be extended using Bayesian hierarchical models to cover this circumstance [12].

Copulas

Copulas are mathematical functions that link multidimensional distributions to their one-dimensional marginal distributions [14]. Briefly, if X and Y are jointly distributed random variables with joint distribution function $H(x, y)$, and marginals $F_X(x)$ and $F_Y(y)$, respectively, then $V = F_X(X) \sim U(0, 1)$ and $W = F_Y(Y) \sim U(0, 1)$ where $U(0, 1)$ indicates the uniform distribution on the unit interval. The distribution function of (V, W) is a copula:

$$\begin{aligned} C(v, w) &= P(V \leq v, W \leq w) \\ &= P(X \leq F_X^{-1}(v), Y \leq F_Y^{-1}(w)) \\ &= H(F_X^{-1}(v), F_Y^{-1}(w)) \end{aligned} \quad (1)$$

The subject is highly mathematical, and of significant importance in statistics because it provides a framework within which different forms of dependence may be studied. In principle, copulas offer an analytical solution to problems estimating unknown joint distributions when only the marginal distributions are available. The link to data fusion is self-evident.

Copulas have received plenty of attention in econometrics, finance, risk analysis, and actuarial science because of their usefulness in studying nonlinear dependence and modeling joint distributions.

At present there does not appear to be a direct connection to the data fusion literature discussed earlier in this article, but it seems clear that the motivating ideas behind copulas could offer some theoretical guidance in problems related to statistical matching and fusion.

What Challenges Does Data Fusion Pose for Risk Analysis?

Data fusion is often seen as an end in itself: the goal is the production of a data set, which is then the subject of subsequent analyses. Challenges for quantitative risk assessment arise in those analyses from the possibility that assumptions used to perform the fusion are not correct. In particular, if data fusion is performed as a preprocessing step without knowledge of the subsequent analyses, the validity of conclusions may be compromised.

In this regard, the approach based on the spatial paradigm has the advantage of providing a well-specified goal that does not depend on future use: accurate estimation of the true, underlying field. The approach also automatically provides uncertainties for those estimates.

An approach based on the statistical matching and missing data paradigms has the advantage of providing a more direct path to a fused data set without resorting to a model-based specification of the underlying field. Its disadvantage is that the resulting fused data set is more vulnerable to the assumptions used to create it precisely because it is not anchored to such a model. Moreover, conditional independence assumptions are required to avoid identifiability problems.

Perhaps the greater risk posed by data fusion is that of *not* performing it. The new classical example is failure to "connect the dots" in analysis of data from disparate sources of intelligence information (e.g., video, network, email, financial, etc.). That particular example is especially difficult because it involves different types of data. The content of this article only deals with numerical information, but clearly even greater challenges arise for multimodal data types.

Summary

We defined the objective of data fusion to be that of combining statistically heterogeneous samples to

construct a new sample that can be regarded as having come from an unobserved joint distribution of interest. The building blocks of a statistical theory of data fusion already exist in various mature areas of statistics. Elements can be found in the literature on statistical matching, missing data, and spatial statistics, and at a more theoretical level, the study of copulas. These threads need to be drawn together to provide a comprehensive framework that is internally consistent and rigorous. We hope that such a framework can provide new, quantitative methods for data fusion, and place both new and existing methods in a common frame of reference so they can be more easily compared and evaluated.

Acknowledgment

The research described in this article was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

- [1] Ressler, S. (2002). *Statistical Matching, Lecture Notes in Statistics*, Springer, New York.
- [2] Winkler, W.E. (2006). *Overview of Record Linkage and Current Research Directions*, Research Report Series Statistics, 2006-2, United States Census Bureau, Statistical Research Division, Washington, DC.
- [3] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.
- [4] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
- [5] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6), 721–741.
- [6] Rubin, D.B. (1976). Inference and missing data, *Biometrika* **63**, 581–592.
- [7] Gilua, Zvi., McCulloch, R.E. & Rossi, P.E. (2006). A direct approach to data fusion, *Journal of Marketing Research* **43**, 1–22.
- [8] Gotway, C.A. & Young, L.J. (2002). Combining incompatible spatial data, *Journal of the American Statistical Association* **97**, 632–648.
- [9] Moulton, B.R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units, *The Review of Economics and Statistics* **72**(2), 334–338.
- [10] Gotway Crawford, C.A. & Young, L.J. (2004). A spatial view of the ecological inference problem, in *Ecological Inference*, O.R. Gary King & M.A. Tanner, eds, Cambridge University Press, London.
- [11] Freedman, D.A. (1999). Ecological inference and the ecological fallacy, Technical Report Number 549, Department of Statistics, University of California, Berkeley.
- [12] Cressie, N.A.C. (1993). *Statistics for Spatial Data*, John Wiley & Sons, New York.
- [13] Banerjee, S., Carlin, B.P. & Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall, New York.
- [14] Nelsen, R.B. (2006). *An Introduction to Copulas*, Springer, New York.

AMY J. BRAVERMAN

Decision Analysis

Decision analysis as a term could describe many methodologies that analyze the elements of a choice that is to be made and recommend a decision. However, by convention – in Anglo-American literature at least – it refers to analyses based upon the subjective expected utility model (see **Subjective Expected Utility**) of rational economic man developed during the nineteenth and early twentieth centuries in the economics literature. In parallel, the Bayesian approach to statistics (see **Bayesian Statistics in Quantitative Risk Assessment**) developed the identical model in its approach to inference and decision [1, 2]. Nonetheless, it should be recognized that there are several quite distinct approaches to decision analysis that are not covered in this article [e.g. [3–5]]. The defining characteristics of the Bayesian decision analytic approach that we describe here is that in certain respects it is explicitly subjective, modeling judgments of uncertainty and preference. Uncertainty is modeled through subjective probability (see **Subjective Probability; Interpretations of Probability**) and preference through utility functions (see **Risk Attitude; Utility Function**). In some decision analyses – though seldom have those that arise in the context of risk analyses – uncertainty is not an issue and the focus is on modeling and exploring preference, particularly in terms of balancing conflicting objectives. In such cases, preferences may be modeled through value functions (see **Value Function; Multiattribute Value Functions**) [6].

It is easy to think of decision analysis as a set of quantitative calculations that draw together subjective and objective information and so identify an “optimal” decision. However, a full understanding of decision analysis requires that one sees it as a supportive process of interactions with the decision makers through which their understanding of the external context, the choice before them, and their own judgments of uncertainty and value all evolve until the way forward becomes clear. Decision analysis may be structured around quantitative calculations, but it is the process that is key; and this process will be discussed in this article.

Bayesian Analysis

The subjective expected utility model that is central to Bayesian decision analysis is very simple. Note that we assume for the present that there is just an individual decision maker; we discuss decision analysis for groups later. One begins with underlying model of the world and the decision maker’s potential interactions with it: $c(a, \theta)$. Here $c(\cdot, \cdot)$ maps a potential *action*, $a \in A$, that the decision maker may take from a set of alternatives A to a *consequence*, $c \in C$, under the assumption that the exogenous *state of the world* may be described by some parameter $\theta \in \Theta$. Note that in principle the consequence set may be very general; but in practice the modeling process usually reduces this to an n dimensional space, where each dimension represents an *attribute* that is important to the decision maker in evaluating the options (see **Multiattribute Modeling**). The function $c(\cdot, \cdot)$, in a sense, embodies the objective aspects of the analysis, encapsulating relevant knowledge drawn from economics, engineering, environmental science, or wherever. The decision maker’s subjective judgments are encoded in two functions: a *subjective probability* distribution (see **Subjective Probability**), $P_\theta(\cdot)$, representing his or her beliefs about the actual state of the world θ and a (multiattribute) *utility* function, $u(\cdot)$, representing his or her preferences for the potential consequences. Subjective expected utility theory then ranks the alternative actions according to the increasing value of the expected utility. Thus the theory prescribes decision making by maximizing expected utility:

$$\max_{a \in A} \int_{\Theta} u(c(a, \theta)) dP_\theta \quad (1)$$

The analysis is simply connected with Bayesian statistics and forecasting because $P_\theta(\cdot)$ may be taken as the posterior distribution for θ after any available data has been analyzed [1, 7]. The basic structure Bayesian decision analysis is thus as shown in Figure 1. Also note that it is seldom the case that decisions are taken in isolation; thus the analysis may cycle through a sequence of decisions [1, 8].

There are, of course, many embellishments to this structure of decision analytic calculations. For instance, the above assumes that the alternative actions are simple entities and thus relates to a decision table model of the underlying choice (see **Decision Modeling**). Most decisions, however,

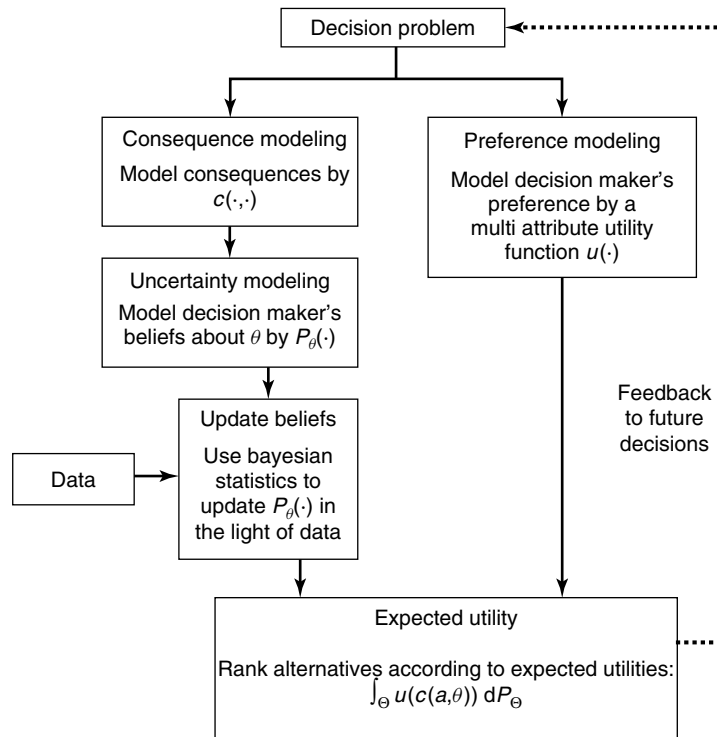


Figure 1 The structure of Bayesian decision analysis [Reproduced from [7] with permission from Springer, © 2003.]

involve contingencies, temporal relationships, and influences between the various actions and uncertainties. Thus the underlying decision model may be *decision tree* (see **Decision Trees**) or *influence diagram* (see **Influence Diagrams**). Also the analysis can give an estimate of the value of information [9]. In essence, this involves calculating the maximum expected utility without any data and the maximum expected utility if data were analyzed and the decision maker's uncertainties updated. The difference indicates the value that the data may provide the decision maker, though it is in terms of the expected utility. Thus it may be necessary to “invert” the utility function to provide a financial value of the information. There are several ways to construct upper bounds, e.g., the value of perfect information or the value of clairvoyance, on the value of information, which can be easier to calculate than the “expected maximum expected utility”, although it should be noted modern decision analytic software usually renders the fuller calculations tractable.

The Decision Analysis Process

The subjective expected utility model – at least as used within decision analysis – is a normative model, i.e., it models a conception of rational choice behavior. It is built on a set of axioms, which encapsulate or, rather, define rationality in decision making [1]. There is no pretense that the model describes actual behavior. Indeed, unguided decision making seldom fits entirely with this conception of rationality (see **Behavioral Decision Studies**). Subjective expected utility is seldom an adequate descriptive model of real behavior. Thus to use decision analysis in practice, there is a need to address the discrepancy between the rationality assumed of the decision maker and the behavior that he or she is likely to exhibit in practice. This means that the subjective expected utility model is embedded within a prescriptive process of decision support in which the decision maker is guided toward the rationality assumed within the model, mindful of the cognitive limitations and behavioral biases that he

or she is likely to exhibit in interacting with the analysis [2]. To achieve this, decision analysis broadly proceeds through the following stages.

Stage 1. *Problem formulation*

Decision problems seldom arrive fully formulated. The decision maker may be aware that a number of issues need resolving, but have little idea of what the options are or even what he or she is trying to achieve. Thus the first step is to explore the issues and concerns with the decision maker and structure a model of them. There are many creative and catalytic ways of interacting with the decision maker that can help here [10–12]. As the issues become clear, they can be structured into decision tree or influence diagram with multiattributed consequences (see **Decision Modeling; Decision Trees; Influence Diagrams; Multiattribute Value Functions**). Note that this activity does not just “build the model”, but also clarifies the decision maker’s thinking and perception of the issues that he or she faces. In knowledge management terms, it both creates and codifies knowledge [13].

Stage 2. *Elicitation*

The decision maker’s beliefs and preferences are elicited and modeled as subjective probabilities and utilities, respectively [14, 15]. This is a reflective process in which the decision maker is guided toward a consistent set of judgments that are modeled by the two functions. It is not a case of fitting functions to the decision maker’s responses to questions. There is mutual convergence; the decision maker’s self understanding and judgments may evolve as the elicitation progresses, guiding him or her toward the rationality assumed within the model.

Stage 3. *Calculation of expected utilities*

With the functions elicited, the quantitative analysis can be conducted. Bayesian methods can update the decision maker’s initial beliefs in the light of any data, the expected utilities calculated, and a ranking of the possible actions produced.

Stage 4. *Sensitivity analysis and critical evaluation*

The output of any analysis might be the spurious result of some imprecision in the inputs; so a sensitivity analysis should be always conducted. As many of the inputs in a decision analysis are based upon judgment and thus limited in their accuracy by human cognitive abilities, there is an even

stronger case for this in a decision analysis. This may be conducted by simply varying one or more input variables deterministically to see how the output ranking changes or by Monte Carlo methods (see **Uncertainty Analysis and Dependence Modeling**) [7, 16]. In fact, this fourth stage is often broadened to explore the decision analysis in many ways, e.g., varying some of the models, introducing dependencies and nonlinearities, and so on. The idea is not just to ensure that the output ranking is robust in the face of reasonable variations in the inputs and modeling assumptions, but also to help the decision maker understand the import of the analysis. Decision analysis should not simply indicate a choice of action; it should primarily aid the decision maker in understanding. It is through this understanding that the decision maker reaches the decision.

Stage 5. *Requisite?*

The final stage is for the decision maker to judge whether the analysis is a requisite for his or her purposes [1, 2, 17]. Has it brought him or her enough confidence to go ahead and decide; or are there still aspects of the situation that he or she feels are unmodeled and which, if included in the analysis, might lead to a different conclusion. In the former case, the analysis stops and the decision maker decides. In the latter case, the analysis cycles returning to an earlier stage and exploring a more sophisticated model.

This description of the decision analytic process is necessarily simplistic. The stages may not be clearly delineated nor the operations confined to a single stage. For instance, French [7] argues that sensitivity analysis should permeate all stages, driving the modeling and shaping of the elicitation procedures, in addition to evaluating the robustness of the conclusions. It should also be mentioned that if the decision maker rejects the guidance of the process, as he or she might, if in the elicitation process he or she cannot provide responses consistent with the underlying Bayesian model, then this form of decision analysis is not for them. Other approaches, such as [3–5] mentioned above, may be more suited to them, though a Bayesian might find them irrational.

Group versus Individual Decision Analysis

The subjective expected utility model is essentially a model of the decision making of a rational individual. It does not apply to groups. Indeed, Arrow

has shown that in effect ideals of rationality and ideals of democracy are in conflict (*see Group Decision*) [1, 18]. On reflection this is not too surprising; decision making requires an expression of will and free will resides in individuals, and not in groups. Groups should be seen as social processes, which translate the decisions of the individual members into an implemented action [18, 19]. In such circumstances, the role of decision analysis is extended so that it does support not only the growth of each group member's understanding, but also supports communication between group members and the growth of shared understanding and – hopefully – consensus. Thus the decision analytic process described in the previous section needs to be embedded in a more discursive and deliberative process within the group, for instance, as in a decision conference or facilitated workshop for a small group (*see Decision Conferencing/Facilitated Workshops*) or a public participation process for larger groups (*see Public Participation*).

It should be recognized that not everybody who takes part in a decision analysis is a decision maker (*see Players in a Decision*). For instance, the decision makers may seek advice of experts; indeed, in very many contexts they would be exceedingly wise to do so. In such cases, the experts provide the subjective probability modeling of uncertainty or at least have a substantial input into this. The incorporation of expert judgment into the elicitation and modeling process requires some subtlety (*see Expert Judgment*) [20]. Similarly, the decision makers may identify a number of stakeholders whose views they may wish to take into account in the preference modeling. Thus the elicitation of the (multiattribute) utility function may reflect the judgments of stakeholders who are not decision makers themselves (*see Supra Decision Maker*).

Discussion

We noted in the introduction that there are other schools of decision analysis. Why do we promote the one described here? There are many reasons. French and Rios Insua [1] argue for the validity of this approach on six grounds:

- it has an axiomatic base that encodes persuasive tenets of rationality;
- there is a lack of counterexamples, which might cause one to question the analysis;
- it is feasible – the calculations are, by and large, straightforward, given modern algorithms and software;
- sensitivity and robustness analyses allow one to explore the interplay between the objective and subjective inputs and their effect on the conclusions;
- with modern graphical methods, it is transparent to users, bringing them understanding and not just a quantitative prescription;
- it is compatible with many worldviews and philosophies, i.e., it can be tailored to the perceptions of the decision makers.

In addition, we may note in the context of this encyclopedia that with its emphasis on modeling uncertainty with probability, decision analysis coheres with the general approach of risk analysis. Indeed, in many respects the literatures of decision and risk analysis are virtually indistinguishable. Perhaps decision analysis is more even handed in treating uncertainty and value with risk analysis, placing more emphasis on uncertainty modeling, but the underlying principles are to all intents identical.

References

- [1] French, S. & Rios Insua, D. (2000). *Statistical Decision Theory*, Kendall's Library of Statistics, Arnold, London.
- [2] French, S. & Smith, J.Q. (eds) (1997). *The Practice of Bayesian Analysis*, Arnold, London.
- [3] Bellman, R.E. & Zadeh, L.A. (1970). Decision making in a fuzzy environment, *Management Science* **17**(4), B141–B164.
- [4] Roy, B. (1996). *Multi-Criteria Modelling for Decision Aiding*, Kluwer Academic Publishers, Dordrecht.
- [5] Saaty, T.L. (1980). *The Analytical Hierarchy Process*, McGraw-Hill, New York.
- [6] Belton, V. & Stewart, T.J. (2002). *Multiple Criteria Decision Analysis: An Integrated Approach*, Kluwer Academic Press, Boston.
- [7] French, S. (2003). Modelling, making inferences and making decisions: the roles of sensitivity analysis, *TOP* **11**(2), 229–252.
- [8] DeGroot, M.H. (1970). *Optimal Statistical Decisions*, McGraw-Hill, New York.
- [9] Howard, R.A. & Matheson, J.E. (2005). Influence diagrams, *Decision Analysis* **2**(3), 127–143.
- [10] Rosenhead, J. & Mingers, J. (eds) (2001). *Rational Analysis for a Problematic World Revisited*, John Wiley & Sons, Chichester.

- [11] Mingers, J. & Rosenhead, J. (2004). Problem structuring methods in action, *European Journal of Operational Research* **152**, 530–554.
- [12] Franco, A., Shaw, D. & Westcombe, M. (2006). Problem structuring methods, *Journal of the Operational Research Society* **57**(7), 757–878.
- [13] Nonaka, I. & Toyama, R. (2003). The knowledge-creating theory revisited: knowledge creation as a synthesising process, *Knowledge Management Research and Practice* **1**(1), 2–10.
- [14] O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, R., Garthwaite, P.H., Jenkinson, D., Oakley, J.E. & Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*, John Wiley & Sons, Chichester.
- [15] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Trade-offs*, John Wiley & Sons, New York.
- [16] Rios Insua, D. & Ruggeri, F. (2000). *Robust Bayesian Analysis*, Springer-Verlag, New York.
- [17] Phillips, L.D. (1984). A theory of requisite decision models, *Acta Psychologica* **56**, 29–48.
- [18] French, S. (2007). Web-enabled strategic GDSS, e-democracy and Arrow's theorem: a Bayesian perspective, *Decision Support Systems* **43**, 1476–1484.
- [19] Dryzek, J.S. & List, C. (2003). Social choice theory and deliberative democracy: a reconciliation, *British Journal of Political Science* **33**(1), 1–28.
- [20] French, S. (1985). Group consensus probability distributions: a critical survey, in *Bayesian Statistics 2*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith, eds, North-Holland, pp. 183–201.

Related Articles

Risk and the Media

Stakeholder Participation in Risk Management Decision Making

Role of Risk Communication in a Comprehensive Risk Management Approach

SIMON FRENCH

Decision Conferencing/Facilitated Workshops

Decision conferencing provides a structured group approach for assessing the risks associated with several options, with expert judgment (*see* **Expert Judgment**) providing some or all of the inputs and models based on decision theory used to combine the inputs into an overall assessment. A facilitated workshop brings together experts about a specific risk criterion, with an impartial facilitator taking the group through the process of discussing the meaning of each criterion, establishing a measurement scales, and appraising options on the scales. At a subsequent decision conference including the key players, the facilitator guides participants through the process of making the final judgments about inputs and trade-offs between criteria that allow the decision theory model to bring together all the inputs to provide an overall ordering of the options.

The primary reason for working with groups of experts and key players is to benefit from their collective experience, knowledge, and wisdom [1]. Face-to-face debate and deliberation circumvents risk assessments that are anchored in one expert's experience, and generate consensus assessments that are different from the averages of individual expert's judgments [2]. In addition, the group-generated assessments carry an authority that is greater than a risk assessment given by just one individual, and participants are committed to the group's product. It is the modeling and group discussion, rather than a watered-down compromise, that makes it possible for the group to achieve a position that often represents a new perspective on the risk issues (*see* **Group Decision**).

In the sections to follow, the workings of a decision conference and its stages, the role of the facilitator, the use of special group decision support rooms, benefits of decision conferencing, why they work, a brief history, when decision conferencing is appropriate, and guidelines for their use are explained. A case study then shows how facilitated workshops and decision conferences enabled the United Kingdom's Committee on Radioactive Waste Management to engage in a public consultation and risk assessment

exercise that led to formulating and recommending 15 policies that were accepted by the Secretary of State for the Environment.

The Decision Conferencing Process

Decision conferencing is a series of intensive working meetings, called *decision conferences*, attended by groups of people who are concerned about some complex issues facing their organization [3]. There are no prepared presentations or fixed agenda; the meetings are conducted as live, working sessions lasting from 1 to 3 days. A unique feature is the creation, on the spot, of a computer-based model that incorporates data and the judgments of the participants in the groups. The model is often based on multicriteria decision analysis (MCDA) [4] (*see* **Multiaattribute Modeling**), which provides ample scope for representing both the many conflicting objectives expressed by participants and the inevitable uncertainty about future consequences. MCDA models are particularly useful for showing how costs, risks, and benefits can be balanced in selecting the best options. The model is a "tool for thinking", enabling participants to see the logical consequences of differing viewpoints and to develop higher-level perspectives on the issues. By examining the implications of the model, then changing it, and trying out different assumptions, participants develop a shared understanding and reach agreement about the way forward.

Stages in a Typical Decision Conference

Four stages typify most decision conferences, though every event is different. The first phase is a broad exploration of the issues. In the second stage, a model of the participants' judgments about the issues is constructed, incorporating available data. All key perspectives are included in the model, which is continuously projected so that all participants can oversee every aspect of creating the model. In the third stage, the model combines these perspectives, reveals the collective consequences of individual views, and provides a basis for extensive exploration of the model, which is always done on-line. Discrepancies between model results and members' judgments are examined, causing new intuitions to emerge, new insights to be generated, and new perspectives to be revealed. Revisions are made and further discrepancies explored; after several iterations, the new results

and changed intuitions are more in harmony. Then the group moves on to the fourth stage summarizing key issues and conclusions, formulating next steps and, if desired, agreeing on an action plan or a set of recommendations. The facilitator prepares a report of the event's products after the meeting and circulates it to all participants. A follow-through meeting is often held to deal with afterthoughts, additional data, and new ideas.

Role of the Facilitators

The group is aided by one or two facilitators from outside the organization who are experienced in working with groups. The main tasks of the facilitators are to see and understand the group life, and to intervene, when appropriate, to help the group deal with the current issues and maintain a task orientation to its work. The facilitators attend to the processes occurring in the group and provide structure for the group's tasks, but refrain from contributing to content. They structure the discussions, helping participants to identify the issues and think creatively and imaginatively. The facilitators help participants in *how* to think about the issues without suggesting *what* to think [5]. By refraining from contributing to the content of discussions, the facilitators maintain their impartiality, which is not always an easy task, especially when they are knowledgeable about the content. In those cases, it is best to use that specialist knowledge to frame questions to the group, rather than providing answers. The facilitator acts as a process consultant, always trying to be helpful, mindful that it is the client who owns the problem and the solution and that everything the facilitator does is an intervention. These three principles are just 3 of the 10 process consultancy principles described by Schein [6].

A key function of the facilitators is to help the group deal with the anxiety that inevitably accompanies group work: the group establishes a life that is not always congruent with the lives of the individuals in the group. A covert "basic assumption" develops: work beneath the surface that hinders accomplishment of the overt task. To date, research has identified five basic assumptions: fighting with each other or the facilitator, pairing of two people who the group expects will protect the group's security, becoming dependent on an individual in the group or on the facilitator to "save" the group, developing an intense

sense of belongingness, and acting as if the individual is the only reality [7, 8]. This is an aspect of the complex workings of groups, explored for many decades at the Tavistock Institute of Human Relations, and the subject of an anthology reporting the research findings of explorations in experiential groups [9].

Group Decision Support Rooms

For a group to work effectively, participants require an environment that is conducive to problem solving. Two basic principles help to ensure lively interaction:

1. all participants should be in direct eye-to-eye contact with each other; and
2. participants should be able to see and read all displays, flip charts, projection screens, and other visual aids.

Most work rooms are equipped with rectangular tables, often arranged in a [] shape. This arrangement provides direct eye contact from one side to the other, with those seated at the top table able to see everyone, but participants lined up on the sides unable to see through each other. Conversations are often directed from one side of the table to the other, less often along the same side, and frequently from the top. A better arrangement is more circular: $\cap$ shape, with an opening at the front to create a space for projecting the output of a computer.

Cabaret style, shown in Figure 1, provides an effective arrangement: several circular tables each seating four to six people. By slightly staggering the tables it becomes possible for the person talking to be seen by anyone, with just a slight shift in body position. In addition, this arrangement provides opportunities for the facilitator to give the same task to each table, with participants working on the task briefly and then reporting back to the whole group. In this way, differences within each group are worked on, often to consensus, with new differences appearing after each group reports back, which are then to be worked on by the whole group. In this way, the cabaret-style seating encourages everyone at a table to participate, everyone gets a chance to have their say, and all perspectives are fairly aired. Groups are often surprised by the diversity of opinion that is revealed, leading participants to realize that their own views are more anchored in their past experience than they had appreciated, leading them to

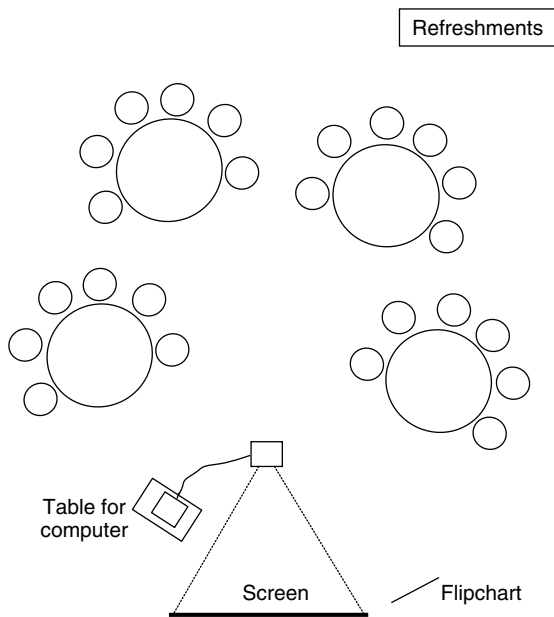


Figure 1 Cabaret-style room layout for a decision conference

become more receptive to alternative points of view. A useful discussion of room arrangements is given by Hickling [10].

Benefits of Decision Conferencing

The marriage in decision conferencing of information technology, group processes, and modeling of issues provides value-added to a meeting that is more than the sum of its parts. Follow-up studies, conducted by the Decision Analysis Unit at the London School of Economics and by the Decision Techtronics Group at the State University of New York, of decision conferences in the United Kingdom and the United States, for organizations in both the private and public sectors, consistently show higher ratings from participants for decision conferences than for traditional meetings [11, 12]. Organizations using decision conferencing report that the process helps them to arrive at better and more acceptable solutions than can be achieved using usual procedures, and agreement is reached more quickly. Many decision conferences have broken through stalemates created previously by lack of consensus, by the complexity of the problem, by vagueness and conflict of objectives, by ownership

in “fiefdoms”, and by failure to think creatively and afresh about the issues.

Why Decision Conferencing Works?

Decision conferencing is effective for several reasons. First, participants are selected to represent all key perspectives on the issues, so agreed actions are unlikely to be stopped by someone else arguing that the group failed to consider a major factor. Second, with no fixed agenda or prepared presentations, the meeting becomes “live”, the group works in the “here and now”, and participants get to grips with the real issues that help to build consensus about the way forward. Third, the model plays a crucial role in generating commitment. All model inputs are generated by the participants and nothing is imposed, so the final model is the creation of the group and is thereby owned by participants. Perhaps most important, the model helps to minimize the threat to individuality posed by the group life: the model reveals higher-level perspectives that can resolve differences in individual views, and through sensitivity analysis shows agreement about the way forward in spite of differences of opinion about details. Fourth, computer modeling helps to take the heat out of disagreements. The model allows participants to try different judgments without commitment, to see the results, and then to change their views. Instant play back of results that can be seen by all participants helps to generate new perspectives and to stimulate new insights about the issues. Wishy-washy compromises are avoided.

A Brief History of Decision Conferencing

Decision conferencing was developed in the late 1970s by Dr. Cameron Peterson and his colleagues at Decisions and Designs, Inc., largely as a response to the difficulty in conducting a single decision analysis for a problem with multiple stakeholders, each of whom takes a different perspective on the issues. The approach was taken up in 1981 at the London School of Economics and Political Science (LSE’s) Decision Analysis Unit by Dr. Larry Phillips, who integrated into the facilitator’s role many of the findings about groups from work at the Tavistock Institute of Human Relations. The service and supporting software continued to be developed throughout the 1980s and 1990s at the LSE in association with International Computers Limited and Krysalis Limited, and now

in the 2000s through Catalyze Limited. As decision conferencing spread around the globe, facilitators felt a need to share experiences, so they created the International Decision Conferencing Forum, which meets annually. Decision conferencing is now offered by about 20 organizations located in the United Kingdom, the United States, Portugal, Australia, New Zealand, and Hungary.

When Decision Conferencing is Appropriate

Decision conferencing can be applied to most major issues facing private organizations, government departments, charities, and voluntary organizations. Topics typically cover operations, planning, or strategy. Risk assessment forms a part of nearly every decision conference, for the generic issue facing all organizations is how best to balance costs, benefits, and risks. Risk is a component in all these decision conference topics: to evaluate alternative visions for the future; to prioritize R&D projects and create added value; to design factories, ships, and computer systems; to resolve conflict between groups; to allocate limited resources across budget categories; to evaluate the effectiveness of government policies, schemes, and projects; to improve utilization of existing buildings and plant; to determine the most effective use of an advertising budget; to assess alternative sites for a technological development; to deal with a crisis imposed by potentially damaging claims in a professional journal; to develop a strategy to respond to a new government initiative; and to create a new policy for health care provision. Any issue that would benefit from a meeting of minds in the organization can be effectively resolved with decision conferencing, which provides a way for “many heads to be better than one”.

Guidelines

Experience shows that decision conferencing works best in organizations when four conditions are met reasonably well. First, the style of decision making in the organization should allow for consultation and deliberation, provided there is adequate time to do so. Second, the organization should be open to change, for decision conferencing is usually experienced as a very different way to deal with complex issues. Third, a climate of problem solving should exist, so that options can be freely explored. Finally, authority and

accountability should be well distributed throughout the organization, and should be neither concentrated at the top nor totally distributed toward the bottom. When these conditions are met, decision conferencing can release the creative potential of groups in ways that enable both the individual and the organization to benefit.

A Case Study

After effectively ignoring the issue of what to do with the accumulating radioactive waste in the United Kingdom, the UK government created the Committee on Radioactive Waste Management (CoRWM) in November 2003. This committee was asked to start with a blank slate, consider all possible policies for managing the UK’s radioactive waste, and recommend by July 2006 on what should be done. But in doing so, they were to work openly with full transparency, to be inclusive, and to make recommendations that inspire public confidence.

From the spring of 2005, this author and colleagues facilitated many workshops with groups of experts who assessed value scores of the options on more than two dozen criteria comprising an MCDA model. A final decision conference with the CoRWM members at the end of March 2006 elicited criterion weights, and many sensitivity analyses looked at how overall weighted preference scores might change with different input assumptions about scores and weights. The model proved to be remarkably robust, with the deep disposal options consistently coming out as more preferred than the temporary storage options. Subsequently, CoRWM formulated 15 recommendations, which were passed on to the government in July 2006. In November 2006 the government accepted the key elements of CoRWM’s recommendations, and work is now ongoing to select a site. Hardly any criticism of this work has appeared in the media, which is a tribute to the openness with which CoRWM conducted their work. The following sections describe this case study in more detail.

The Initial Work

CoRWM members were chosen by the Department of the Environment, Food, and Rural Affairs (DEFRA) from over 200 applications. The committee represented great diversity of perspectives on the

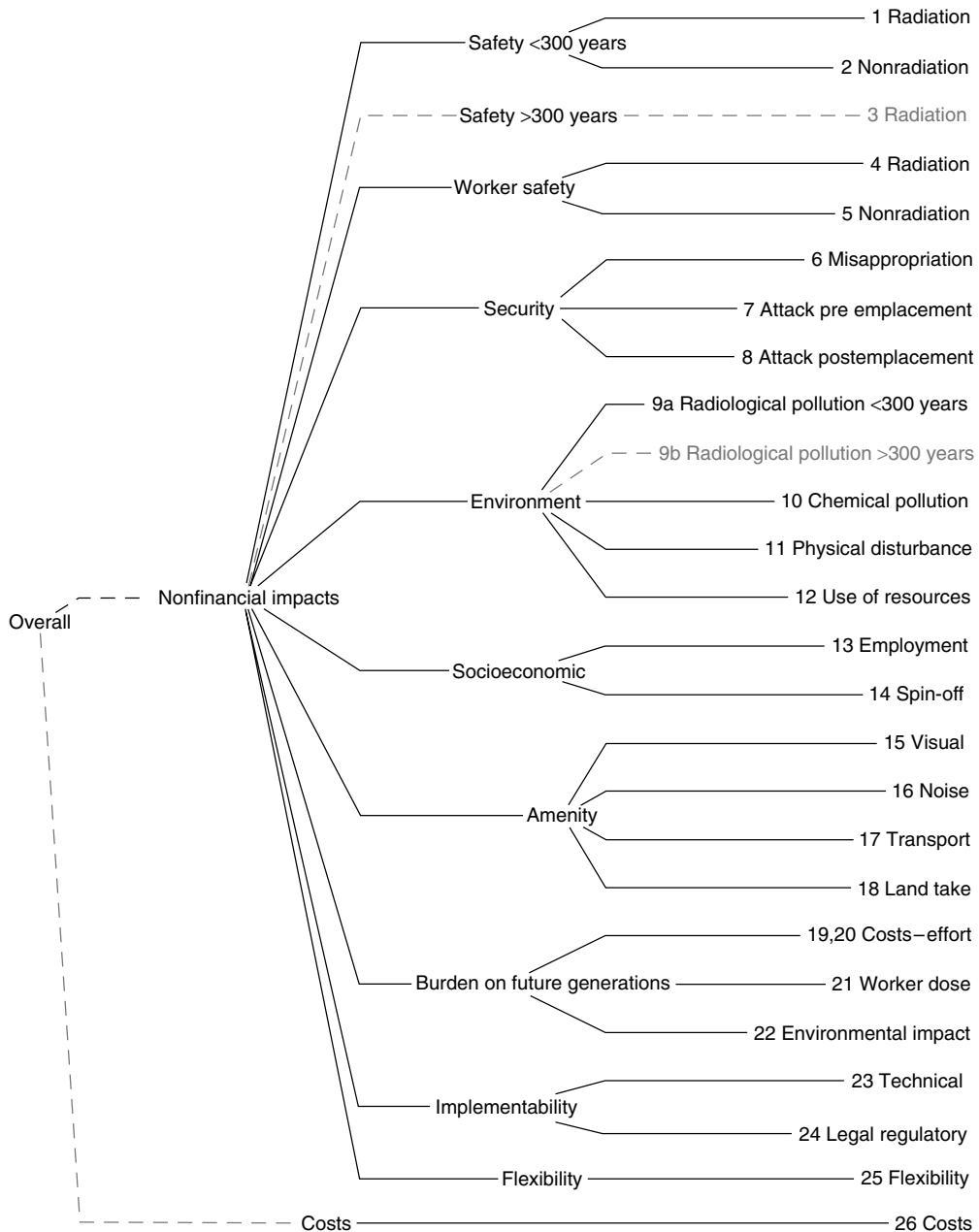


Figure 2 Value tree for the radioactive waste case

issue of radioactive waste disposal, from enthusiastic nuclear supporter to Greenpeace skeptic. The committee first established a methodology, analytic-deliberative discourse [13], which relies on both holistic assessment and analytic modeling of values.

Then they set to work engaging the public in eliciting the views of citizens and stakeholders from different cross sections of society about the options for managing the wastes. Eventually, 14 viable options were grouped in three categories: interim storage,

geological disposal, and nongeological disposal, the latter two assumed to be permanent disposal categories.

CoRWM then identified 10 objectives, which they called *headline criteria*, which should be met as far as possible by an option. Criteria under those headline criteria provided practical realizations of the objectives, enabling the options to be scored meaningfully. Figure 2 shows the completed MCDA value tree. Note that the word “risk” does not appear; the many impact subcriteria effectively cover all aspects of the risk of managing radioactive waste, but using terms that can be defined with sufficient clarity so that the options can be scored on the subcriteria.

The Workshops

The committee selected specialists who attended several facilitated workshops in 2005 to clarify definitions of the criteria and to establish equal-interval assessment scales. An example of a scale is shown in Table 1.

The nine-point scale was constructed by first asking the experts to define a minimally acceptable definition of “1” and a maximally feasible definition of “9” such that all options would fall within the range from 1 to 9. The next step defined point “5” such that the value difference between 1 and

5 was the same as between 5 and 9. For this and many other scales, further definitions for points 3 and 7 were defined so as to provide equal value increments as well. Thus, this process effectively defined value functions by incorporating them into the measurement scale, effectively a Likert scale [14].

The expert workshops resulted in many changes to the definitions of the criteria, including the merging of two criteria that were deemed as not mutually preference independent (Costs and Effort, under the headline criterion “Burden on future generations”). Much discussion and debate attended the formulation of nearly every subcriterion, as the different experiences of the experts were revealed and their conflicting perspectives aired, eventually leading to agreement about how the subcriterion should be defined.

In later workshops, the same experts scored the options on the subcriteria. The process of committing to a number between 1 and 9 led to further discoveries of differences in viewpoint, and the ensuing discussion attempted to resolve those differences to arrive at a consensus score. When consensus could not be reached, the median value was input to the computer model, with the range noted and later subjected to sensitivity analysis.

In the meantime, CoRWM engaged the public in scoring and also in swing weighting the scales (*see Multiattribute Modeling*). The latter was a process

Table 1 The Flexibility subcriterion for the radioactive waste case, showing the nine-point scale for scoring the options

Subcriterion 28: Flexibility	
Extent to which the option is expected to allow for future choice and respond to unforeseen or changed circumstances over the 300-year time span	
9	System is fully monitored and adaptable, and the waste is easily retrievable using the existing system
8	
7	Key system elements can be monitored and are moderately adaptable, and the waste retrieval requires some modification of the system
6	
5	Some system elements can be monitored, adaptability is limited, and waste retrieval is moderately difficult
4	
3	Few system elements are monitored, little adaptability in the system, and it is difficult to retrieve the waste
2	
1	Monitoring options is severely restricted, the system is not adaptable, and waste retrieval is very difficult

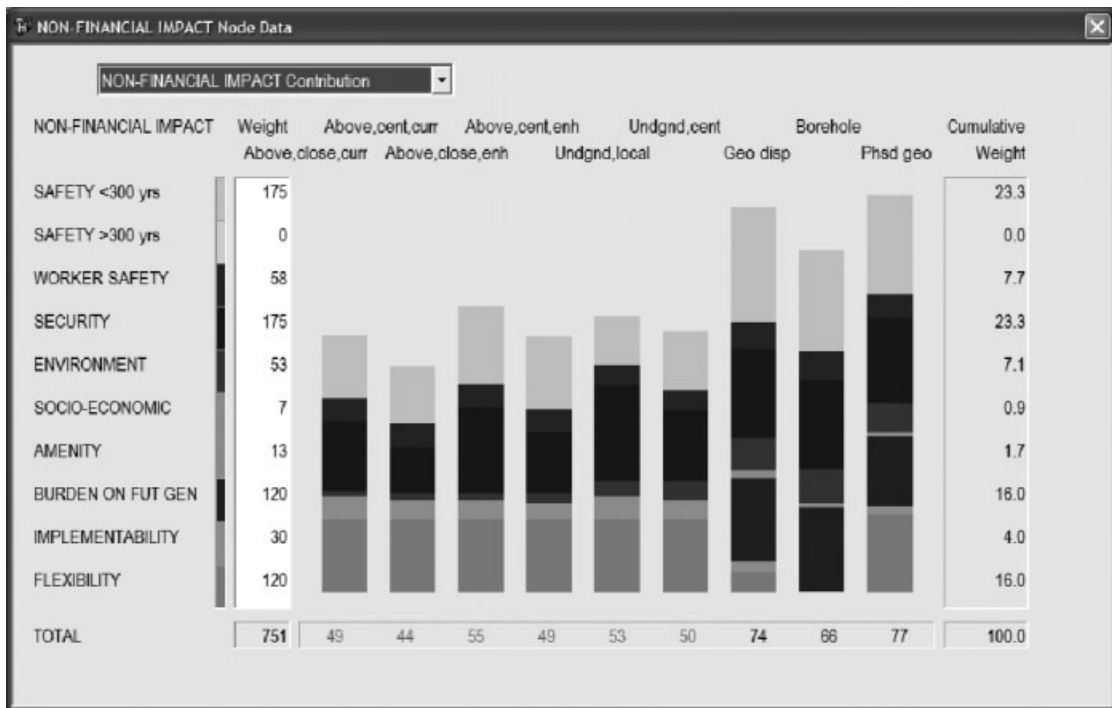


Figure 3 Final result for the impacts criterion in the radioactive waste case. The six left bars are for storage options and the three right bars are for disposal options

that compared the swings in value from 1 to 9 on one scale as compared to another. One CoRWM member summarized the views of citizens and stakeholders about scoring and weighting, and these views were subsequently used in sensitivity analyses.

The Final Decision Conference

At the end of March 2006, a final decision conference with CoRWM members provided the opportunity to complete the MCDA model by weighting the criteria. The cabaret-style seating placed the members at two tables, with weights discussed separately at the tables. Each group reported its views and a whole-group discussion followed, usually to consensus. When consensus could not be reached, recording the extreme views left them to be tested in sensitivity analyses.

Figure 3 shows the final result for the high-level wastes, the most dangerous waste stream, for the nine options relevant to that waste stream. Only the impacts criteria are shown, with the stacked bar graph

displaying the contribution of each headline criterion, and the numbers at the right giving the relative weights associated with each headline criterion. It is clear that the three geological disposal options score higher, overall, than the six storage options. Subsequent sensitivity analysis rarely changed this picture. Substituting optimistic scores for the storage options and pessimistic scores for the disposal options on all criteria for which the experts could not agree about the scores very slightly reduced the difference between the storage and disposal options, but the relative heights of the bar graphs remained unchanged. Testing the robustness of this result against differences of opinion among the CoRWM members' judgments about weights did not change the picture. Nor did it shift when scores and weights representing different views of citizens and stakeholders were tested in the model. Overall, a remarkably robust result emerged.

A week later, the CoRWM committee met again to engage in a holistic analysis, which largely confirmed the results of the MCDA. Taking these appraisals

into account, along with deep consideration of ethical issues, the committee then set about drafting its recommendations, which were made available to the public for comment. By July 2006 a final set of 15 recommendations was forwarded to the government. These revolved around three main themes: geological disposal as the end state, improved interim storage, and a new approach to implementation that engages local communities to participate willingly. In November the Secretary of State, David Miliband, accepted the suggested policies. The Implementation Planning Group, established by DEFRA, is now taking the lead on the next steps. CoRWM's work was conducted in full public view, with all reports archived at that time on <http://www.corwm.org.uk>, now a valuable resource for all aspects of this project. A major conclusion from this project is that analytic-deliberative discourse methods, like those embodied in decision conferencing, can be made to work on important risk assessment projects.

References

- [1] Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many are Smarter Than the Few*, Little Brown, London.
- [2] Phillips, L.D. (1999). Group elicitation of probability distributions: are many heads better than one? in *Decision Science and Technology: Reflections on the Contributions of Ward Edwards*, J. Shanteau, B. Mellors & D. Schum, eds, Kluwer Academic Publishers, Norwell, pp. 313–330.
- [3] Phillips, L.D. (2007). Decision conferencing, in *Advances in Decision Analysis: From Foundations to Applications*, W. Edwards, R.F. Miles & D. von Winterfeldt, eds, Cambridge University Press, Cambridge.
- [4] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley & Sons, New York.
- [5] Phillips, L.D. & Phillips, M.C. (1993). Facilitated work groups: theory and practice, *Journal of the Operational Research Society* **44**, 533–549.
- [6] Schein, E.H. (1999). *Process Consultation Revisited: Building the Helping Relationship*, Addison-Wesley, Reading.
- [7] Bion, W.R. (1961). *Experiences in Groups*, Tavistock Publications, London.
- [8] Lawrence, G., Bain, A. & Gould, L. (1996). The fifth basic assumption, *Free Associations* **6**(37), 28–55.
- [9] Trist, E. & Murray, H. (eds) (1990). *The Social Engagement of Social Science: A Tavistock Anthology*, Free Association Books, London.
- [10] Hickling, A. (1990). Decision spaces: a scenario about designing appropriate rooms for 'activity-based' decision management, in *Tackling Strategic Problems: The Role of Group Decision Support*, C. Eden & J. Radford, eds, Sage, London, pp. 167–177.
- [11] Chun, K.-J. (1992). *Analysis of Decision Conferencing: A UK/USA Comparison*, London School of Economics & Political Science, London.
- [12] McCart, A.T. & Rohrbaugh, J. (1989). Evaluating group decision support system effectiveness: a performance study of decision conferencing, *Decision Support Systems* **5**, 243–253.
- [13] Renn, O. (1999). A model for an analytic-deliberative process in risk management, *Environmental Science and Technology* **33**, 3049–3055.
- [14] Likert, R. (1932). A technique for the measurement of attitudes, *Archives of Psychology* **140**, 44–53.

Related Articles

Risk and the Media

Scientific Uncertainty in Social Debates Around Risk

Stakeholder Participation in Risk Management Decision Making

LAWRENCE D. PHILLIPS

Decision Modeling

In its simplest form, Bayesian decision modeling can be formulated as follows: a decision maker (DM) needs to choose a single action $a \in A$ where A is the set of acts – called the *action space* – she has available. The consequence $c(a, \theta)$ of this choice a will depend on the state of nature $\theta \in \Theta$ where Θ – called the *state space* – is the set of all such states the DM believes might happen. When both $A = \{a_1, a_2, \dots, a_m\}$ and $\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}$ are finite sets, the consequences $\{c(a_i, \theta_j) = c_{ij} : i = 1, 2, \dots, m, j = 1, 2, \dots, n\}$ can be expressed as a table called a *decision table* as shown in Table 1.

Such tables need not be algebraic. Ideally, the acts, the states of nature, and the consequences are all given in common language and in simple problems can be elicited directly from the DM's verbal description of her problem.

Example 1 There is a suggestion that an infection has entered a herd of cattle. If it has, then there is a small possibility that the infection may be detected entering the food chain. The authorities could choose to continue to observe the herd, but otherwise do nothing, or order the immediate culling of the herd. This problem could be represented by the decision table (Table 2).

Here, the consequence c_{11} , to the authorities, is the cost of the observation. Consequence c_{12} is the cost of the observation but also, given that the infection

enters the food chain, the cost incurred by the mass withdrawal of foodstuff from the market and the costs to farmers associated with the fear of beef products and the consequent loss of jobs in the industry. Consequences c_{21} and c_{22} are the same and consist of the cost of the cull and the costs to all farmers associated with a loss of confidence in beef products, albeit after any health risk has been demonstrably swiftly addressed.

The example above illustrates several points. First, while decision modeling, it is necessary to be clear *whose* consequences we are considering. We need to consider the consequences for those in charge (owners or their agents) of the decisive act. In the problem above, the authorities are the owners of the act – (not the farmer of the possibly infected cattle or beef farmers, in general). Second, a decision has to be made as to the precise lists of legitimate attributes of the consequences. Note, in our example, we have omitted any explicit evaluation of health risk. If health risks might endure after a mass withdrawal of foodstuffs, such an evaluation would normally need to be included in any decision analysis. Third, as far as possible the state space should contain events that are eventually resolved. Thus, whether the infection is detected as entering the food chain is resolvable, although, whether it actually entered may not be. Considerable effort often needs to be expended to define states in sufficient detail and clarity so that they are resolvable (see [1, 2]). In particular, states need to be defined in sufficient detail so that the consequences arising from the state and each possible act are clear (see [2] for further discussion of this issue).

Fourth, the separation of acts and states of nature is useful in the sense that it often helps the DM to discover acts she had not previously considered. In the example above, these might include a partial cull of the herd. This augmentation of the problem often necessitates new states being added. The description thus dynamically evolves until the DM believes that it represents her problem satisfactorily, or is *requisite* (see [3]). Using semantics that allow the DM to develop her description so that it faithfully represents what she believes is the structure of her problem, without getting into early quantification, is an essential aspect of a decision analysis. The decision table is the simplest such framework. More expressive alternative frameworks are the decision

Table 1 A decision table

	States of nature						
	θ_1	θ_2	$\dots$	θ_j	$\dots$	θ_n	
Actions	a_1	c_{11}	c_{12}	$\dots$	c_{1j}	$\dots$	c_{1n}
	a_2	c_{21}	c_{22}		c_{2j}		c_{2n}
	$\vdots$			$\ddots$			$\vdots$
	a_i	c_{i1}	c_{i2}		c_{ij}		c_{in}
	$\vdots$					$\ddots$	$\vdots$
	a_m	c_{m1}	c_{m2}	$\dots$	c_{mj}	$\dots$	c_{mn}

Table 2 Decision table for Example 1

	No transfer of infection	Entered food chain
Observe	c_{11}	c_{12}
Cull herd	c_{21}	c_{22}

tree (see **Decision Trees**) and the influence diagram (see **Influence Diagrams**), see below.

Probabilistic Uncertainty and Optimal Decisions

The value of the state of nature θ is usually at best only partially known to the DM when she needs to act. However, being a Bayesian she is willing to articulate her beliefs about the states of nature probabilistically: by a density (or if Θ is discrete, a probability mass function) $p(\theta)$ called her *prior density* (mass function). For each consequence $c \in C$, she will assign a score $u(c)$ – called her *utility score* for c . The utility score will assign a numerical value that reflects how good, in her own eyes, a consequence is. Thus, in particular, if she thinks that consequence c_2 is at least as preferable as a consequence c_1 then her utility score $u(c_2) \geq u(c_1)$. It is common to assign a value zero to the utility of the worst consequence and a value one to the most desirable consequence, although this is not formally necessary. How the real-valued utility function is formally defined and elicited is discussed elsewhere (see **Utility Function**).

Because the state of nature is typically not fully known when the act needs to be chosen, it is often not immediately clear which act is the best choice. An act that is good in one state of nature may not be good in another. Bayesian decision theory states that for a DM to be rational (in a sense defined in e.g. [1, 4–6]), she should choose an act $a^* \in A$, her space of possible actions, so that a^* – often called her *Bayes decision* – maximizes her expected utility $\bar{u}(a)$ on the set A , i.e.,

$$a^* = \arg \max \bar{u}(a) \quad (1)$$

When θ is discrete and takes only a countable set of values, $\bar{u}(a)$ can be written as

$$\bar{u}(a) = \sum_{\theta \in \Theta} u(c(a, \theta)) p(\theta) \quad (2)$$

while when θ has an absolutely continuous density $p(\theta)$

$$\bar{u}(a) = \int_{\theta \in \Theta} u(c(a, \theta)) p(\theta) d\theta \quad (3)$$

When the axioms that logically lead the DM to a Bayes decision are violated, it can be demonstrated that there will always be scenarios where the DM's preferences over acts are inconsistent (see e.g., [7]). So, there are strong arguments for encouraging the

DM to choose an act to maximize her expected utility. However, it is also true that DMs often exhibit preferences that are not expected utility maximizing. Indeed, there is even evidence to suggest that without coaxing, it is common for DMs not to think in terms of actions, states of nature, and their consequences at all (see, for example [8])! So this Bayesian decision methodology is not primarily descriptive, but prescriptive. The main use of Bayesian decision analysis is to encourage a DM to make wise choices and choose explicable acts in the face of uncertainty (see e.g., [7]).

Using Data to Guide Decisions

On many occasions, it is possible to arrange to take measurements $\mathbf{x} \in \mathbb{X}$ of a random vector $\mathbf{X}$; often at a cost in money or resources. For example, in the risk scenario described above, it might be possible to sample the herd and check this sample for any signs of the infection. Our choices are then whether to exploit this opportunity to sample, if so, which measurements to take, and given that we take measurements, to decide how best to act in the light of them. In particular, we need to help the DM choose a *decision rule* $d(\mathbf{x}) \in \mathbb{D}$ that specifies the act $a \in A$ she takes as a function of the measurements $\mathbf{x} \in \mathbb{X}$. Again Bayesian decision theory prescribes that the DM should choose a *Bayes decision rule* $d^*(\mathbf{X})$ where

$$d^*(\mathbf{X}) = \arg \max \bar{u}(d(\mathbf{X})) \quad (4)$$

and where, when $(\mathbf{x}, \theta)$ are discrete,

$$\bar{u}(d(\mathbf{X})) = \sum_{\mathbf{x} \in \mathbb{X}} \left\{ \sum_{\theta \in \Theta} u(c(d(\mathbf{x}), \theta)) p(\theta | \mathbf{x}) \right\} p(\mathbf{x}) \quad (5)$$

while when $(\mathbf{x}, \theta)$ has an absolutely continuous density $p(\mathbf{x}, \theta)$,

$$\bar{u}(d(\mathbf{X})) = \int_{\mathbf{x} \in \mathbb{X}} \left\{ \int_{\theta \in \Theta} u(c(d(\mathbf{x}), \theta)) p(\theta | \mathbf{x}) d\theta \right\} \times p(\mathbf{x}) d\mathbf{x} \quad (6)$$

Here $p(\theta | \mathbf{x})$ – the posterior mass function or density – is usually calculated using Bayes Rule

$$p(\theta | \mathbf{x}) = \frac{p(\mathbf{x} | \theta) p(\theta)}{p(\mathbf{x})} \quad (7)$$

where $p(\theta)$ is the DM's prior mass function or density, capturing her beliefs before seeing the resulting

measurement $\mathbf{x}$, $p(\mathbf{x}|\theta)$ specifies her probability of learning that $X = \mathbf{x}$ when the state of nature is θ , and

$$p(\mathbf{x}) = \sum_{\theta \in \Theta} p(\mathbf{x}|\theta)p(\theta) \quad (8)$$

when $(\mathbf{x}, \theta)$ is discrete and

$$p(\mathbf{x}) = \int_{\theta \in \Theta} p(\mathbf{x}|\theta)p(\theta) d\theta \quad (9)$$

when $(\mathbf{x}, \theta)$ is absolutely continuous. It is simple to check that in regular cases, for example, when the spaces $\mathbb{X}$ and Θ are both finite and $p(\mathbf{x}, \theta) > 0$ for all $\mathbf{x} \in \mathbb{X}$ and $\theta \in \Theta$ that for each $\mathbf{x} \in \mathbb{X}$, the DM should choose d so that $d(\mathbf{x})$ maximizes.

$$\bar{u}(d(\mathbf{x})) = \sum_{\theta \in \Theta} u(c(d(\mathbf{x}), \theta))p(\theta|\mathbf{x}) \quad (10)$$

When for any $\mathbf{x} \in \mathbb{X}$, θ is absolutely continuous with a density $p(\theta|\mathbf{x})$ then $\bar{u}(d(\mathbf{x}))$ is given by

$$\bar{u}(d(\mathbf{x})) = \int_{\theta \in \Theta} u(c(d(\mathbf{x}), \theta))p(\theta|\mathbf{x}) d\theta \quad (11)$$

and the decision rule $d(\mathbf{x})$ which maximizes this for each $\mathbf{x} \in \mathbb{X}$ is also a Bayes decision rule. Thus (in such regular cases), the DM should plan to choose an act that maximizes her posterior expected utility for each possible value $\mathbf{x} \in \mathbb{X}$ she might observe.

Issues and Developments

This formulation of Bayesian decision theory follows most closely the approach taken by Savage [4] who separated out the action space and state space as described above. This works very well for simple decision problems, but the semantics of the decision table are restrictive. Not all decisions may be possible or sensible, given certain states of nature and states are often impossible, given a certain action is taken. For example, in the setting described above, inoculating the herd after it is culled would be absurd. Furthermore, learning that the herd are all dead makes culling it impossible. This can make the decision table at best ambiguous and difficult to explain. Furthermore, decisions are often taken sequentially, where each new act provides new information about the states of nature. Thus, in the example above, we might choose a decision rule that first uses a crude test for the presence of the infection and on the basis of the results of this test, a more precise but costly test may be used. In these sorts of scenario,

the decision table is again not the best framework for representing the broad structure of such a decision problem. Finally, if we have a continuous state space, then the decision table can no longer be used to summarize the structure of the problem, while other frameworks, such as the influence diagram can. For all these reasons, various alternative representations of decision problems have been developed; two of the more popular and established representations are the decision tree (see **Decision Trees**) and the influence diagram (see **Influence Diagrams**).

References

- [1] De Finetti, B. (1974). *Theory of Probability*, John Wiley & Sons, Vol. 1.
- [2] Howard, R.A. (1988). Decision analysis: practice and promise, *Management Science* **34**(6), 679–695.
- [3] Phillips, L.D. (1984). A theory of requisite decision models, *Acta Psychologica* **56**, 29–48.
- [4] Savage, L.J. (1972). *The Foundations of Statistics*, 2nd Edition, Dover Publications.
- [5] Smith, J.Q. (1988). *Decision Analysis: A Bayesian Approach*, Chapman & Hall.
- [6] Robert, C. (1994). *The Bayesian Case*, Springer–Verlag, Berlin.
- [7] French, S. & Rios Insua, D. (2000). *Statistical Decision Theory, Kendall's Library of Statistics* 9, Arnold.
- [8] Beach, L.R. (1990). *Image Theory: Decision Making in Personal Organizational Contexts*, John Wiley & Sons.

Further Reading

- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufman Publishers, San Mateo.
- Poole, D. & Zhang, N.L. (2003). Exploiting contextual independence in probabilistic inference, *Journal of Artificial Intelligence Research* **18**, 263–313.
- Raiffa, H. (1968). *Decision Analysis*, Addison–Wesley.

Related Articles

Bayes' Theorem and Updating of Belief

Bayesian Statistics in Quantitative Risk Assessment

Decision Analysis

Interpretations of Probability

JIM Q. SMITH AND PETER THWAITES

Decision Trees

Although it is one of the oldest representations of a decision problem, the decision tree is a very versatile and general framework within which one can calculate and provide an explanation for a Bayes decision rule in a given decision problem (see **Decision Modeling**). It is possible to modify the representation so that it represents decision problems with continuous variables, but it is most expressive and useful when both the state space and the decision space are finite: the only case we consider here. Unlike many of its competitors – like the influence diagram (see **Influence Diagrams**) – it is flexible and expressive enough to represent asymmetries within both the decision space and outcome space explicitly through its topology. Of course, this flexibility can also be a drawback. Because the complete set of possible unfoldings of situations is expressed explicitly, the decision tree can be topologically complicated even for relatively simple problems. Nevertheless, the framework it provides is still extremely valuable.

The decision tree represents a problem sequentially, encoding step by step which options are open to the decision maker (DM) at a given time, and how the situation the DM might find himself or herself in could develop. So decision trees are particularly powerful in scenarios where each decision rule $d(\mathbf{x})$ is naturally specified as an ordered *decision sequence* $d(\mathbf{x}) = (d_1, d_2, \dots, d_k)$ where new information, pertinent to the states of nature, is gathered between the component decisions $d_i, i = 1, 2, \dots, k$: see the example below.

Once the tree coding how information is acquired through different decisions is elicited, the topology of the tree can be supplemented with probabilities and utilities, and this information utilized to calculate a Bayes decision rule. So the decision tree gives not only an evocative and explicit representation of important *qualitative* features of a decision problem, like the possible unfoldings, the relative timing of decisions, and the possible acquisition of new information in the light of each component decision, but can also be used as a formal framework for calculating the best course of action in the light of the *quantitative* information provided by the DM. We illustrate this process below.

It should be noted that a decision problem has many decision tree representations, with different

trees having different purposes. There are three types of trees: *normal form tree*, the *causal tree*, and the *dynamic programming tree*. In this chapter, we focus mainly on the third and the most common type of decision tree – the dynamic programming tree. Such a tree is designed so that, besides providing a useful description of a decision problem, it can also be elaborated with probabilities and utilities and then used as a framework for a particularly simple and transparent method of calculating a Bayes decision rule called *backwards induction* or *rollback*. Such trees enable the DM to identify explicitly a Bayes decision rule even for very inhomogeneous problems. Software is now widely available supporting both the construction of the tree and the rollback calculations illustrated below that enable optimal decision rules to be identified. Because of their expressiveness and versatility, decision trees continue to play a central role in the analysis of discrete decision problems.

Drawing a Decision Tree

The following example is a simplification of the type of decision problem faced by users of a forensic science laboratory (see [1]).

Example 1 The police believe with probability one that a suspect S is guilty. To help convict the suspect, they would like the forensic science service to examine articles of the suspect's clothing for the existence of fibre that matches fibre at the crime scene. The culprit was seen by several witnesses to be wearing a skirt and blouse identical to the one owned by the suspect. If the suspect had sat on a sofa at the crime scene then the probability she picked up a fibre from the sofa that will be found thorough an analysis of the skirt or the blouse are respectively 0.8 and 0.5 where these events are independent. If the suspect has not sat on the sofa the police believe that the probability of finding a matching fibre is zero. On the basis of witness statements the police believe the probability the culprit sat on the sofa is 0.8. The probability they obtain a conviction given they find this match is 0.9 whilst if no such evidence is found then their probability of conviction is 0.1. The police can arrange the examination of the suspect's skirt or blouse or both (in either order). The cost for immediately analyzing both is \$1600. The police could also analyze the skirt alone for \$1200, the blouse for \$800 with these rates also

applying if on analyzing one item of clothing they subsequently decide to analyze the other. Additional costs to the police of going to court are \$2000 but \$0 if they release the suspect without charge. The police's utility $u(x_1, x_2)$ has attributes of cost x_1 and probability of conviction x_2 and has been elicited and found to take the form

$$u(x_1, x_2) = 0.2u_1(x_1) + 0.8x_2 \quad (1)$$

where

$$u_1(x_1) = (4000)^{-1} (4000 - x_1) \quad (2)$$

Such a problem can be usefully represented and the optimal decision discovered using a dynamic programming decision tree. The nodes of the tree, other than those at the tips of its branches, are called *situations* and each situation is either a decision node and represented by $\square$ or a chance node, represented by $\circ$. At a situation, a collection of circumstances can happen. Each edge emanating from a situation is labeled by such a circumstance. If the situation is a decision node, then the edges emanating out of this vertex label the different possible decisions that can be made at that point in time. If the situation is a chance node, then the edges coming out of this vertex label the different possible outcomes that could take place: either subsets of states or observations. We draw the tree with its root (the first situation faced by the DM) on the far left of the page. We then sequentially introduce new situations in an order, consistent with when they are observed or enacted. Conversely, if the value represented by an edge emanating from a chance node $v(c)$ is not known when d_2 is committed to, then $v(c)$ must be introduced after $v(d_2)$ in the tree: i.e., further from the root node.

To illustrate this construction, consider the problem above. The first situation facing the police is a decision. Do they release the suspect immediately or pursue the case in some way? The edge associated with not taking the suspect to court is labeled "release", while the edge associated with not doing this is labeled "pursue". If the police choose not to release the suspect, then, with no new information, they are faced with the choice of how to use the forensic science service (if at all) in the problem above. Here the emanating edges and receiving vertices can

be labeled

$$e_0 = \{\text{go to court now}\}$$

$$e_s = \{\text{pay for the analysis of the skirt}\}$$

$$e_b = \{\text{pay for the analysis of the blouse}\}$$

$$e_2 = \{\text{pay for the analysis of both together}\}$$

So the first part of the tree is given in Figure 1.

Note that we could equally well start with the alternative tree given by Figure 2.

However, the algorithm for calculating a Bayes decision described below always gives the same answer irrespective of the particular choice of the dynamic programming tree [2]. On taking the decision e not to investigate the potential forensic evidence, the situation also labeled by e is whether

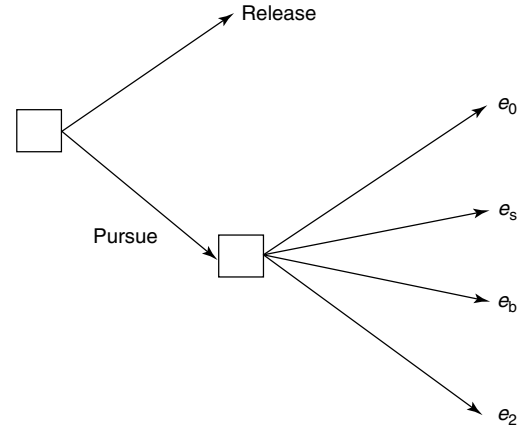


Figure 1 Beginning to draw a decision tree

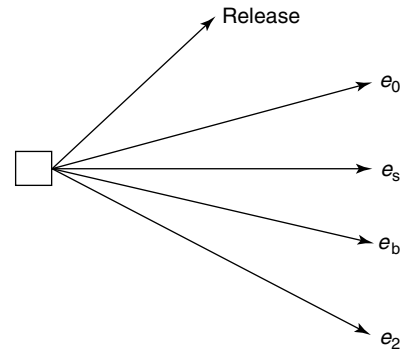


Figure 2 An alternative start to this decision tree

the court subsequently finds the suspect guilty g or innocent i . If they choose to pay for the investigation e_s of the skirt, then the result of this investigation could be positive (+) with a matching fiber being found, or negative (-). So draw two edges out of node e_s to vertices labeled + and -. If a fiber is found on the skirt – i.e., if the police reach situation $(e_s, +)$ then they have their evidence and there is no point in paying for the blouse to be investigated. They therefore need to decide whether to take the suspect to court e or release her $\bar{e}$. On arriving at $(e_s, +, e)$, the court finds the suspect guilty or innocent with the probabilities provided above.

However, if the police learn that there is no fiber on the skirt, i.e., arrive at the situation $(e_s, -)$, then they can decide to release the suspect, take the suspect to court immediately $(e_s, -, e)$, or pay for the investigation of the blouse $e_{s,b}$. On reaching the situation $(e_s, -, e_{s,b})$, a matching fiber is found (+) on the blouse or not found (-), leading by an edge to a situation $(e_s, -, e_{s,b}, +)$ or $(e_s, -, e_{s,b}, -)$. This, in turn, leads to the decision to release ($\bar{e}$), or go to court, i.e., leads to one of the two situations $(e_s, -, e_{s,b}, +, \bar{e})$, $(e_s, -, e_{s,b}, +, e)$ or $(e_s, -, e_{s,b}, -, \bar{e})$, $(e_s, -, e_{s,b}, -, e)$.

Alternatively, the police could choose to investigate the blouse first and then decide on whether to subsequently check the skirt. Proceeding exactly as for the branch starting with e_s but permuting the indices b and s gives us the unfolding of situations after e_b . Finally, on choosing (e_2) to investigate both garments immediately, and depending on whether a fiber is found (+) or not (-) the decision of whether to go to court has to be made. The full tree for this problem is given in Figure 3.

The *leaf nodes* or *terminal nodes* are the ones a maximum distance from the root node: technically the non-root nodes connected to only one other node. For transparency only selected nodes are labelled in the tree above.

Some Comments on Drawing a Tree

1. Even in this small problem, we note that the tree is very bushy. However, since such trees are now mostly drawn electronically, this is not such a problem, because we are then able to zoom in and out of parts of the tree as we draw or examine it.
2. Like the decision table, the tree provides a useful qualitative framework that can be elicited
3. Each terminal vertex (or equivalently each root-to-leaf path) labels a possible unfolding of circumstances as seen by the DM: here, the police. Thus, all possible consequences are represented on the tree, making the description of the problem appreciable to a client. This transparency means that, in practice, the tree is often modified while it is being elicited as the client remembers various possible turns of events not originally articulated. For instance, in the example above, the client might tell you that there is no time to wait for the results of one fiber test and then perform another: so any act that involves $e_{b,s}$ or $e_{s,b}$ is infeasible. This new information is simple to accommodate: we simply erase paths containing these edges. Alternatively, the police may have to follow a protocol that always examined a skirt if a blouse was analyzed. Again this protocol is easy to obey: we simply remove all paths containing these edges. Conversely, as the tree is drawn, other possibilities may come to mind – here, for example, the possibility of forensic analysis of a different piece of evidence. This type of elaboration can be simply incorporated into the tree. The tree gives natural support to the evolutionary dynamics of the modeling process, allowing the client to continue to develop a model until it is requisite [3].
4. The tree can be made simpler by not representing in it those decision rules that are clearly

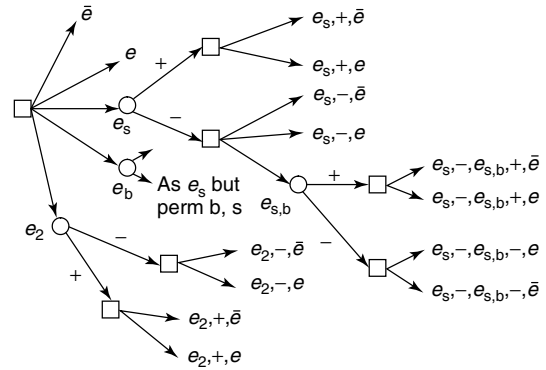


Figure 3 The full tree for Example 1

always going to be suboptimal. Here, for example, the decision $(e_s, +, \bar{e})$ of paying to investigate the skirt and then, when a fiber is found, to release the suspect clearly wastes the cost of the investigation: just releasing the suspect $\bar{e}$, must obviously be a better option than this. However, some care needs to be exercised when pruning the tree in this way, since it is quite easy to discard decision rules that might actually be good. This is particularly true if you plan to augment your problem to perform a sensitivity analysis. When in doubt, we suggest you keep the rules in and let them be deleted by rollback algorithms described below. However, sometimes, problems are so complicated, it is necessary to simplify the tree. Various formal techniques, using measures of expected value of information, or a normal form analysis can be used to simplify the tree so that only promising decision rules are compared [2, 4]. Alternatively, we can select only a small subset of the possible decision rules that appear reasonably dense in the space of possible decisions and hope that one of these is close to being optimal.

5. Like all Bayesian decision models, the efficacy of the representation is limited by the imagination of the people providing the description of the problem. However, because of the simplicity and the familiarity of the representation of possible events as a path through time, it often oils the imagination, so that the client's understanding of various intricacies of the process increases or becomes more secure as the tree is drawn.
6. Because a dynamic programming tree is drawn primarily to identify promising decisions, not all variables in the problem are necessarily represented in the tree, but only those that influence the consequences in the problem. In the illustrative example, for instance, whether the culprit sat on the sofa is not represented here. Note, however, that it does appear implicitly and is used to evaluate probabilities on the unfolding root-to-leaf paths.

Embellishing the Tree with Quantitative Information

Once the tree is drawn, we can embellish it by appending *terminal utilities* to the leaves of the root-to-leaf paths and probabilities to the edges emanating

from its chance nodes. Thus, for example, consider the root-to-leaf path ending with the leaf $(e_s, +, e)$ and labeling the unfolding that the police chose to check the skirt which was found to have on it a matching fiber that they then go to court. The financial cost x_1 in this case is the cost of the examination plus the other costs of going to court (\$1200 + \$2000) and the probability of a successful conviction is 0.9 so the utility of this consequence is

$$U(x_1, x_2) = 0.2 (4000)^{-1} (4000 - 3200) + 0.8 \times 0.9 = 0.76 \quad (3)$$

The terminal utilities associated with this problem are given in Table 1.

We can add the probabilities that a chance situation develops into an adjacent situation and append these to the edges emanating from each chance node in the tree. Because the elicited probabilities in the problem usually condition in an order consistent with how the client believes things happen rather than condition on the order things are observed, these are often not elicited directly, but often need to be calculated, using the law of total probability

Table 1 Terminal utilities

Leaf	Cost x_1	Conviction probability x_2	Utility
$\bar{e}$	0	0	0.2
e	2000	0.1	0.18
$e_s, +, \bar{e}$	1200	0	0.14
$e_s, +, e$	3200	0.9	0.76
$e_s, -, \bar{e}$	1200	0	0.14
$e_s, -, e$	3200	0.1	0.12
$e_s, -, e_{s,b}, +, \bar{e}$	2000	0	0.1
$e_s, -, e_{s,b}, +, e$	4000	0.9	0.72
$e_s, -, e_{s,b}, -, e$	4000	0.1	0.08
$e_s, -, e_{s,b}, -, \bar{e}$	2000	0	0.1
$e_b, +, \bar{e}$	800	0	0.16
$e_b, +, e$	2800	0.9	0.78
$e_b, -, \bar{e}$	800	0	0.16
$e_b, -, e$	2800	0.1	0.14
$e_b, -, e_{b,s}, +, \bar{e}$	2000	0	0.1
$e_b, -, e_{b,s}, +, e$	4000	0.9	0.72
$e_b, -, e_{b,s}, -, e$	4000	0.1	0.08
$e_b, -, e_{b,s}, -, \bar{e}$	2000	0	0.1
$e_2, -, \bar{e}$	1600	0	0.12
$e_2, -, e$	3600	0.1	0.1
$e_2, -, \bar{e}$	1600	0	0.12
$e_2, +, e$	3600	0.9	0.74

and/or Bayes rule (see **Bayes' Theorem and Updating of Belief**). This is called *preprocessing* in [5]. In the example above, we can calculate the probabilities associated with the chance nodes labeled $(e_s, e_b, e_2, (e_s, -, e_{s,b}), (e_b, -, e_{b,s}))$.

Thus, if E ($\bar{E}$) denotes the event that the culprit sat (did not sit) on the sofa then, by the law of total probability

$$\begin{aligned} P(+|e_s) &= P(+|e_s, E)P(E) + P(+|e_s, \bar{E})P(\bar{E}) \\ &= 0.8 \times 0.8 + 0 \times 0.2 = 0.64 \\ &\Leftrightarrow P(-|e_s) = 0.36 \end{aligned} \quad (4)$$

Notice that the probability on the complementary edge $P(-|e_s) = 1 - P(+|e_s) = 0.36$, so we can quickly derive the probability of no fiber match from the probability of a positive identification. Similarly, if $N(S)$ and $N(B)$ denote respectively, the events that no matching fiber is found on the skirt (blouse) then

$$\begin{aligned} P(+|e_2) &= P(+|e_2, E)P(E) \\ &= \{1 - P(N(S) \cap N(B)|e_2, E)\} P(E) \\ &= \{1 - P(N(S)|e_2, E)P(N(B)|e_2, E)\} P(E) \\ &= \{1 - 0.2 \times 0.5\} \times 0.8 = 0.72 \end{aligned} \quad (5)$$

by independence. More work is needed to calculate the edge probabilities on the longer root-to-leaf paths. For example, consider the probability $P(-|e_s, -, e_{s,b}, -, \bar{e})$. Then, by our assumption of independence, and from the probabilities given above

$$\begin{aligned} P(N(B)|N(S)) &= \frac{P(N(S) \cap N(B))}{P(N(S))} \\ &= \frac{P(N(S) \cap N(B)|E)P(E) + P(N(S) \cap N(B)|\bar{E})P(\bar{E})}{P(N(S))} \\ &= \frac{P(N(S) \cap N(B)|E)P(E) + P(\bar{E})}{P(N(S))} \\ &= \frac{P(N(S)|E)P(N(B)|E)P(E) + P(\bar{E})}{P(N(S))} \\ &= \frac{0.2 \times 0.5 \times 0.8 + 0.2}{0.36} \\ &= 0.778 \end{aligned} \quad (6)$$

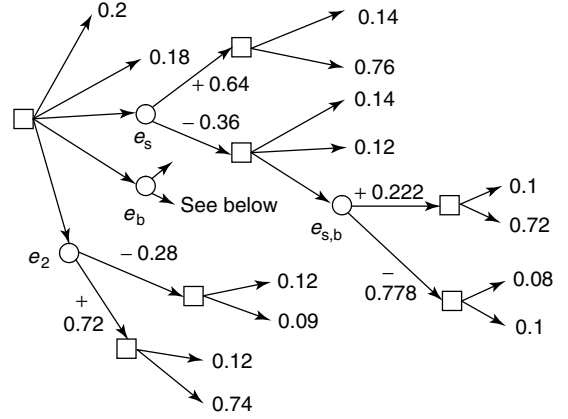


Figure 4 The completed tree (1)

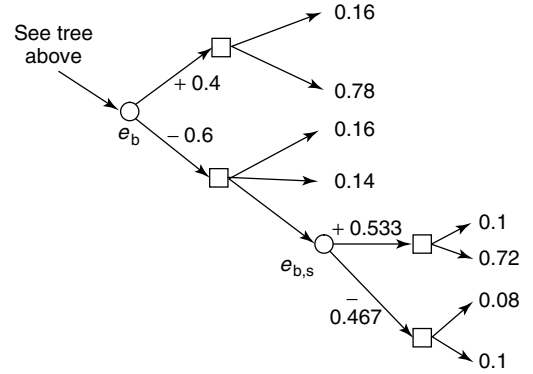


Figure 5 The completed tree (2)

The completed trees are given in Figures 4 and 5.

Folding Back a Tree to Find Optimal Strategies

Once the tree is embellished with edge probabilities and terminal utilities, it is straightforward to use it to calculate a Bayes, or expected utility optimizing decision. The solution algorithm follows a simple and self apparent principle. If you know the optimal decision rule(s) for a particular situation, then an optimal strategy requires you to make an optimal decision if you should reach that situation, then an optimal decision now must be one that plans to make one of these optimal decisions, if that situation is reached. It is straightforward to check that any decision satisfying this principle will give you a

Bayes decision rule: i.e., an optimal policy (see e.g. [2, 5]).

To illustrate how this works for a decision problem described by a tree, consider the possibility that the police choose to get the skirt analyzed, do not find a matching fiber, decide then to pay to check the blouse and obtain a positive result. The DM is then in the situation $(e_s, -, e_{s,b}, +)$, and can choose not to go to court (with an expected utility of 0.1) or to go to court (with an expected utility of 0.72). Clearly, when faced with this decision, any rational DM would choose to go to court. So, on observing this sequence of events, we know this is what the DM would do. Thus, provided the DM plans to choose rationally in the future, the expected utility on reaching $(e_s, -, e_{s,b}, +)$ is 0.72. Append this expected utility with the decision node $(e_s, -, e_{s,b}, +)$ and cross out the $\bar{e}$ edge out of this situation: it should never be used. Similarly on observing $(e_s, -, e_{s,b}, -)$, the DM notes that proceeding to go to court has an associated expected utility of 0.08 while releasing the suspect has a value 0.1. So clearly the DM should plan to release the suspect under this contingency. Cross out the edge labeled e and transfer to the decision node $(e_s, -, e_{s,b}, -)$ the expected utility for the release. Do this for all the penultimate decision nodes in the tree: i.e., the decision nodes all of whose edges lead to a calculated expected utility.

Each of these decision nodes has now inherited an associated expected utility. Now consider the chance nodes leading by an edge to these decision nodes. We can assign an expected utility to these nodes because we have calculated the probabilities of proceeding along each of the edges emanating from such a situation. Thus, consider the chance node $(e_s, -, e_{s,b})$. With probability 0.222 the police will obtain a utility 0.72 and with probability 0.778 a utility of 0.1. So using the formula for expectation, the expected utility associated with reaching this chance node – and then subsequently acting optimally – is

$$0.222 \times 0.72 + 0.778 \times 0.1 = 0.238 \quad (7)$$

Append this expected utility to this chance node and repeat this averaging procedure for all such chance nodes whose edges lead to a node whose associated expected utility has been calculated.

Continuing in this way, calculate expected utilities, appending to situations their expected utilities and crossing out any decision edges associated with suboptimal acts. Thus, for the decision node

$(e_s, -)$ we have calculated all expected utilities $\bar{u}$ associated with not going to court $\bar{e}$ ($\bar{u}(\bar{e}) = 0.14$), going to court e with no more analysis ($\bar{u}(e) = 0.14$), and proceeding to check the blouse for fibers $e_{s,b}$ ($\bar{u}(e_{s,b}) = 0.238$). Clearly, choosing the $e_{s,b}$ course is the best one. So transfer this utility to this node and delete the other edges. Since we know we will take to court if we obtain a match on the skirt with utility 0.76, the expected utility associated with e_s is

$$0.64 \times 0.76 + 0.36 \times 0.238 = 0.572 \quad (8)$$

The completed tree is given in Figures 6 and 7. Note that most software will provide such output automatically, using the algorithm described above.

Having proceeded to this point, we can read the Bayes decision rule. Start at the root and read along the undeleted paths to the remaining leaves,

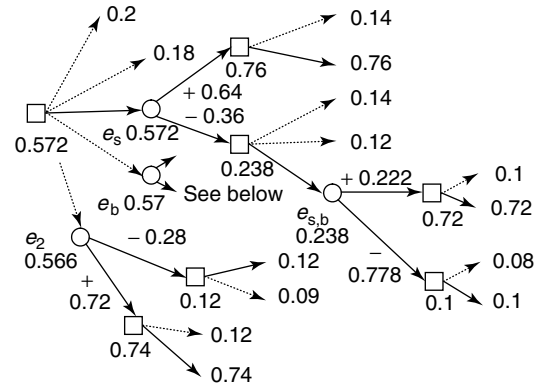


Figure 6 The tree evaluated completely (1)

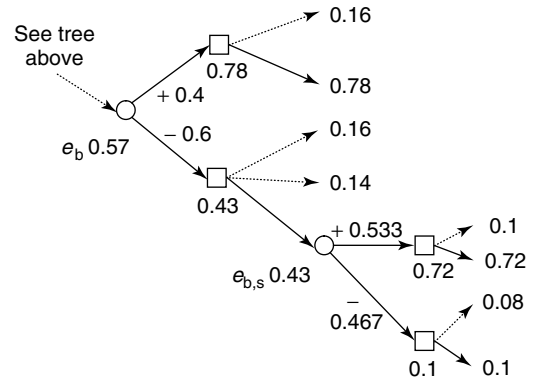


Figure 7 The tree evaluated completely (2)

translating into English as we go. Thus, in our example, the DM should take the undeleted edge e_s which is to pay for a search of a matching fiber in the skirt. If this is successful, +, the DM immediately takes the case to court. If a matching fiber is not found —, then the DM should pay to check the blouse. If a matching fiber is found on the blouse, the suspect should be taken to court, but if a match is not found, then the suspect should be released. This decision rule is utility maximizing for this problem.

Other Types of Decision Tree

There are two other types of decision trees that need to be mentioned. The first is the *normal form tree* (Figure 8) [2] and derives directly from Savage's description of a decision problem and the decision table (see **Decision Modeling**). The root node is a decision node with outgoing edges labeling each of the (many) decision *rules*. The next situation has outgoing edges labeled by each state of nature $\theta \in \Theta$. All subsequent situations are also chance nodes representing possible outcomes of experiments performed within the decision rule.

Such trees are very bushy, but are often manageable to write down after the first edge. Below is part of a normal form tree following the decision rule $d'(\mathbf{x})$: "Pay for the analysis of the skirt. If a fiber is found then take the suspect to court, otherwise do not." and letting E denote the event that the culprit sat on the sofa.

Note that there will inevitably be much repetition in such trees. However, there is also a computational advantage to normal form trees, since probabilities

typically appear in their natural causal order. They can therefore often be elicited directly and do not need to be calculated through formulae like the law of total probability or Bayes rule, as they were in the dynamic programming tree. This is useful both for transparency and also with respect to a sensitivity analysis where it is the elicited probabilities (and not functions of these) we would like to perturb. The expected utility associated with each decision rule is simple to calculate using the usual formula for expectation. Thus, in the example above,

$$\begin{aligned} \bar{u}(d'(\mathbf{x})) &= 0.8 \times 0.8 \times 0.76 + (0.8 \times 0.2 + 0.2) \\ &\times 0.14 = 0.537 \end{aligned} \quad (9)$$

In this way, we can calculate the expected utility associated with each decision rule. A Bayes rule will then be a decision rule with the highest expected utility.

Such trees are especially useful in understanding which decision rules are optimal for certain prior inputs and how sensitive a Bayes decision is to these probabilities. Thus, for example, we may want to investigate how sensitive the choice of decision rule is to the specification of the probability that the culprit sat on the sofa. It is easy to check that in this problem, for every possible decision rule, $\bar{u}(d'(\mathbf{x}))$ is a linear function of $P(E)$. Ideas of Pareto optimality then allow us to determine which decision rules could be optimal for some value of this probability and which ranges of the probability made which candidate Bayes decisions optimal (see, for example [2, 4]). Often, the number of such candidate decision rules is small. This type of analysis is called a *normal form analysis*.

A second tree, here called a *causal tree* [6] demands that the order of the nodes in the root-to-leaf paths of the tree is consistent with the order in which situations happen in the modeled context. This is often descriptively the most powerful decision tree and can be used for defining and analyzing various causal hypotheses. Such a tree does not necessarily support the direct implementation of backward induction/roll back. One typical root-to-leaf path on the example above can be written as the sequence of labeled edges: the culprit sits on sofa, a detectable fiber is transferred to her skirt, the police decide to analyze the skirt, the police detect the fiber on the skirt, the police take the suspect to court, the court finds the suspect guilty. Because each root-to-leaf path in this tree tries to represent how things happen

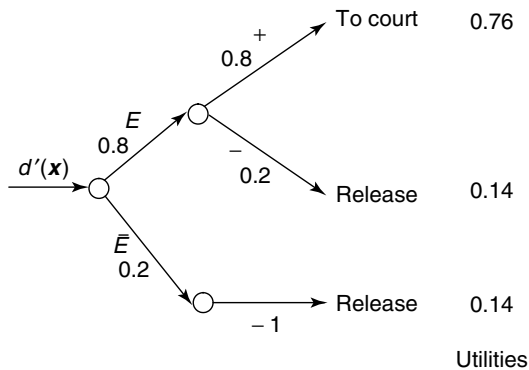


Figure 8 A normal form tree

rather than the order in which they are observed, this tree is often the most amenable to augmentation of hypotheses. In particular, unlike with many other trees, most of the probabilities associated with other possibilities can remain unaltered. For example, if we decide to consider a new possibility that the suspect dry cleaned her skirt so that any fibers would be found only with a small probability, then the path below could be changed to: culprit sits on sofa, a detectable fiber is transferred to her skirt, the suspect *does not dry clean her skirt*, the police decide to analyze the skirt, the police detect the fiber on the skirt, the police take the suspect to court, the court finds the suspect guilty. The associated path also appears (having very small probability): culprit sits on sofa, a detectable fiber is transferred to her skirt, the suspect *dry cleans her skirt*, the police decide to analyze the skirt, the police detect the fiber on the skirt, the police take the suspect to court, the court finds the suspect guilty.

Relationships to Other Graphical Decision Problem Representations

One problem with trees of all kinds is that although they represent asymmetric problems very well, nowhere in their topology is an explicit representation of hypotheses about dependence relationships between the state and measurement variables in the system. Such qualitative information is often elicited. For example, in the case study above, we have elicited for the police the qualitative information that the discovery of a fiber on the skirt is independent of the discovery of a fiber on the blouse, given the suspect sat on the sofa. Such statements tell us that many probabilities labeling the edges of a normal form or causal tree are equal to each other. However, this important information is not represented in the topology of the tree. On the other hand, it can be very easily represented by a Bayes Net, influence diagram (see **Influence Diagrams**), or other topological structures (see e.g. [7–9] and references therein). Elsewhere, we discuss a decision analysis (see **Decision Analysis**) using influence diagrams (see **Influence Diagrams**). Unlike decision trees, these structures can code continuous problems as effectively as discrete ones. Influence diagrams are, nevertheless, no panacea. In particular, they are only powerful when the decision problem has

close to a symmetric structure (in a sense defined in **Influence Diagrams**). So recent developments, both descriptive and algorithmic, often toggle between the two representations, simultaneously exploiting their different expressiveness (see e.g. [10]).

References

- [1] Puch, R.O. & Smith, J.Q. (2004). FINDS: a training package to assess forensic fibre evidence, *Advances in Artificial Intelligence*, Springer 420–429.
- [2] Raiffa, H. (1968). *Decision Analysis*, Addison–Wesley.
- [3] Phillips, L.D. (1984). A theory of requisite decision models, *Acta Psychologica* **56**, 29–48.
- [4] Smith, J.Q. (1988). *Decision Analysis: A Bayesian Approach*, Chapman & Hall.
- [5] Raiffa, H. & Schlaifer, R. (1961). *Applied Statistical Decision Theory*, MIT Press.
- [6] Shafer, G.R. (1996). *The Art of Causal Conjecture*, MIT Press, Cambridge.
- [7] Jaegar, M. (2004). Probability decision graphs – combining verification and AI techniques for probabilistic inference, *International Journal of Uncertainty Fuzziness and Knowledge Based Systems* **12**, 19–42.
- [8] Jensen, F.V., Neilsen, T.D. & Shenoy, P.P. (2004). Sequential influence diagrams: a unified asymmetry framework, in *Proceedings of the Second European Workshop on Probabilistic Graphical Models*, P. Lucas, ed, Leiden, pp. 121–128.
- [9] Thwaites, P.A. & Smith, J.Q. (2006). Non-symmetric models, chain event graphs and propagation, *Proceedings of the 11th IPMU Conference*, Paris, pp. 2339–2347.
- [10] Salmaron, A., Cano, A. & Moral, S. (2000). Importance sampling in Bayesian Networks using probability trees, *Computational Statistics and Data Analysis* **24**, 387–413.

Further Reading

Jensen, F.V. & Neilsen, T.D. (2007). *Bayesian Networks and Decision Graphs*, 2nd Edition, Springer-Verlag, New York.

Vomelelova, M. & Jensen, F.V. (2002). An extension of lazy evaluation for influence diagrams avoiding redundant variables in the potentials, in *Proceedings of the First European Workshop on Probabilistic Graphical Models*, J.A. Gamez & A. Salmeron, eds, Cuenca.

Related Articles

Bayesian Statistics in Quantitative Risk Assessment

Utility Function

JIM Q. SMITH AND PETER THWAITES

Default Correlation

A firm or an obligor is said to default on its debt when it cannot adhere to the payment schedule or other restrictions specified in the contractual agreement with its lenders. Examples of this include missed payment of interest or principal, a filing for bankruptcy as well as a distressed exchange. Joint default risk, or correlation between defaults, refers to the correlation between default events or default (or survival) times. The term is also sometimes applied to the correlation between changes in the default probabilities of individual firms.

Although default risk for an individual firm has been studied for several decades, it is only recently that the academic and the investment communities have focused on attempts to understand and model joint default risk. This interest has been driven primarily by the development of the credit derivatives market with financial products like collateralized debt obligations (CDO) whose payoff explicitly depends (*see Nonlife Insurance Markets; Credit Value at Risk*) on correlated defaults. Issuance of cash CDO reached (*see Potency Estimation; Credit Risk Models; From Basel II to Solvency II – Risk Management in the Insurance Sector*) \$268 billion in 2005 and that of synthetic CDO exceeded \$300 billion (2006 Structured Credit Insight, Morgan Stanley).

There is ample empirical evidence that defaults across firms are not independent. Figure 1 from Das *et al.* [1] graphs the number of defaults in North American firms by month over the period 1979 to 2004. The periods of significant clustering of default events coincide, as might be expected, with times of economic and financial stress like the 1991–1992 and 2001 US recessions.

There are at least three hypotheses regarding why defaults may be correlated. First, defaults may be correlated because firms are exposed to common or correlated risk factors. For example, Figure 1 suggests a common dependency on the economy. Second, there may be “contagion” between firms that have a contractual or other relationships with each other so that the default of one of the firms impacts the probability of default of the other firms to which it is related. The collapse of Penn Central Railway in 1970 – at the time, the largest railroad in the United States – resulted in the default of other

regional railroads. Third, default of firms may be correlated because of “frailty”. That is, suppose that there is a common default covariate that impacts the default risk of firms but it cannot be observed (or is observed imperfectly). Then, the default of a firm reveals information of this missing covariate and thus causes the default probability of other firms to be updated. Thus, even though there are no contagion-type effects, the default event of a firm impacts the probability of default of another firm. An example of frailty is the possible presence of accounting irregularities as implicated in the bankruptcy of Enron, whose default brought upon increased scrutiny of other firms.

There have been two main approaches to modeling and calibrating joint default risk. The first approach starts by modeling the default risk of an individual firm and then introducing joint default risk, for example, by allowing for a correlated default covariate across firms. The two principal theoretical setups for modeling the default risk of an individual firm are the structural model of Merton [2] and extensions thereof, and the reduced form models of Madan and Unal [3], Duffie and Singleton [4], and others. The second approach assumes that default probabilities of individual firms are exogenously available, and thus directly models joint default risk by allowing for a copula correlation.

Correlated Default in Structural Models

In Merton [2], default is (*see Credit Value at Risk*) assumed to occur only at maturity if the firm value is less than the face value of debt. Let the time t firm value V_t follow a geometric Brownian motion,

$$dV_t = \mu V_t dt + \sigma V_t dW_t \quad (1)$$

where μ is the instantaneous expected rate of return for the firm, σ is the volatility or diffusion parameter, and W_t is standard Brownian motion. The firm’s capital structure consists of equity, S_t , and a zero-coupon bond, B_t with (*see Numerical Schemes for Stochastic Differential Equation Models; Simulation in Risk Management*) maturity $T \geq t$ and face value of F . Then, the firm defaults if $V_T < F$. Let the

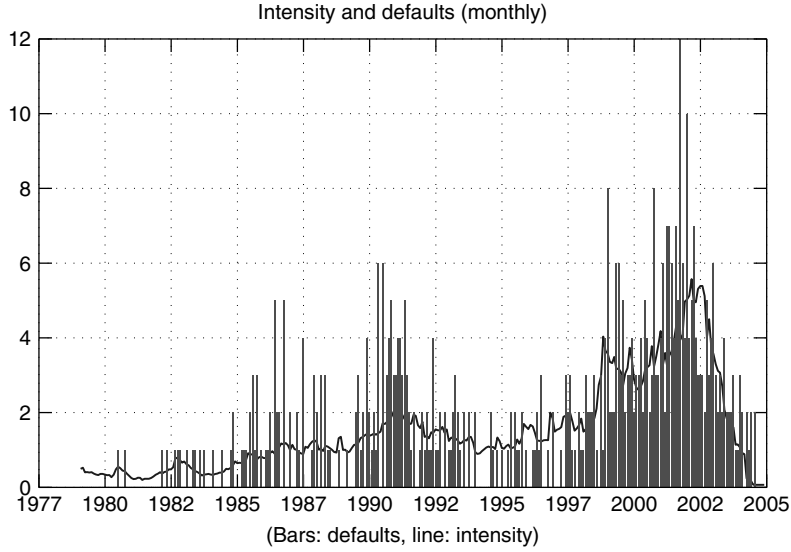


Figure 1 Intensities and Defaults. Aggregate (across firms) of monthly default intensities and number of defaults by month, from 1979 to 2004. The vertical bars represent the number of defaults, and the line depicts the intensity. The figure is taken from Das *et al.* [1] and explanation for the computation of the aggregate default intensity is provided therein [Reproduced from [1] with permission from The American Finance Association, © 2007.]

probability of default be $P[V_T < F]$. We can explicitly derive the probability,

$$\begin{aligned} P[V_T < F] &= P\left[V_{it}e^{\left(\mu - \frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}z} < F\right] \\ &= P[z < -DT D] \\ &= \Phi(-DT D) \end{aligned} \quad (2)$$

where z is a standard normal variate, $\Phi[\cdot]$ is the cumulative normal distribution, and $DT D$ is the “distance-to-default” (see **Credit Risk Models; Default Risk**) defined as,

$$DT D = \frac{\ln\left(\frac{V}{F}\right) + \left(\mu - \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}} \quad (3)$$

Empirically, estimates of the $DT D$ have been shown to be predictive of defaults (see, for example, Duffie *et al.* [5]) and used in commercial applications by Moody’s KMV.

As the only stochastic variable in the Merton [2] model is the firm value, correlation between defaults between two firms occur through the correlation

between their respective firms’ values. Let each firm follow a diffusion process,

$$dV_{it} = \mu_i V_{it} dt + \sigma_i V_{it} dW_{it} \quad i = 1, 2 \quad (4)$$

where W_{1t} and W_{2t} are correlated with correlation coefficient ρ . Let P_1 and P_2 be their respective probabilities of default as defined in equation (2) and the joint probability of default as the probability that at maturity both $V_{1,T} < F_1$ and $V_{2,T} < F_2$,

$$\begin{aligned} P(V_{1,T} < F_1, V_{2,T} < F_2) &= \int_{-\infty}^{DT D_1} \int_{-\infty}^{DT D_2} \frac{1}{2\pi\sqrt{1-\rho^2}} \\ &\quad \times \exp\left[\frac{z_1^2 - 2\rho z_1 z_2 + z_2^2}{2(1-\rho^2)}\right] dz_1 dz_2 \end{aligned} \quad (5)$$

The correlation between defaults, ρ^D , can then be computed as,

$$\rho^D = \frac{P(V_{1,T} < F_1, V_{2,T} < F_2) - P_1 P_2}{\sqrt{P_1(1-P_1)P_2(1-P_2)}} \quad (6)$$

Equation (6) explicitly relates the default correlation to asset correlations. Estimates of asset correlations are provided in Akhavan *et al.* [6].

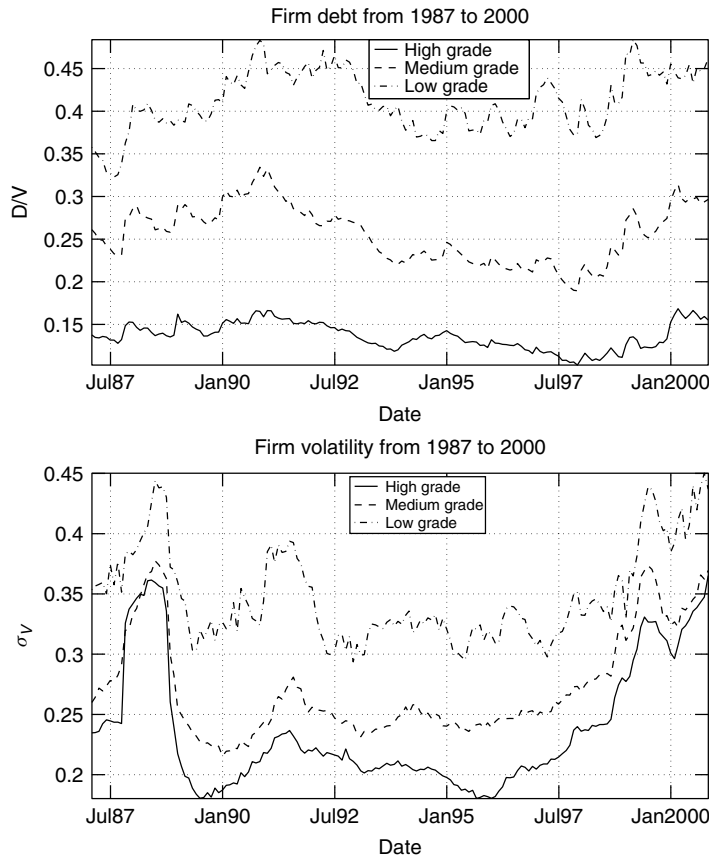


Figure 2 Debt and Volatility over 1987–2000. The plot from Das *et al.* [9] graphs the median debt fraction (D/V) and the firm return volatility (σ_V) of firms in each rating category for each of the 166 months over the period January 1987 to October 2000. The firm value and the firm return volatility are estimated on a monthly frequency using Merton [2]. The debt is measured as the sum of current debt plus half the long-term debt. Firms with Moody's rating single-A or higher are classified as high-grade, Ba and Baa are classified as medium-grade, and single-B and C as low grade [Reproduced from [9] with permission from Institutional Investor, Inc. Journals Group, © 2006.]

The Merton model can be extended to allow for default prior to maturity by considering default to have occurred the first time the firm value breaches a threshold [7]. Zhou [8] provides an explicit derivation of the joint default probability of two firms for such a model.

Although modeling the probability of joint default as a function of the correlation between the firms' asset value is intuitive, in practice, defaults may also be correlated as changes in firms' debt levels as well as in firms' volatilities are correlated. Figure 2 from Das *et al.* [9] plots, for each of three rating classes, the median debt level and firm volatility. It is clear that both debt levels and firm volatilities

vary over time and that these changes are correlated across firms in the economy. In particular, it appears especially important to account for the correlation of volatilities. Although it is possible to extend a Merton type model, given the difficulty of implementing and calibrating such extensions in a parsimonious framework, recent efforts have tended to focus on reduced form models.

Reduced Form Models

In reduced form models, given a properly defined probability space, default is an unpredictable event

occurring at the first jump time τ of a counting process governed by an intensity λ_t . That is, the conditional probability of default over a period between t and $t + dt$, conditional on surviving up to time t , is (in the limit) $\lambda_t dt$. If λ_t itself is stochastic, then τ is said to be doubly stochastic. The probability of survival of a firm up to time T is then equal to,

$$P(\tau > T) = E \left[\exp \left(\int_0^T \lambda_t dt \right) \right] \quad (7)$$

Explicit representations of equation (7) are possible for certain specifications for λ_t . In particular, when the stochastic differential equation for λ_t is an affine diffusion, then it is known from the results of Duffie *et al.* [10] that the solution of equation (7) is,

$$P(\tau > T) = \exp(a(t) + b(t)\lambda_t) \quad (8)$$

where $a(t)$ and $b(t)$ are solutions to a set of Riccati ordinary differential equations. Specific solutions for common specifications are available in Duffie *et al.* [10].

The doubly stochastic framework can be extended to n firms by allowing each firm to have its own intensity process, λ_{it} , $i = 1, \dots, n$. The first default time $\tau = \min(\tau_1, \dots, \tau_n)$ has the intensity $\Lambda_t = \sum_{i=1}^n \lambda_{it}$. Therefore, the conditional probability of survival of all firms up to time T is also given by equation (7) with Λ_t substituted for λ_t . That is, consider the joint survivorship event defined as $\{\tau_1 > t_1, \dots, \tau_n > t_n\}$ for any set of times $t_1 \leq t_2 \leq \dots \leq t_n$. For any time $t < t_1$,

$$P(\tau_1 > t_1, \dots, \tau_n > t_n) = E \left(\exp \left(- \int_t^{t_n} \Lambda_s ds \right) \right) \quad (9)$$

where Λ is the sum of the intensities of the n firms surviving at time t .

In the doubly stochastic framework, conditional on the information at time t , each default event is independent. The only source of correlation between default times is from the correlation between the intensities, λ_{it} . This provides a natural setting for testing the hypothesis of whether common or correlated factors affecting the default intensities are sufficient to explain the distribution of defaults. Das *et al.* [1] provide the theoretical framework for such a test.

Proposition Das *et al.* [1] Suppose that $(\tau_1, \dots, \tau_n)$ is doubly stochastic with intensity $(\lambda_1, \dots, \lambda_n)$. Let $K(t) = \#\{i : \tau_i \leq t\}$ be the cumulative number of defaults by t , and let $\Lambda_t = \int_0^t \sum_{i=1}^n \lambda_{iu} 1_{\{\tau_i > u\}} du$ be the cumulative aggregate intensity of surviving firms to time t . Then $J = \{J(s) = K(\Lambda^{-1}(s)) : s \geq 0\}$ is a Poisson process with rate parameter 1.

The empirical tests are based on the following corollary to the proposition.

Corollary Under the conditions of the proposition, for any $c > 0$, the successive numbers of defaults per bin,

$$J(c), J(2c) - J(c), J(3c) - J(2c), \dots$$

are iid Poisson distributed with parameter c .

That is, given a cumulative aggregate intensity (as, for example, plotted in Figure 1), the sample period can be divided into nonoverlapping time bins such that each bin contains an equal cumulative aggregate intensity of c . The doubly stochastic assumption then implies that the numbers of defaults in the successive bins are independent Poisson random variables with parameter c . Das *et al.* [1] test this implication using standard tests of the Poisson distribution.

Figure 3 plots the distribution for bin size 8 along with the corresponding theoretical Poisson distribution. It is apparent that the empirical distribution does not look like the theoretical distribution. In particular, the right tail is too “fat” compared with the Poisson distribution. The doubly stochastic property is rejected, suggesting that contagion- or frailty-type effects may also be important for understanding the clustering of defaults in the economy.

Modeling Correlated Default Using Copula

Let the marginal distribution providing the survival time (see **Comonotonicity; Simulation in Risk Management; Extreme Value Theory in Finance; Copulas and Other Measures of Dependency**) or default probability of each firm be given. The marginal distribution can be linked using a “copula” function, C , to create a joint distribution. The copula is a multivariate distribution function such that every marginal distribution is uniform on $[0, 1]$, and allows the separation of the marginal distribution from the

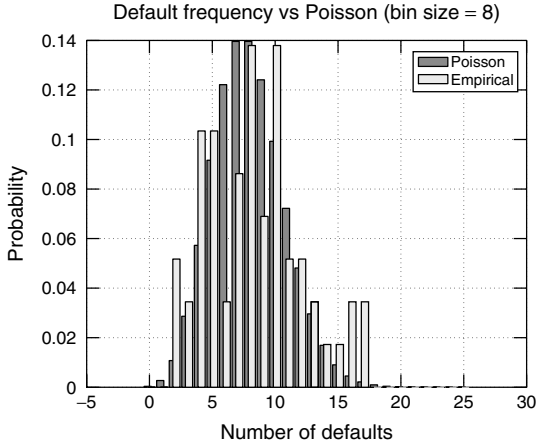


Figure 3 *Default distributions.* Das *et al.* [1]. The empirical and theoretical distributions of defaults for bin size 8. The theoretical distribution is Poisson [Reproduced from [1] with permission from The American Finance Association, © 2007.]

dependencies. Sklar [11] establishes that any multivariate distribution can be written in the form of a unique copula C if each of the univariate distributions are continuous.

A wide range of copula distributions have been proposed. The most commonly used copula as introduced by Li [12] is the Gaussian copula where,

$$C(z_1, \dots, z_n) = \Phi_n(\Phi^{-1}(z_1), \Phi^{-1}(z_2), \dots, \Phi^{-1}(z_n); \Sigma) \quad (10)$$

where $\Phi_n()$ is the multivariate normal distribution. The marginal distributions of z_i , $i = 1, \dots, n$ are Gaussian, and the copula links them together with the correlation matrix Σ . Another multivariate distribution that could be used as an alternative to the Gaussian is the t copula. There are also copulas that do not derive from multivariate distributions. These include a set of copulas called Archimedean copulas. Examples of the Archimedean copulas include the Gumbel and the Clayton copulas. The Gaussian copula is the most commonly used implementation in industry.

Li [12] provides illustrations of how the copula may be used to price a security that depends on joint default risk. Consider a first-to-default swap written on a portfolio of N names. For each firm i , let the distribution function for the survival time T_i be $F_i(t)$.

Assuming a Gaussian copula, the joint distribution of the survival times is,

$$F(t_1, \dots, t_N) = \Phi_N(\Phi^{-1}(F_1(t_1)), \Phi^{-1}(F_2(t_2)), \dots, \Phi^{-1}(F_N(t_N)); \Sigma) \quad (11)$$

Now to simulate defaults using the copula:

- Simulate N random numbers $z_1, z_2, \dots, z_N$ from an N -dimensional normal distribution with covariance matrix Σ .
- Obtain $T_1, T_2, \dots, T_N$ as $T_i = F_i^{-1}(\Phi(z_i))$, $i = 1, \dots, N$.

Given the default times from each simulation, the first-to-default time is the minimum of the default times of the N firms.

The copula methodology has been widely adapted in the industry as it does not require the development of a default risk model for an individual firm nor does it require a detailed modeling of the dependencies beyond the specification of the copula. That is, the dependency of defaults or survival times can be because of any common or correlated factors, contagion, or frailty. The use of copula may be considered a quick fix when sufficient empirical data is not available, or the underlying “true” economic model is unknown. This in turn is also the weakness of these models. They are difficult to calibrate to data and may not provide any additional economic insight.

Extensions

Given the empirical evidence that common and correlated factors may not be sufficient to model joint default risk and the limitation of copula-based models, there has been some recent effort in accounting explicitly for contagion (*see Credit Risk Models; Role of Alternative Assets in Portfolio Construction*) and frailty. As of yet, this effort should be considered in its infancy, especially for modeling contagion for which empirical evidence (apart from the well-known case of Penn Central) is scarce but see Davis and Lo [13]. There has been relatively more progress in the development of frailty models (for example, Giesecke [14] and Collin-Dufresne *et al.* [15]).

In particular, Duffie *et al.* [16] extend the model of Duffie *et al.* [5] by incorporating a frailty variable as an additional unobservable common default covariate. Empirically, the model improves upon the model without frailty by more closely explaining the right tail of defaults in the economy. For example, the actual number of defaults in the five-year period from January 1998 and December 2002 is 195. The model without frailty assigns almost zero probability to defaults being greater than 175. In contrast, for the model incorporating frailty, the realized number of defaults falls close to the 96th percentile. Thus, it appears that a model that allows for an unobservable variable may be of some use, especially when the right tail of defaults needs to be modeled.

Conclusion

In summary, there has been considerable progress in the modeling and the understanding of joint default risk. Earlier, efforts were made to extend structural models by allowing the stochastic variables to be correlated but these have been proved to be difficult to calibrate and implement in practice. More recent efforts have been toward extending reduced form models, where typically the correlation between survival times is driven only by the correlation between default intensities. Such a model does reasonably well, but underestimates the defaults in the right tail of the empirical distribution suggesting a role for frailty or contagion. There is ongoing effort especially in extending the existing models for frailty.

The rapid growth of the credit derivatives markets has required models that can be implemented for pricing securities. The industry has standardized in the use of the Gaussian copula for pricing products sensitive to joint default risk like CDOs. However, empirical evidence indicates that such a model fits only imperfectly (for example, the model fitted over different tranches of a CDO results in different correlations). It appears likely that the Gaussian copula will play a role similar to that of the Black–Scholes implied volatility in the equity markets in that it provides a convenient and easily implementable standard for comparative pricing.

In summary, there is much ongoing effort in both academic and the industry to understand and model joint default risk, and it is likely to remain for some time an area of rapid development.

References

- [1] Das, S., Duffie, D., Kapadia, N. & Saita, L. (2007). Common failings: how corporate defaults are correlated, *Journal of Finance* **62**(1), 93–118.
- [2] Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *The Journal of Finance* **29**, 449–470.
- [3] Madan, D.B. & Unal, H. (1998). Pricing the risk of default, *Review of Derivatives Research* **2**, 121–160.
- [4] Duffie, D. & Singleton, K. (1999). Modeling term structures of defaultable bonds, *Review of Financial Studies* **12**, 687–720.
- [5] Duffie, D., Saita, L. & Wang, K. (2006). Multi-period corporate default prediction with stochastic covariates, *Journal of Financial Economics* **83**, 635–665.
- [6] Akhavan, J.D., Kocagil, A.E. & Neugebauer, M. (2005). *A Comparative Empirical Study of Asset Correlations*, Working paper, Fitch Ratings, New York.
- [7] Black, F. & Cox, J.C. (1976). Valuing corporate securities: some effects of bond indenture provisions, *Journal of Finance* **31**, 351–367.
- [8] Zhou, C. (2001). An analysis of default correlation and multiple defaults, *Review of Financial Studies* **14**, 555–576.
- [9] Das, S., Freed, L., Geng, G. & Kapadia, N. (2006). Correlated default risk, *Journal of Fixed Income* **16**(2), 7–32.
- [10] Duffie, D., Pan, J. & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions, *Econometrica* **68**, 1343–1376.
- [11] Sklar, A. (1959). *Fonctions de Repartition a Dimensions et Leurs Marges*, Publications de l'Institut de Statistique de l'Universite de Paris, 229–231.
- [12] Li, D.X. (2000). On default correlation: a copula approach, *Journal of Fixed Income* **9**, 43–54.
- [13] Davis, M. & Lo, V. (2001). Infectious default, *Quantitative Finance* **1**, 382–387.
- [14] Giesecke, K. (2004). Correlated default with incomplete information, *Journal of Banking and Finance* **28**, 1521–1545.
- [15] Collin-Dufresne, P., Goldstein, R. & Helwege, J. (2003). *Is Credit Event Risk Priced? Modeling Contagion Via the Updating of Beliefs*, Working paper, Haas School, University of California, Berkeley.
- [16] Duffie, D., Eckner, A., Horel, G. & Saita, L. (2007). *Frailty Correlated Default*, Working paper, Stanford University.

Further Reading

Schönbucher, P. & Schubert, D. (2001). *Copula Dependent Default Risk in Intensity Models*, Working paper, Bonn University.

Default Risk

In modern finance, default risk (*see Credit Risk Models; Options and Guarantees in Life Insurance; Informational Value of Corporate Issuer Credit Ratings*) is typically interpreted as the probability of default, and default occurs when a debtor firm misses a contractually obligatory payment such as a coupon payment or repayment of principal. The occurrence of default is always costly for the lender and often triggers bankruptcy proceedings, in which case the assets of the firm are liquidated and the proceeds of the sale used to pay creditors according to prespecified rules. Defaults are more prone to occur where the general economic environment is poor; losses in such circumstances are then particularly painful.

Assessing default risk helps potential lenders determine whether they should give loans in the first place; if they do so, it also helps them to decide how much to charge for default risk, what covenant conditions, if any, to impose on the loan agreement, and how to manage the default risk through the course of the loan period. The assessment of default risks is also critical in the valuation of corporate bonds and credit derivatives such as basket-default swaps.

There is an important distinction between default risk under the actual (or P) probability measure and that under the risk-neutral (or Q) probability measure (*see Equity-Linked Life Insurance; Premium Calculation and Insurance Pricing; Mathematical Models of Credit Risk*). In the former, default risk is assessed in terms of the probabilities of actual default occurrences. In the latter, default risk is assessed in terms of risk-neutral probabilities. Typically, a decision to lend would be assessed using the P measure, but decisions relating to pricing would be assessed under the Q measure. The Q measure would also be used when backing out estimates of default risk from the prices of traded securities.

When default occurs, the actual loss inflicted depends on the loss given default (LGD) rate. The LGD is a proportion between 0 and 1 and is typically expressed as $1 - R$, where R is the nonnegative recovery rate (*see Compensation for Loss of Life and Limb; Credit Scoring via Altman Z-Score; Credit Value at Risk*), the proportion of the loan subsequently recovered after default has

taken place. The expected loss per dollar exposure is therefore

$$PD \times LGD \quad (1)$$

where PD is the default probability under the P measure. Note that the expected loss is sensitive to what we assumed about the recovery rate R . The size of the loss also depends on the creditor's exposure at default (EAD), typically represented by the notional size of the loan. Thus, the expected loss is equal to

$$EAD \times PD \times LGD \quad (2)$$

Defaults are often modeled as Poisson processes. However, when dealing with losses on portfolios of default-risky instruments, it is important to take account of correlations between defaults; this gives rise to the notion of the concentration risk in a portfolio. Since the default process is a binary one (i.e., a particular loan either defaults or it does not), it follows that the distribution of losses on a loan portfolio (typically) has a small left hand tail starting from zero, a relatively small positive mean, and a long heavy right hand tail.

Credit Ratings

One of the oldest means of assessing default risk is through credit ratings. A credit rating is an expert assessment of the likelihood that a debt will be paid in full and on time. These ratings are traditionally produced by independent rating agencies (*see Enterprise Risk Management (ERM); Informational Value of Corporate Issuer Credit Ratings*) such as Moody's or Standard & Poor's (S&P) and are usually provided for the major debt issues of large corporates or sovereigns. So, for example, in the S&P system there are 10 ratings (AAA, AA, A, BBB, BB, B, CCC, CC, C, and D), with AAA being the highest rating and D indicating that default has actually occurred. The "A" and "BBB" ratings indicate high-quality issues and are often referred to as *investment grade*; the others indicate lower quality issues and are known as *junk bonds*; and "C" rated issues indicate that the issue is more or less vulnerable to default.

Ratings are based on factors such as the rating agency's assessments of the economic prospects of the issuing firm, the quality of its management, the risks implicit in its lines of business, and the

degree of security offered on the debt issue. Typically, the rating agency will go through publicly available information such as prospectus, annual reports, and regulatory filings and meet with the debt issuer to obtain further confidential information. It may or may not give the issuer the opportunity to respond to its preliminary findings. It will then usually publish the rating and any supplementary analysis on which it is based, and charge the issuer a fee for the rating. In other cases, the rating might be carried out completely independent of the firm being rated, and ratings might be revealed on a confidential basis to the agency's fee-paying clients.

Whether published or not, the rating has a critical bearing on a debt issue's reception in the marketplace: a poor rating will discourage many investors from taking it up, and those that do take it up will demand higher yields to compensate for the associated default risk. A poor rating on a bond will therefore lead to a lower bond price, and the desire to avoid such an outcome gives an issuer an incentive to cooperate with the rating agency and persuade them of its creditworthiness.

More recently, many financial institutions have developed their own internal risk rating systems, which they apply to their loans. These produce estimates of the probability of default on a particular loan, but many also produce estimates of the loss given default. Combined with information about exposure at default, this information enables a bank to estimate the expected loss on a loan. These models help banks manage their loan portfolios on a consistent and transparent basis, and are mainly used to rate credit risks and determine the capital charges to apply against them. For example, the capital charge might be equal to the expected loss multiplied by some adjustment factor that takes account of concentration risk and possible error in the model.

Contingent Claims Approaches

Some of the most important approaches to default risk are the contingent claim approaches (*see Actuary; Longevity Risk and Life Annuities; Credit Scoring via Altman Z-Score*) – sometimes also known as *structural approaches* – that originate from a seminal article by Merton [1]. The motivation for these approaches builds on the limited liability rule that allows shareholders to default on their obligations

provided they hand over all the firm's assets to creditors. The firm's liabilities are therefore contingent claims issued against the firm's assets, and under the basic model default occurs at maturity when the value of the firm's assets falls short of the value of its liabilities. Subject to certain simplifying assumptions, default risk can then be modeled using regular option pricing methods.

To illustrate this model in a simple form, consider a firm with assets V . These are financed by equity S and a single debt instrument with face value F maturing at time T . The market value of this instrument is B . Under standard assumptions, the value of the firm at time $t = 0$ is given by

$$V_0 = S_0 + B_0 \quad (3)$$

Default occurs at time T in the event such that $V_T < F$. Merton's insight was to regard the option to default as a put option on the firm's assets. If we make the additional assumptions needed to justify the Black–Scholes option pricing model (*see Risk-Neutral Pricing: Importance and Relevance; Statistical Arbitrage*) – a Gaussian diffusion process, and so on – then the value of this put, P_0 , can be shown to be

$$P_0 = -N(-d_1)V_0 + Fe^{-rT}N(-d_2) \quad (4)$$

where

$$d_1 = \frac{\ln(V_0/Fe^{-rT}) + \sigma^2T/2}{\sigma\sqrt{T}}, d_2 = d_1 - \sigma\sqrt{T} \quad (5)$$

and where $N(\cdot)$ is the cumulative standard normal distribution, r is the risk-free rate, and σ is the volatility rate of the firm's assets. We can now see that the cost of this option depends on the firm's leverage ratio Fe^{-rT}/V_0 (or the present value of its debt divided by the present value of its assets), σ , the term to maturity of the debt T , and r .

This option pricing approach allows us to recover other parameters of interest. For example, the probability of default under the Q measure is

$$\Pr[V_T < F] = N\left(\frac{\ln(F/V_0) - (r - \sigma^2/2)T}{\sigma\sqrt{T}}\right) \quad (6)$$

This tells us that the probability of default is increasing with F , decreasing with V_0 , and (for $V_0 > F$, as seems reasonable) increasing with σ , all of which accord with economic intuition.

An implication of a positive default probability is that the yield on the debt must exceed the risk-free rate to compensate the lenders for the default risk they are bearing. Under the Merton model (*see Credit Value at Risk*), the resulting default spread π – the difference between the yield on the debt and the risk-free rate – is given by

$$\pi = -\frac{1}{T} \ln \left(N(d_2) + \frac{V_0}{Fe^{-rT}} N(-d_1) \right) \quad (7)$$

which tells us that an increase in the leverage ratio leads to a rise in the spread. Since d_1 and d_2 themselves depend on σ , we can also infer that the spread rises with the volatility. Both of these results are again in line with economic intuition.

The Merton model can be extended in many ways. Extensions include first-passage time models, in which the firm is allowed to default not just at a fixed time T , but at some random time when the V hits a critical threshold; jump-diffusion models, in which the assumption of Gaussian diffusion is supplemented with a jump process in the value of the firm; and models with stochastic or endogenous default barriers. The univariate Merton model can also be generalized to multivariate cases: these allow us to determine the default risk associated with a portfolio of bonds or corporate debt.

A particularly influential version of the Merton model is the KMV model [2]. This model was developed in the 1990s by a private firm, KMV, named after its founders Kealhofer, McQuown and Vasicek. This model is widely used and [3] reports that 40 of the biggest 50 banks in the world use it. The model itself is a relatively straightforward extension of the Merton model, but its distinctive strength lies in its empirical implementation using a large proprietary database and in the quality of the empirical research underlying this implementation. Once the capital structure of the firm is specified, the KMV model allows one to estimate expected default frequencies (EDFs) for any chosen horizon period. The main modeling innovation in the KMV model is to introduce an additional state variable, the so-called distance to default, which is the number of standard deviations that a firm at any given time is away from default: the greater the distance to default, the lower the probability of default. The use of this variable allows for a more sophisticated relationship between default probability and asset value than is the case in the Merton model: it can accommodate the impact

of factors such as heavy-tailed asset values, more sophisticated capital structures, and the possibility that default will not automatically lead to liquidation. The actual inference from distance to default to default probability is obtained using KMV's proprietary software calibrated to their database of historical defaults.

The KMV model appears to perform well in practice. Evidence suggests that the EDFs of firms that subsequently default rise sharply a year or two before default, and that changes in EDFs also anticipate the credit downgrades of traditional rating agencies. Ratings agencies are often slow to adjust their ratings, and this can be a problem if the credit quality of a firm deteriorates suddenly, which often happens if a firm is close to default. The EDFs estimated by the KMV model will tend to react more quickly to such changes, because they are reflected in the share price and therefore in the firm's estimated distance to default. Another advantage of the KMV model is that its EDFs tend to reflect the current macroeconomic environment, which rating agencies tend to ignore. This means that KMV forecasts are more likely to be better predictors of short-term default probabilities than those based on credit ratings.

On the other hand, contingent claims approaches require information about firm values and volatilities that is often not available. Estimates of firm values have to be obtained from information contained in the firm's stock price and balance sheet, and such estimates are usually dependent on questionable assumptions about factors such as capital structure and the distribution of asset values. There are also downsides to the fact that KMV model takes account of stock price valuations: first, the KMV approach is difficult to implement for firms that are not publicly traded, precisely because it requires information on stock values; in addition, if the stock market is excessively volatile, the KMV model can produce estimates of default probabilities that are also excessively volatile. Estimates of volatilities are also difficult to obtain and are dependent on questionable assumptions about factors such as capital structure. For their part, multivariate versions of these models are also dependent on simplistic treatments of statistical codependence – most typically, that we can model codependencies between defaults using Pearson correlations, which implicitly assumes that the underlying risk factors are multivariate

elliptical – and these correlations are also difficult to estimate using the limited data available. A final weakness of many of these contingent claims models is that they ignore credit migration, i.e., they treat default probabilities as fixed over the horizon period, and this ignores the tendency of ratings to migrate over time.

Credit Migration Approaches

The next class of models are those built on a credit migration (*see* **Credit Migration Matrices**) or transition matrix that represents the set of probabilities that any given credit rating will change to any other over the horizon period. These probabilities can be inferred from historical default data, from the credit ratings themselves, or from models such as KMV. Typically, these probabilities will suggest that a firm with any given rating is most likely to retain that rating by the end of the horizon period; however, there are small probabilities of upgrades or downgrades to adjacent ratings, and even smaller probabilities of changes to more distant ratings. To give a typical example, an A rated bond might have a 0.05% probability of migrating upward to AAA, a 2.5% probability of migrating to AA, a 90% probability of remaining at A, a 5% probability of being downgraded to BBB, a 0.5% probability of being downgraded to BBB, and very low or negligible probabilities of lower downgrades.

The best-known migration model, the CreditMetrics model, considers the forward value of a loan at the end of the horizon period for each possible end-horizon credit rating: its value if it ends up at AAA, AA, and so forth, down to its value if it defaults, which would also be conditional on recovery rate assumptions [4]. These values are found by discounting the loan's cash flows at a rate equal to the risk-free end-horizon forward rate plus an estimate of the credit spread for that rating. The combination of transition probabilities and end-horizon loan values enables the modeler to apply value-at-risk (VaR) analysis and obtain the credit VaR as the current loan value minus the relevant quantile of the distribution of forward loan values.

In the case of a single loan taken out by a particular obligor, the only transition probabilities needed are those for that obligor's current rating. More generally, calculations for a portfolio of loans would be based on a complete transition matrix: in

the case of the CreditMetrics model, this would be an 8×8 matrix showing the probabilities that any current AAA, AA, A, BBB, BB, B, CCC, and D rating would change to any other.

However, we cannot reasonably expect default and migration probabilities to remain constant over time. There is therefore a need for an underlying structural model that ties these to more fundamental economic variables. CreditMetrics responds to this need by invoking the Merton model and making associated simplifications to obtain these probabilities from the joint distribution of the equity returns of obligor firms. An alternative solution is offered by the CreditPortfolioView model [5, 6]: this model relates these probabilities to macroeconomic variables such as the unemployment rate, the economic growth rate, the level of long-term interest rates, and savings rates. This model therefore captures the notion that the credit cycle closely follows the business cycle – when the economy worsens, both downgrades and defaults increase, and *vice versa*.

However, all these models involve simplistic assumptions of one sort or another and are highly reliant on suitable datasets.

Intensity Models of Default Risk

There are also intensity-based models, sometimes also referred to as reduced-form models. These do not specify the process leading to default, as such; instead, default is modeled as an exogenous process that occurs randomly in accordance with a fitted intensity or hazard function. These models are fundamentally empirical and do not embody any economic theory of default: they make no attempt to relate default risk to capital structure and make no assumptions about the causes of default. To give an illustration, one simple intensity-based model postulates that the observed credit spread (π) is equal to the product of the default probability and the loss given default, *viz*:

$$\pi = PD \times LGD \quad (8)$$

As an aside, one will note that equation (8) is similar to equation (1). However, whereas equation (1) is an exercise under the P measure, equation (8) would be an exercise under the Q measure, because it would be used to back out default probabilities from observed bond prices. Any spread depends on both the maturity

of the loan and the credit rating of the borrower, and is usually observable. The probabilities of default for different maturities and credit ratings can then be recovered if we can estimate or specify the loss given default (or the recovery rate).

Since intensity-based models are based on empirical debt prices, they are able to reflect complex default term structures better than some of the earlier approaches considered. However, corporate bond markets can be rather illiquid, and the pricing data available are often inaccurate. These data problems mean that, even though they are made to fit the available data, we cannot take for granted that fitted intensity models will necessarily provide better estimates of “true” default probabilities than other methods.

Another common type of intensity model postulates an exogenous default process that would be fitted to historical default data. Subject to certain conditions, we can model defaults by a Poisson process. However, a basic Poisson process is unlikely to be accurate because it assumes the default intensity parameter to be constant and we would expect it to be correlated with the credit cycle. To get around this problem, we can model defaults as driven off a Poisson process with a stochastic intensity that takes account of business cycle and other macroeconomic factors that affect it. A good example is the CreditRisk+ model of Credit Suisse Financial Products, which models default rates as functions of gamma-distributed background factors [7]. The CreditRisk+ model is attractive because it yields closed-form solutions for default risks, which are very useful computationally; it is also attractive in so far as the only information it requires for any instrument is information about default rates and exposures.

General Problems and Future Objectives in Default Risk Modeling

While different default risk models have their own characteristic strengths and weaknesses, all models of default risk have generic weaknesses. One weakness is their exposure to model risk: it is often difficult to empirically discriminate between alternative models, so we can never be sure that any model we choose will be the “best” one. This is so even where we are dealing with approaches in which the “model” is

wholly or partially implicit, as is the case with credit ratings: any approach to default risk must be based on at least some assumptions that have to be taken on trust and cannot be tested. However, even within the confines of any given model (or approach), all are subject to parameter risk: there will be certain critical parameters whose values cannot be precisely determined even if we had unlimited amounts of past data. This is because parameters are, in principle, calibrated rather than estimated: even if we had “perfect” parameter estimates based on past data – which is never the case – model calibration requires us to make judgments over the extent to which the values of those same parameters might differ from their estimated historical values. Calibration always requires judgment, which is necessarily subjective and open to error. Furthermore, data are inevitably limited in supply and often flawed in quality, so reliable estimates of “past” parameter values can be difficult to obtain. Evidence also suggests that estimates of default probabilities are likely to be sensitive to the values of key parameters such as volatilities and correlations. Correlations are especially troublesome, not only because data are scarce, but also because the correlations used in practice are obtained using assumptions that are empirically doubtful. Estimates of expected losses on bond portfolios are also likely to be sensitive to the values of these same parameters, as well as to assumptions about recovery rates.

Much more work remains to be done. One major objective of ongoing research is to build models that adequately capture the market risks associated with default-risky positions: most current models of default risk ignore the market risks involved and ignore interactions between market and credit risk factors. Another important objective is to develop default risk models that can provide better valuations of default-risky positions, especially positions in credit derivatives. The explosive growth in the credit derivatives market suggests that practitioners are making good progress in this front. Nonetheless, the development of more sophisticated default risk models will be an important factor in the development of more advanced credit derivatives and in the future growth of the credit derivatives market.

References

- [1] Merton, R.C. (1974). On the pricing of corporate debt, *Journal of Finance* **28**, 449–470.

- [2] Kealhofer, S. & Bohn, J.R. (2001). *Portfolio Management of Default Risk*, KMV working paper, at www.kmv.com.
- [3] Berndt, A., Douglas, R., Duffie, D., Ferguson, F. & Schranz, D. (2004). *Measuring Default Risk Premia from Default Swap Rates and EDFs*, Preprint, Stanford University.
- [4] Morgan, J.P. (1997). *CreditMetrics Technical Document*, New York.
- [5] Wilson, T. (1997). Portfolio credit risk I, *Risk* **10**, 111–117.
- [6] Wilson, T. (1997). Portfolio credit risk II, *Risk* **10**, 56–61.
- [7] Credit Suisse Group (1997). *CreditRisk+: A Credit Risk Management Framework*, Credit Suisse Financial Products, New York.

Further Reading

- Crosbie, P.J. & Bohn, J.R. (2002). *Modeling Default Risk*, KMV working paper, at www.kmv.com.
- Crouhy, M., Galai, D. & Mark, R. (2001). *Risk Management*, McGraw-Hill, New York.
- Lando, D. (2004). *Credit Risk Modeling: Theory and Applications*, Princeton University Press, Princeton.
- McNeil, A.J., Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management*, Princeton University Press, Princeton.

KEVIN DOWD

Degradation and Shock Models

Stochastic degradation (*see Common Cause Failure Modeling*) in engineering, ecological, and biological systems is naturally modeled by increasing (decreasing) stochastic processes to be denoted by $W_t, t \geq 0$. We are interested in modeling stochastic degradation as such and in considering the first passage times, when this degradation reaches the predetermined or random level D . The latter interpretation can be useful for risk and safety assessment, when D defines some critical safety region. When, for instance, degradation implies the diminishing resistance to loads in some structures, it can result not just in an “ordinary” failure, but in a severe catastrophic event.

The main focus of this paper is on shock models (*see Nonlife Insurance Markets; Repair, Inspection, and Replacement Models; Mathematics of Risk and Reliability: A Select History*), which, in many instances, form the basis for the corresponding stochastic modeling of degradation. Shocks are generally interpreted as instantaneous harmful random events that represent danger to human beings, the environment, or the economic values. Therefore, probabilistic description of shock processes is important in analysis of different risks.

In the section titled “‘Direct’ Models of Degradation”, some simple models of degradation, not related to shocks, are considered. In the sections titled “Shot-Noise Process” and “Asymptotic Cumulative Shocks Modeling” the cumulated shock models are discussed, and simple asymptotic results are presented. In many applications, the number of shocks in the time interval of interest is large, which makes it possible to apply asymptotic methods rather effectively. In the section titled “Noncumulative Shock Models”, some important noncumulative shock models, especially in safety and risk assessment, are briefly described. The impact of a single (noncumulative) shock can be disastrous or even fatal in some settings, therefore this case needs careful attention. Finally, some weaker criterions of failures or of critical values of deterioration caused by shocks are discussed in the section titled “Weaker Criterions of Failures”.

In this presentation, we try to focus on the meaning of the models and interpretations of results, but the

subject is rather technical and one cannot avoid certain mathematical details.

“Direct” Models of Degradation

In this section we briefly define several approaches that are most often used in engineering practice for degradation modeling. The simplest one that is widely used is the path model, the stochastic nature of which is described either by the additive or by the multiplicative random variable.

$$W_t = \eta(t) + Z \quad (1)$$

$$W_t = \eta(t)Z \quad (2)$$

where $\eta(t)$ is an increasing, continuous function ($\eta(0) = 0, \lim_{t \rightarrow \infty} \eta(t) = \infty$) and Z is a nonnegative random variable with the cumulative distribution function (cdf) $G(z)$. Therefore, the sample paths (realizations) for these models are monotonically increasing. Thus, the “nature” of the stochastic process is simple and meaningful: let the failure (catastrophe) be defined as *reaching* by $W_t, t \geq 0$ the degradation threshold $D > 0$ and T_D be the corresponding time to failure random variable with the cdf $F_D(t)$. It follows, e.g., for model (2) that

$$\begin{aligned} F_D(t) &= \Pr(W_t \geq D) = \Pr\left(Z \geq \frac{D}{\eta(t)}\right) \\ &= 1 - G\left(\frac{D}{\eta(t)}\right) \end{aligned} \quad (3)$$

Example 1 Let $\eta(t) = t$ and assume the Weibull law for Z : $G(z) = 1 - \exp\{-(\lambda z)^k\}$, $\lambda, k > 0$. Then, in accordance with relation (3)

$$F_D(t) = \exp\left\{-\left(\frac{\lambda D}{t}\right)^k\right\} \quad (4)$$

which is often called the *inverse Weibull distribution* [1]. Specifically, when $\lambda = 1, k = 1$,

$$F_D(t) = \exp\left\{-\frac{D}{t}\right\} \quad (5)$$

It is clear that the value at $t = 0$ for this distribution should be understood as

$$F_D(0) = \lim_{t \rightarrow 0} F_D(t) = 0 \quad (6)$$

The inverse-Weibull distribution is a convenient and simple tool for describing threshold models with a linear function $\eta(t)$.

Assume now that the threshold D is a random variable with the cdf $F_0(d) = \Pr(D \leq d)$ and let, at first, degradation be modeled by the deterministic, increasing function $W(t)$ ($W(0) = 0, \lim_{t \rightarrow \infty} W(t) = \infty$). Or, equivalently, the problem can be reformulated in terms of the fixed threshold and random initial value of degradation. Denote by T the random time to failure (degradation reaches the threshold value). As events $T \leq t$ and $W(t)$ are equivalent, similar to equation (3) [2], we have

$$F(t) \equiv \Pr(T \leq t) = \Pr(D \leq W(t)) = F_0(W(t)) \quad (7)$$

where the last equality is owing to the fact that the cdf of D is $F_0(d)$. Substituting d by $W(t)$, finally results in equation (7).

Now, let a deterministic degradation $W(t)$ in equation (7) turn to a stochastic process $W_t, t \geq 0$. To obtain the corresponding distribution of the time to failure, in this case, we must obtain the expectation of $F_0(W_t)$ with respect to the process $W_t, t \geq 0$

$$F(t) = E[F_0(W_t)] \quad (8)$$

This equation is too general, as the stochastic process is not specified. The following example considers the path model for $W_t, t \geq 0$.

Example 2 Let, e.g., $F_0(d) = 1 - \exp\{-\lambda d\}$ and $W_t = \eta(t)Z$, where Z is also exponentially distributed with parameter μ . Direct integration in equation (8) gives

$$\begin{aligned} F(t) &= E[1 - \exp\{-\lambda \eta(t)Z\}] \\ &= \int_0^\infty (1 - \exp\{-\lambda \eta(t)z\}) \mu \exp\{-\mu z\} dz \\ &= 1 - \frac{\mu}{\mu + \lambda \eta(t)} \end{aligned} \quad (9)$$

Probably, the most popular and well investigated stochastic process is the Wiener process (see **Life-time Models and Risk Assessment**). The Wiener process with drift is also often used for modeling wear, although its sample paths are not monotone (but the mean of the process is a monotonically increasing function). It is defined as

$$W_t = \mu t + X(t) \quad (10)$$

where $\mu > 0$ is a drift parameter and $X(t)$ is a standard Wiener process: for the fixed $t \geq 0$ the random variable $X(t)$ is normally distributed with zero mean and variance $\sigma^2 t$. It is well known [3] that the first passage time T_D

$$T_D = \inf_t \{t, W_t > D\} \quad (11)$$

is described in this case by the inverse Gaussian distribution (see **Bonus–Malus Systems**)

$$\begin{aligned} \bar{F}_D(t) &= \Pr(T_D > t) = \Phi\left(\frac{D - \mu t}{\sqrt{t}\sigma}\right) \\ &\quad - \exp\{-2D\mu\} \Phi\left(\frac{D + \mu t}{\sqrt{t}\sigma}\right) \end{aligned} \quad (12)$$

$$E[T_D] = \frac{D}{\mu}, \text{Var}(T_D) = \frac{D\sigma^2}{\mu^3} \quad (13)$$

where $\Phi(t)$, as usual, denotes the cdf of the standard normal random variable.

Another popular process for modeling degradation is the gamma process. Although the estimation of the parameters for the degradation models driven by the gamma process is usually more complicated than for the Wiener process, it better captures the desired monotonicity. The gamma process is a stochastic process $W_t, W_0 = 0$ with independent nonnegative increments having a gamma cdf with identical scale parameter [4]. The increment $W_t - W_\tau$ has a gamma distribution with a shape parameter $v(t) - v(\tau)$ and a scale parameter u , where $v(t)$ is an increasing function and $v(0) = 0$. Therefore, W_t is gamma distributed with a shape parameter $v(t)$ and a scale parameter u and

$$E[W_t] = \frac{v(t)}{u}, \text{Var}(W_t) = \frac{v(t)}{u^2} \quad (14)$$

The first passage time T_D , which for the monotonically increasing processes is just the time of reaching D , is described in this case by the following distribution [5]:

$$F_D(t) = \Pr(T_D \leq t) = \Pr(W_t \geq D) = \frac{\Gamma(v(t), Du)}{\Gamma(v(t))} \quad (15)$$

where $\Gamma(a, x) = \int_x^\infty t^{a-1} e^{-t} dt$ is an incomplete gamma function for $x > 0$. Thus, deterioration, which takes place in accordance with the gamma process, can be effectively modeled in this case.

A natural way of modeling additive degradation is via the sum of random variables, which represent the degradation increments

$$W_t = \sum_{i=1}^n X_i \quad (16)$$

where $X_i, i = 1, 2, \dots, n$ are positive independent and identically distributed (i.i.d.) random variables, with a generic variable denoted by X , and n is an integer.

Example 3 [3] Assume that a mechanical item is employed in a number of operations and each of them increases the item's wear on X units. Assume that X is normally distributed with $E[X] = 9.2$ and $\sqrt{\text{Var}(X)} = 2.8$, which guarantees that for practical reasons this random variable can be considered as a positive one. Let the initial wear be zero. The item is replaced by a new one if the total degree of wear exceeds 1000. What is the probability of replacing the item after 100 operations are performed?

Taking into account the property stating that the sum of n i.i.d. normally distributed random variables is a normally distributed random variable with parameters $nE[X]$ and $\sqrt{\text{Var}(X)n}$, we obtain

$$\begin{aligned} \Pr\left(\sum_{i=1}^{100} X_i \geq 1000\right) &= 1 - \Phi\left(\frac{1000 - 920}{28}\right) \\ &= 1 - \Phi(2.86) = 0.021 \end{aligned} \quad (17)$$

Thus the probability of an item's replacement after 100 cycles is 0.021.

The next step to a more real stochastic modeling is to view n as a random variable N or some point process $N_t, t \geq 0$. The latter is counting point events of interest in $[0, t], t \geq 0$. The result is called the *compound point process*

$$W_t = \sum_{i=1}^{N_t} X_i \quad (18)$$

Denote by $Y_i, i = 1, 2, \dots$ the sequence of inter-arrival times for $N_t, t \geq 0$. If $Y_i, i = 1, 2, \dots$ are i.i.d (and this case will be considered in what follows) with a generic variable Y , then the Wald's equation [3] immediately yields

$$E[W_t] = E[N_t]E[X] \quad (19)$$

where, specifically for the compound Poisson process with rate m , $E[N_t] = mt$. It can be shown [6] that under certain assumptions, the stationary gamma process ($v(t) = vt$) can be viewed as a limit of a specially constructed compound Poisson process. An important modification of equation (18) of the next section surprisingly results in the fact that the limiting distribution of W_t is gamma with a shape parameter, which already does not depend on t .

Relation (18) has a meaningful interpretation via the shocks – instantaneous external events causing random amount of damage (degradation) $X_i, i = 1, 2, \dots$ and eventually resulting in an accumulated damage W_t . Numerous shock models were considered in the literature (see [7] and references therein). Apart from degradation modeling in various applications, shock models also present a convenient tool for analyzing certain nonparametric properties of distributions [8]. The explicit expressions for W_t usually cannot be obtained. Some results in terms of the corresponding Laplace transforms can be found in [9–11], but effective asymptotic results exist and will be discussed in the section titled “Asymptotic Cumulative Shocks Modeling”.

Shot-Noise Process

A meaningful modification of the cumulative shock model (18) is given by the shot-noise process [12, 13]. In this model, the additive input of a shock of magnitude X_i in $[t, t + dt]$ in the cumulative damage W_t is decreased in accordance with some decreasing (nonincreasing) response function $h(t - s)$. Therefore, it is defined as

$$W_t = \sum_{i=1}^{N_t} X_i h(t - \tau_i) \quad (20)$$

where $\tau_1 < \tau_2 < \tau_3, \dots$ is the sequence of the corresponding arrival (waiting) times in the point process. This setting has a lot of applications in electrical engineering, materials science, health sciences, risk, and safety analysis. For instance, cracks due to fatigue in some materials tend to close up after the material has borne a load that has caused the cracks to grow. Another example is the human heart muscle's tendency to heal after a heart attack [14]. Thus the inputs of shocks in accumulative damage decreases with time.

Equivalently, definition (20) can be written as

$$W_t = \int_0^t Xh(t-u) dN_u \quad (21)$$

where $dN_u = N(t, t+du)$ denotes the number of shocks in $[t, t+dt)$. Firstly, we are interested in the mean of this process. As $X_i, i = 1, 2, \dots$ are independent from the point process $N_t, t \geq 0$ and assuming that $E[X] < \infty$

$$\begin{aligned} E[W_t] &= E[X] \int_0^t h(t-u) dN_u \\ &= E[X] \int_0^t h(t-u)m(u) du \end{aligned} \quad (22)$$

where $m(u) = dE[N_u]/du$ is the rate (intensity) of the point process. For the Poisson process, $m(u) = m$ and

$$E[W_t] = mE[X] \int_0^t h(u) du \quad (23)$$

Therefore, distinct from equation (19), asymptotically the mean accumulative damage is finite, when the response function has a finite integral

$$\lim_{t \rightarrow \infty} E[W_t] < \infty \text{ if } \int_0^\infty h(u) du < \infty \quad (24)$$

which has an important meaning in different engineering and biological applications. It can be shown directly from definitions that, if $E[X^2] < \infty$,

$$\begin{aligned} \text{Cov}(W_{t_1}, W_{t_2}) &= mE[X^2] \int_0^{t_1} h(t_1-u)h(t_2-u) du; \\ &\quad t_1 \geq t_2 \end{aligned} \quad (25)$$

The central limit theorem also takes place for sufficiently large m in the following form [15, 16]:

$$\frac{W_t - E[W_t]}{(\text{Var}(W_t))^{1/2}} \xrightarrow{D} N(0, 1), t \rightarrow \infty \quad (26)$$

where the sign “ D ” means convergence in distribution and, as always, $N(0, 1)$ denotes the standard normal distribution. The renewal case gives similar results.

$$\lim_{t \rightarrow \infty} E[W_t] = \frac{1}{E[X]} \int_0^\infty h(u) du \quad (27)$$

Example 4 Consider a specific exponential case of the response function $h(u)$ and the Poisson process of shocks with rate m

$$W_t = \sum_{i=1}^{N_t} X_i \exp\{\alpha(t - \tau_i)\} \quad (28)$$

By straightforward calculations, [12] using the technique of moment generating functions, it can be shown that for t sufficiently large the stationary value W_∞ has a gamma distribution with mean $m/\lambda\alpha$ and variance $m/\lambda^2\alpha$

$$F_D(t) = \Pr(T_D \leq t) = \Pr(W_t \geq D) = \frac{\Gamma(m/\alpha, D\lambda)}{\Gamma(m/\alpha)} \quad (29)$$

It is well known from the properties of the gamma distribution that as m/λ increases, it converges to the normal distribution, therefore there is no contradiction between this result and asymptotic relation (26).

Asymptotic Cumulative Shocks Modeling

In many applications, the number of shocks in the time interval of interest is large, which makes it possible to apply asymptotic methods. As already stated, exact nonasymptotic results in shocks modeling exist only for the simplified settings. Therefore, the importance of asymptotic methods should not be underestimated, as they usually present, in practice, simple and convenient relations.

The results of this section are based on the methodology of renewal processes, marked point processes, and random walks, but technical details will be skipped (the reader is referred to the corresponding references) and the focus will be on final asymptotic relations. As in the section titled “‘Direct’ Models of Degradation”, consider a family of nonnegative, i.i.d., two-dimensional random vectors $\{(X_i, Y_i), i \geq 0\}$, $X_0 = 0, Y_0 = 0$, where $\sum_{i=1}^n X_i$ is the accumulated damage after n shocks and $Y_i, i = 1, 2, \dots$ the sequence of the i.i.d. interarrival times for the renewal process. Recall that the renewal process is defined by the sequence of the i.i.d. interarrival times. Specifically, when these times are exponentially distributed, the renewal process reduces to the Poisson one. We shall assume for simplicity that X and Y are independent, although the case of dependent variables can be also considered [7]. Let $0 < E[X], E[Y] < \infty, 0 <$

$\text{Var}(X), \text{Var}(Y) < \infty$. It follows immediately from equation (19) and the elementary renewal theorem [3] that

$$\lim_{t \rightarrow \infty} \frac{E[W_t]}{t} = \lim_{t \rightarrow \infty} \frac{E[N_t]E[X]}{t} = \frac{E[X]}{E[Y]} \quad (30)$$

The corresponding central limit theorem can be proved using the theory of stopped random walks [7]

$$\frac{W_t - (E[X]/E[Y])t}{(E[Y])^{-3/2}\sigma t^{1/2}} \longrightarrow N(0, 1), \quad t \rightarrow \infty \quad (31)$$

where $\sigma = \sqrt{\text{Var}(E[Y]X - E[X]Y)}$.

The important relation (31) means that for large t the random variable W_t is approximately normally distributed with expected value $(E[X]/E[Y])t$ and variance $(E[Y])^{-3}\sigma^2(E[X])^2t$. Therefore, only $E[X]$, $E[Y]$, and σ should be known for the corresponding asymptotic analysis, which is very convenient, in practice.

Similar to equation (30),

$$\lim_{t \rightarrow \infty} \frac{E[T_D]}{D} = \lim_{D \rightarrow \infty} \frac{E[N_D]E[Y]}{D} = \frac{E[Y]}{E[X]} \quad (32)$$

where N_D denotes the random number of shocks to reach the cumulative value D . Equation (31) can be now written for the distribution of the first passage time T_D [7] as under:

$$\frac{T_D - (E[Y]/E[X])D}{(E[X])^{-3/2}\sigma D^{1/2}} \longrightarrow N(0, 1), \quad D \longrightarrow \infty \quad (33)$$

This equation implies that for large threshold D the random variable T_D has approximately a normal distribution with expected value $(E[Y]/E[X])D$ and variance $(E[X])^{-3}\sigma^2D$. Therefore, the results of this section can be easily and effectively used in safety and reliability analysis.

We apply equation (33) to the setting of Example 3 [3]. We additionally assume that $E[Y] = 6$ and $\sqrt{\text{Var}(Y)} = 2$ (in hours). The parameter σ in this case is 0.0024916. The question is, what is the probability that a nominal value of 1000 is only exceeded after 600 h? Applying equation (33) we have

$$\begin{aligned} \Pr(T_{1000} \geq 600) \\ &= 1 - \Phi \left(\frac{600 - \frac{6}{9.2} 10^3}{(9.2)^{-3/2} \cdot 24.916 \cdot \sqrt{0.1}} \right) \\ &= 1 - \Phi(-1.848) = 0.967 \end{aligned} \quad (34)$$

Noncumulative Shock Models

Let the shocks occur in accordance with a renewal process or a nonhomogeneous Poisson process. Each shock, independently of the previous history, leads to a system failure with probability θ , and is survived with a complementary probability $\bar{\theta}$. Assume that a shock is the only cause of the system's failure. We see that there is no accumulation of damage and the fatal damage can be a consequence of a single shock. A number of problems in reliability, risk, and safety analysis can be interpreted by means of this model. Sometimes, this setting is referred to as an *extreme shock model* [7].

Let, as previously, $\{Y_i\}_{i \geq 1}$ denote the sequence of i.i.d. lifetime random variables with a generic cdf $F(t)$, and let B be a geometric variable with parameter θ and $S(t, \theta) \equiv \Pr\{T \leq t\}$ denote the probability of the system's failure in $[0, t]$ (the cdf of T). Straightforward reasoning results in the infinite sum for this probability of interest

$$S(t, \theta) = \theta \sum_{k=1}^{\infty} \bar{\theta}^{k-1} F^{(k)}(t) \quad (35)$$

where $F^{(k)}(t)$ is a k -fold convolution of $F(t)$ with itself.

Special complicated numerical methods should be used for obtaining $S(t, \theta)$ in the form of equation (35). Hence, it is very important, for practical assessment of safety or reliability, to obtain simple approximations and bounds. It is well known (see, e.g. [17]) that, as $\theta \rightarrow 0$, the following convergence in distribution takes place:

$$S(t, \theta) \rightarrow 1 - \exp \left\{ -\frac{\theta t}{E[Y]} \right\}, \quad \forall t \in (0, \infty) \quad (36)$$

Specifically, when $\{Y_i\}_{i \geq 1}$ is a sequence of exponentially distributed random variables with constant failure rate λ , equation (36) turns into an exact relation

$$S(t, \theta) = 1 - \exp\{-\theta\lambda t\} \quad (37)$$

Thus, equation (36) constitutes a very simple asymptotic exponential approximation. In practice, the parameter θ is not usually sufficiently small for using approximation effectively (equation (36)), therefore the corresponding bounds for $S(t, \theta)$ can be very helpful.

The simplest (but rather crude) bound that is useful, in practice, for the survival function (see **Insurance Applications of Life Tables; Risk Classification/Life**) $\bar{S}(t, \theta) \equiv 1 - S(t, \theta)$ can be based on the following identity:

$$E[\bar{\theta}^{N_t}] = \sum_{k=0}^{\infty} \bar{\theta}^k (F^{(k)}(t) - F^{(k+1)}(t)) = \bar{S}(t, \theta) \quad (38)$$

Finally, using Jensen's inequality, we have

$$\bar{S}(t, \theta) = E[\bar{\theta}^{N_t}] \geq \bar{\theta}^{E[N_t]} \quad (39)$$

Let $\{Y_n\}_{n \geq 1}$, now, be a sequence of interarrival times for the nonhomogeneous Poisson process with rate $\lambda(t)$. Thus, given a realization $\sum_1^k Y_i = y$, $k = 1, 2, \dots$, the interarrival time Y_{k+1} has the following distribution:

$$F(t|y) = 1 - \frac{\bar{F}(t+y)}{\bar{F}(y)} \quad (40)$$

Similar to equation (35), consider the following geometric-type sum:

$$S_P(t, \theta) = \Pr\{T \leq t\} = \theta \sum_{k=1}^{\infty} \bar{\theta}^{k-1} F_P^{(k)}(t) \quad (41)$$

where $F_P^{(k)}$ denotes the cdf of $\sum_1^k Y_i$ and the subindex "P" stands for "Poisson".

It turns out [18], that, in this case, the cdf $S_P(t, \theta)$ can be obtained exactly in a very simple form, even for the time dependent $\theta(t)$. Let $\lambda(t, H_t)$ denote the complete intensity function for some general orderly point process [19], where H_t is a history of the process up to time t . Conditional rate of termination $\lambda_c(t, H_t)$, which is, in fact, a conditional failure rate for the lifetime T , can be defined *via* the following equation:

$$\begin{aligned} \lambda_c(t, H_t) dt &= \Pr\{T \in [t, t + dt) | H_t, T(H_t) \geq t\} \\ &= \theta(t) \lambda(t, H_t) dt \end{aligned} \quad (42)$$

The condition $T(H_t) \geq t$ means that all shocks in $[0, t)$ were survived. At the same time, it is clear that for the specific case of the Poisson process of shocks equation (42) becomes

$$\lambda_c(t, H_t) = \theta(t) \lambda(t) \quad (43)$$

and eventually, we arrive at a simple exponential representation.

$$S_P(t, \theta(t)) = 1 - \exp \left\{ - \int_0^t \theta(u) \lambda(u) du \right\} \quad (44)$$

Thus $\theta(t) \lambda(t) \equiv r(t)$ is a failure rate for our lifetime T , in this setting.

Weaker Criteria of Failures

In the previous section, the system could be killed by a shock, and it was assumed to be "as good as old", if a shock was survived. Assume that we are looking now at the process of nonkilling shocks, but a failure of a system can still occur when the shocks are "too close" and the system had not recovered from the consequences of a previous shock. Therefore, the time for recovering should be taken into account. It is natural to assume that it is a random variable τ with a cdf $R(t)$. Thus, if the shock occurs while the system still had not recovered from the previous one, then a failure (disaster, catastrophe) occurs. As previously, for the Poisson process of shocks with rate $\lambda(t)$, we want to obtain the probability of a failure-free performance in $[0, t)$, $\bar{S}_P(t)$. Consider the following integral equation for $\bar{S}_P(t)$ [20]:

$$\begin{aligned} \bar{S}_P(t) &= \exp \left\{ - \int_0^t \lambda(u) du \right\} \left(1 + \int_0^t \lambda(u) du \right) \\ &+ \int_0^t \lambda(x) \exp \left\{ - \int_0^x \lambda(u) du \right\} \left[\int_0^{t-x} \lambda(y) \right. \\ &\times \exp \left\{ - \int_0^y \lambda(u) du \right\} R(y) \hat{S}(t-x-y) dy \left. \right] dx \end{aligned} \quad (45)$$

where the first term in the right-hand side is the probability that there was not more than one shock in $[0, t)$ and the integrand defines the joint probability of the following events:

- the first shock occurred in $[x, x + dx)$
- the second shock occurred in $[x + y, x + y + dy)$
- the time between two shocks y is sufficient for recovering (probability $R(y)$)
- the system is functioning without failures in $[x + y, t)$.

By $\hat{S}(t)$ in equation (45), we denote the probability of system's functioning without failures in $[0, t)$, given that the first shock had occurred at $t = 0$. Similar to equation (45)

$$\begin{aligned} \hat{S}(t) = & \exp \left\{ - \int_0^t \lambda(u) du \right\} + \int_0^t \lambda(x) \\ & \times \exp \left\{ - \int_0^x \lambda(u) du \right\} R(x) \hat{S}(t-x) dx \end{aligned} \quad (46)$$

Simultaneous equations (45) and (46) can be solved numerically. On the other hand, for the constant failure rate λ , these equations can be easily solved *via* the Laplace transform. Obtaining the Laplace transform of $\hat{S}(t)$ for this specific case from equation (46), we finally arrive at

$$\tilde{\bar{S}}_P(s) = \frac{s[1 - \lambda \tilde{R}(s + \lambda)] - \lambda^2 \tilde{R}(s + \lambda) + 2\lambda}{(s + \lambda)^2 [1 - \lambda \tilde{R}(s + \lambda)]} \quad (47)$$

where $\tilde{\bar{S}}_P(s)$ and $\tilde{R}(s)$ denote Laplace transforms of $\bar{S}_P(t)$ and $R(t)$, respectively.

Example 5 Exponentially distributed $\tau : R(t) = 1 - \exp\{-\mu t\}$. Then

$$\tilde{R}(s + \lambda) = \frac{\mu}{(s + \lambda)(s + \lambda + \mu)} \quad (48)$$

and

$$\tilde{\bar{S}}_P(s) = \frac{s + 2\lambda + \mu}{s^2 + s(2\lambda + \mu) + \lambda^2} \quad (49)$$

Performing the inverse Laplace transform

$$\bar{S}_P(t) = A_1 \exp\{s_1 t\} + A_2 \exp\{s_2 t\} \quad (50)$$

where s_1, s_2 are the roots of the denominator in equation (49) and are given by

$$s_{1,2} = \frac{-(2\lambda + \mu) \pm \sqrt{(2\lambda + \mu)^2 - 4\lambda^2}}{2} \quad (51)$$

and

$$A_1 = \frac{s_1 + 2\lambda + \mu}{s_1 - s_2}, A_2 = -\frac{s_2 + 2\lambda + \mu}{s_1 - s_2} \quad (52)$$

Equation (50) gives an exact solution for $\bar{S}_P(t)$. In applications, it is convenient to use simple

approximate formulas. Consider the following reasonable assumption

$$\frac{1}{\lambda} \gg \bar{\tau} \equiv \int_0^\infty (1 - R(x)) dx \quad (53)$$

Relation (53) means that the mean interarrival time in the shock process is much larger than the mean time of recovery, and this is often the case, in practice. In the study of repairable systems, the similar case is usually called the *fast repair*. Therefore, using this assumption, equation (50) results in the following approximate relation:

$$\bar{S}_P(t) \approx \exp\{-\lambda^2 \bar{\tau} t\} \quad (54)$$

Assume now that time of recovery for the homogeneous Poisson process of shocks is a constant τ . In this case, straightforward reasoning defines the probability of survival as the following sum:

$$\bar{S}_P(t) = \exp\{-\lambda t\} \sum_{k=0}^{[t/\tau]} \frac{(\lambda(t - (k-1)\tau))^k}{k!} \quad (55)$$

where notation $[\cdot]$ means the integer part.

Another possible generalization of the shock models is to consider two independent shock processes: the process of harmful shocks with rate λ_h and the process of healing (repair) shocks with rate λ_r . A system fails if we have two successive harmful shocks, but if a harmful shock is followed by a healing one, it is not a failure. In fact, mathematically, the problem can be described by the equations similar to (45–46) and solved *via* the Laplace transform. Similar to equation (54), the fast repair approximation, in this case, is given by

$$\bar{S}_P(t) \approx \exp \left\{ - \frac{\lambda_h^2}{\lambda_h + \lambda_r} t \right\} \quad (56)$$

which is a simple and effective approximation formula.

References

- [1] Bae, S., Kuo, W. & Kvam, P. (2007). Degradation models and implied lifetime distributions, *Reliability Engineering and System Safety* **92**(5), 601–608.
- [2] Finkelstein, M.S. (2003). A model of biological aging and the shape of the observed hazard rate, *Lifetime Data Analysis* **9**, 93–109.
- [3] Beichelt, F. & Fatti, L. (2002). *Stochastic Processes and their Applications*, Taylor & Francis.

- [4] Abdel-Hameed, M. (1975). A gamma wear process, *IEEE Transactions on Reliability* **4**(2), 152–153.
- [5] Noortwijk, J., van der Weide, J., Kallen, M. & Pandey, M. (2007). Gamma processes and peaks-over-threshold distributions for time-dependent reliability, *Reliability Engineering and System Safety* **92**(12), 1651–1658.
- [6] Dufresne, F., Gerber, H. & Shiu, E. (1991). Risk theory with the gamma process, *ASTIN Bulletin* **21**(2), 177–192.
- [7] Gut, A. & Husler, J. (2005). Realistic variation of shock models, *Statistics and Probability Models* **74**, 187–204.
- [8] Barlow, R. & Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing. Probability Models*, Holt, Rinehart and Winston.
- [9] Shanthikumar, J. & Sumita, U. (1983). General shock models associated with correlated renewal sequences, *Journal of Applied Probability* **20**, 600–614.
- [10] Shanthikumar, J. & Sumita, U. (1984). Distribution properties of the system failure time in a general shock model, *Advances in Applied Probability* **16**, 363–377.
- [11] Sumita, U. & Shanthikumar, J. (1985). A class of correlated cumulative shocks models, *Advances in Applied Probability* **17**, 347–366.
- [12] Ross, S. (1996). *Stochastic Processes*, John Wiley & Sons, New York.
- [13] Rice, J. (1977). On generalized shot noise, *Advances in Applied Probability* **9**, 553–565.
- [14] Singpurwalla, N. (1995). Survival in dynamic environment, *Statistical Science* **10**, 86–103.
- [15] Papoulis, A. (1971). High density shot noise and Gaussianity, *Journal of Applied Probability* **8**, 118–127.
- [16] Lund, R., McCormic, W. & Xiao, U. (2004). Limiting properties of Poisson shot noise processes, *Journal of Applied Probability* **41**, 911–918.
- [17] Kalashnikov, V. (1997). *Geometric Sums: Bounds for Rare Events with Applications*, Kluwer Academic Publishers.
- [18] Finkelstein, M.S. (2003). Simple bounds for terminating Poisson and renewal processes, *Journal of Statistical Planning and Inference* **113**, 541–548.
- [19] Cox, D.R. & Isham, V. (1980). *Point Processes*, Chapman & Hall.
- [20] Finkelstein, M.S. & Zarudnij, V. (2000). A shock process with a non-cumulative damage, *Reliability Engineering and System Safety* **71**, 103–107.

MAXIM FINKELSTEIN

Dependent Insurance Risks

The financial consequences of insurance contracts are uncertain (risky), simply due to their contingent nature; that is, exchanges of money, such as benefit payments and sometimes premiums, are based upon unforeseen, contingent events. Practical insurance contracts are complex and span many dimensions, including the number of risks covered, types of coverages, and payment structure. Because of these many dimensions, it is helpful to understand their interrelationships and how different aspects of a risk may be dependent. Further, from an insurer's viewpoint, diversification of risks is at the heart of risk pooling, a fundamental risk management mechanism. The success of diversification depends on the dependencies among insurance risks in an insurer's portfolio.

Actuaries and other financial analysts generally use the language of probability to assess uncertainty associated with insurance contracts. This article follows this paradigm and uses the idea of random variables (or collections of random variables, stochastic processes) to represent uncertain outcomes associated with insurance contracts. Dependencies among insurance risks can naturally be represented using ideas of dependencies among random variables. This article explores these concepts in the context of actuarial and insurance risks.

What Are Insurance Risks?

Because of the many dimensions of an insurance risk, it is useful to discuss types of dependencies in terms of categories of insurance risks. As with supply and demand in economics, insurance risks are represented on an individual and a collective basis. The former is known as an *individual risk model*, a "micro" representation of an individual contract. The latter is known as a *collective risk model*, a "macro" representation of a portfolio, or collection of contracts.

It is also useful to think of insurance coverages in terms of the length of the contract. Contracts with long-term coverages include most life, annuity, and retirement systems policies. Here, the time of the benefit payment, and for some contracts the

amounts of payment, often depends on the length of the survival time of the insured. Because of the potentially long length of time between payments and the occurrence of an insured event, present values are critical for valuing long-term contracts. In contrast, most individual property/casualty, health, and group policies can be thought of as "short-term" coverages. They are characterized by the reduced role of present values. In addition, the benefit amount is typically a random variable. Its value will depend in health insurance on the services provided, and in property insurance on the extent of the property damage.

Why Is Dependence Important?

The degree of dependence impacts insurance risk measures that are used to guide enterprises that manage these risks. Two broad classes of summary measures are those that (1) quantify the *financial solvency* (see **Solvency**) of a firm and (2) those for valuation of insurance contracts (see **Fair Value of Insurance Liabilities**). The latter class is further subdivided into (a) valuation at contract initiation, the pricing function, and (b) valuation following contract initiation, the liability or reserving function. Financial solvency measures are analyzed from an insurer's viewpoint and so dependencies among portfolios of insurance risks take precedence. Epidemics, floods, and earthquakes are risks that are pandemic or widespread risks that may influence many contracts within an insurer's portfolio, thus inducing dependencies. The capital investment environment is another example of a pandemic risk where an upturn or downturn in returns from capital investments affects the value of all contracts in a portfolio simultaneously, thus inducing dependencies. From an insurer's viewpoint, relationships among different categories of insurance, also known as *lines of business*, are also critical for establishing capital adequacy standards. Many lines of business are positively related, often due to a common economic environment that they share. Insurers often seek lines of business that are negatively related, which is useful for diversification of risks. For example, improving mortality may induce lower claims for life insurance and more claims for annuities.

Individual insurance contracts span many dimensions, they may: (a) cover more than a single person, structure or firm, (b) provide benefits that vary over

different types of claims, (c) provide benefits that vary by the cause of a claim, and (d) provide benefit payments immediately as well as over a longer period of time. For the first dimension, in annuities and life insurance, benefit payments are determined based on the survival of two or more covered lives. Intuitively, pairs of individuals exhibit dependence in mortality because they share common risk factors. These factors may be purely genetic, as in the case of a grandparent and grandchild, or environmental, as in the case of a married couple. In automobile insurance, one may vary the amount of coverage based on the type of claim such as damage to the automobile or injury to a third party. It seems intuitively plausible that a more serious accident will induce larger claims for both types of coverages, inducing a dependency among claim types. In pensions associated with an individual's employment, a person may leave employment owing to disability, voluntary or involuntary separation from work or permanent retirement from the labor force. A poor economy can increase each of these three causes of leaving employment, thus inducing a dependency among the causes of exiting. This is important from an insurer's viewpoint because each cause will mean a different benefit payment.

Statistical Models of Dependence

There are several approaches for quantifying relationships among random variables. The goal of each method is to understand the joint distribution function of random variables; distribution functions provide all the details on the possible outcomes of random variables, both in isolation of one another and as they occur jointly. We categorize these methods as (a) a "variables-in-common", (b) a conditioning, and (c) a direct approach. As we will see, there is substantial overlap in these approaches in that multivariate distributions for modeling dependencies can be derived using more than one approach.

In insurance, the "variables-in-common" technique is used extensively for generating dependent random variables. With this technique, a common element serves to induce dependencies among several random variables. One way to induce a dependency is through a latent (unobserved) variable that is common to all random variables. For example, we might assume that Y_1 and Y_2 are two independent (underlying) lifetimes, that Z represents the

time until a common disaster (such as an epidemic, earthquake, or flood) and that $X_1 = \min(Y_1, Z)$ and $X_2 = \min(Y_2, Z)$ are observed lifetimes. The common latent variable Z induces a dependency between X_1 and X_2 . For example, for further discussions of this "joint-life" mortality structure and its relevance to life insurance pricing, see [1]. More generally, one might assume that the distribution functions of X_1 and X_2 each depend on a latent variable Z and that, conditional on Z , the distribution functions are independent. For example, given Z , assume that X_1 and X_2 have independent exponential distributions with mean $1/Z$. Then, if Z has a gamma distribution, it is straightforward to show (unconditionally) that X_1 and X_2 have a bivariate Pareto distribution.

Another common method of modeling dependent distributions is to use conditioning arguments. That is, if one is interested in the joint distribution of X_1 and X_2 , one specifies the marginal distribution of X_1 and the conditional distribution of X_2 , given X_1 . Regression analysis can be viewed as a type of modeling based on conditional distributions. In regression, it is customary to identify a variable of interest, such as claim size, and model its distribution conditional on other variables, such as policyholder characteristics including age, gender, driving experience, and so forth. Based on the number of applications, regression analysis is probably the most widely used applied statistical method in the social sciences, including insurance. As another example, many models in time series analysis rely on conditional distributions. For example, we could think of X_2 as the current value of a series and X_1 as the prior value of the same series. Then, the distribution of X_2 given X_1 is a Markovian model of a time series.

A third approach is to directly model the joint distribution of a collection of random variables. In terms of prevalence of applications, the normal distribution is by far the most widely used multivariate distribution. Other multivariate distributions have found their way into applications but are used much less frequently. These include the multivariate t -, Pareto, and Poisson distributions.

In insurance, it has become common to use "*copula*" models (see **Copulas and Other Measures of Dependency**) to represent dependence. To define a copula, begin by considering " p " uniform (on the unit interval) random variables, $U_1, \dots, U_p$. Assume that $U_1, \dots, U_p$ may be related through their joint

distribution function

$$\Pr(U_1 \leq u_1, \dots, U_p \leq u_p) = C(u_1, \dots, u_p) \quad (1)$$

Here, we call the function C a *copula*. Further, U is a (*ex ante*) uniform random variable whereas, u is the corresponding (*ex post*) realization. To complete the construction, we may select arbitrary marginal distribution functions $F_1(x_1), \dots, F_p(x_p)$. Then, the function

$$C(F_1(x_1), \dots, F_p(x_p)) = F(x_1, \dots, x_p) \quad (2)$$

defines a multivariate distribution function, evaluated at $(x_1, \dots, x_p)$, with marginal distributions $F_1, \dots, F_p$. Compared to the classic multivariate distribution approach, the primary advantage of using copulas is that one can use different marginal distributions. Thus, for example, one might use a short- or medium-tail distribution such as a gamma for the first marginal distribution for automobile damage. A longer tail distribution, such as a Pareto, could be used for the second distribution, which represents injury claims. See, for example [2] or [3] for more information on copulas.

Actuarial Applications of Dependence Modeling

Having described types of risks and their importance in practice as well as approaches to modeling dependence, we are now in a position to describe specific models of dependence used in actuarial and insurance models. This list is necessarily incomplete and is the subject of vigorous ongoing research.

Tabular Summaries of Insurance Risks. It is customary in mortality to present tabular information by age and gender, as well as by the time since purchasing insurance (known as a *select* and *ultimate decomposition*), size of the insurance benefit, and so forth. As another example, rating tables in automobile insurance are typically presented based on age, gender, type of driving (work *versus* pleasure), average amount of driving, and so forth. These traditional risk tables summarize an average risk by one or more factors of interest. Although this approach is limited (only averages are portrayed, only information by factors portrayed is available), it also clearly indicates a long-standing interest in actuarial studies in dependence modeling.

Linear Regression Studies. Risk tables are often created using more formal statistical methods such as linear regression. Linear regression is a widely used tool in statistical modeling. It provides a disciplined approach for deciding which combinations of factors are important determinants of a risk. The goal of this methodology is to estimate a regression function, which is a conditional mean of the risk given the rating factors. Thus, although certainly more flexible than simple risk tables, it focuses on only one aspect of a conditional distribution, the mean.

Other (Nonlinear) Regression Methods. When the underlying risk is not approximately normally distributed, other types of nonlinear regression models can be used to study insurance data. In *mortality studies*, it is customary to use survival models that incorporate special long-tail and censored features of lifetime data. In studying frequencies of accident claims, count regressions are employed, typically using Bernoulli, Poisson, and negative binomial distributions. For studying claim severities, medium-tail distributions such as the gamma are employed. For the analyst, a convenient fact is that Bernoulli, Poisson, negative binomial, and gamma regressions are special cases of the generalized linear model family of regression routines (see, for example [4]). For univariate risks that appear over time, such as interest rates and inflation, time series regression models are used to forecast future values and for *stochastic control* in insurance (see **Stochastic Control for Insurance Companies**). For risks that are observed both over time and in the cross section, longitudinal/panel data models are of interest. For example, we might have workers' compensation claims from a cross section of small firms, observed over 5–10 years. In this regard, panel data methods have been shown to be desirable statistical techniques as well as produce desirable *credibility estimates*. Panel data methods are often used to study insurance corporate behavior, by examining cross sections of firms over time.

Dependence between Frequency and Severity. In regression studies, one focuses on a single risk as the primary measure of interest and uses other factors to explain or predict this risk. When considering the number of claims (frequency) and the amount of claims (severity), there are two measures of interest suggesting the use of a bivariate risk measure. However, the relationship between these

two measures is hierarchical, in the sense that one typically first considers the number of claims and then, conditional on the number, the claims amount is considered. This is the traditional *collective risk model* (see **Collective Risk Models**). Thus, even the dimension of the claim vector is dependent on the frequency. It is traditional to first condition on the claims number and to then assume that the claims severity marginal distribution does not depend on the frequency. Specifically, suppose that N is the claims number random variable and $C_1, \dots, C_N$ are the corresponding claim amounts. Then, one models the joint distribution of $(N, C_1, \dots, C_N)$ as the marginal frequency distribution times the conditional distribution of severity, given frequency,

$$\begin{aligned} f(N = n, C_1, \dots, C_N) &= f(N = n) \\ &\quad \times f(C_1, \dots, C_N | N = n) \\ &= f(N = n) \\ &\quad \times \prod_{i=1}^n f(C_i | N = n) \quad (3) \end{aligned}$$

It is customary to assume that the conditional distribution of severity, given frequency $f(C_i | N = n)$ does not depend on frequency.

Other Multivariate Dependencies. In other actuarial applications, there are many examples where this natural hierarchy does not exist and more general dependence structures are useful. We have already discussed the case where more than one life is involved in an insurance contract. Modeling the dependence among lives can either be done using “variables-in-common” techniques or a copula approach. Similarly, modeling dependence among causes of failure can be analyzed using these techniques. However, it is probably fair to say that dependence causes are more commonly analyzed using *multistate models in life insurance mathematics*.

Particularly in property and casualty insurance, it is common for an event to trigger several types of claims under contract. For example, in automobile insurance, an accident may cause damage to one’s own vehicle, damage to another motorist’s vehicle, as well as injury to the drivers. More severe accidents

will cause each of these three types of claims to be larger, inducing a dependency among them. Because injury types of claims tend to be longer tailed compared to vehicle damage claim types, it is natural to use a copula approach to model the dependency yet using different marginal distributions for each claim type.

When examining portfolios of insurance contracts broken down by line of business, the statistical model framework is similar. Some lines of business have relatively short tails (term *life insurance*) whereas, others have long tails (medical malpractice insurance) and thus it is desirable to allow each line of business to have its own marginal distribution. However, there are often relationships among these lines due to a common corporate structure (marketing, underwriting, and so forth) and thus it is important to understand their dependence. The copula approach is becoming the favored approach for this type of analysis.

Concluding Remarks

Concepts of dependence are critical to almost every mathematical and probabilistic model of insurance operations. In this article, we have sought to introduce the concepts that are most directly related to widely known statistical methods. Concepts of dependence also take on critical roles in other actuarial and insurance models including those of *stochastic ordering of risks*, *comonotonic risks* (see **Comonotonicity**), neural nets, fuzzy set theory, and others.

References

- [1] Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A. & Nesbitt, C.J. (1997). *Actuarial Mathematics*, Society of Actuaries, Schaumburg.
- [2] Nelsen, A.E. (1999). *An Introduction to Copulas, Lecture Notes in Statistics 139*, Springer-Verlag.
- [3] Frees, E.W. & Valdez, E. (1998). Understanding relationships using copulas, *North American Actuarial Journal* 2(1), 1–25.
- [4] Haberman, S. & Renshaw, A.E. (1996). Generalized linear models and actuarial science, *The Statistician* 45(4), 407–436.

EDWARD W. FREES

Design of Reliability Tests

What are Reliability Tests?

Reliability tests provide the theoretical and practical tools whereby it can be ascertained that the probability and capability of components, products, and systems to perform their required functions in specified environments for the desired function period without failure or the desired and specified reliability and life have indeed been designed and manufactured in them [1].

It is very essential that for any industry and technology to be competitive, in today's highly competitive world marketplace, all companies have to know the reliability of their products, have to be able to control it, and have to produce them at the optimum reliability level (*see* **Reliability Optimization; Quantitative Reliability Assessment of Electricity Supply**) that yields the minimum life-cycle cost (*see* **Repair, Inspection, and Replacement Models; Reliability Integrated Engineering Using Physics of Failure**) to the user. How to design and apply reliability tests (*see* **Reliability of Consumer Goods with "Fast Turn Around"; Hazard and Hazard Ratio**) in industry is a very fast growing and very important field in consumer and capital goods industries, in space and defense industries, and in National Aeronautics and Space Administration (NASA) and Department of Defense (DoD) agencies.

Reliability Test Objectives

The general goal of reliability tests is to determine if the performance of components, equipment, and systems, either under closely controlled and known stress conditions (*see* **History and Examples of Environmental Justice; Reliability Demonstration; Stress Screening**) in a testing laboratory or under field use conditions, with or without corrective and preventive maintenance, and with known operating procedures, is within specifications for the desired function period, and if it is not, whether it is the result of a malfunction or of a failure that requires corrective action. To achieve this goal, tests are done to determine the failure rate, the mean life, and the reliability of components, equipment, and systems and their associated confidence limits at desired confidence levels. On the

basis of the results obtained, the next step is to determine the pattern of recurring failures, the causes of failures, the underlying times-to-failure distribution, and the associated stress levels. Thus we can provide guidelines as to whether corrective actions should be taken and what these should be. Next the reliability test is repeated to provide reevaluation of the reliability wise performance of the units after corrective actions are taken to assure that these actions are the correct ones and are as effective as intended.

In some cases where high reliability is required after a certain development period, i.e., designing the spacecraft in the Apollo Project, it is required to determine the growth in the mean life and/or the reliability of units during their research, engineering, and development phase; and whether the growth rate is sufficient to meet the mean life and/or the reliability requirements of the specifications by the time the life and/or the reliability need to be demonstrated.

Another important goal of reliability tests is to provide management with the reliability test results so that the requested information is presented in a very easily understood form throughout the complete evaluation of a component, equipment, or system, to appreciate the value of the reliability testing efforts and make the right decisions for the improvement of the mean time between failures (MTBFs) (*see* **Reliability Growth Testing; Availability and Maintainability**), failure rate, and/or the reliability of company products.

Reliability Test Types

Developmental testing occurs during the early phases of the product's life cycle, usually from project inception to product design release to manufacturing. It is vital to be able to characterize the reliability of the product as it progresses through its initial design stages so that the reliability specifications will be met by the time the product is ready to be released to production. With a multitude of design stages and changes that could affect the product's reliability, it is necessary to closely monitor how the product's reliability grows and changes as the product's design matures. There are a number of different test types during this phase of a product's life cycle to provide useful reliability information [2–4]:

1. Component-level testing (see Enterprise Risk Management (ERM); Expert Judgment; No Fault Found; Probabilistic Risk Assessment)

Although component-level testing can continue throughout the development phase of a product, it is most likely to occur very early in the process. This may be owing to the unavailability of parts in the early stages of the development program. There may also be a special interest in the performance of a specific component if it has been radically redesigned or if there is a separate or individual reliability specification for that component. In many cases, component-level testing is undertaken to begin characterizing a product's reliability even though full system-level test units are unavailable or prohibitively expensive. The system-level reliability can be modeled on the basis of the configuration of the components and the result of component reliability testing, if sufficient understanding exists to characterize the interaction of the components.

2. System-level testing

Although the results of component-level tests can be used to characterize the reliability of the entire system, the ideal approach is to test the entire system, particularly if that is how the reliability is specified. Although early system-level test units may be difficult to obtain, it is advisable to perform reliability tests at the system level as early in the development process as possible. At the very least, comprehensive system-level testing should be performed immediately prior to the product's release for manufacturing in order to verify the designed-in reliability. During such system-level reliability testing, the units under test should be from a homogeneous population and should be devoted solely to the specific reliability test.

3. Environmental testing (see Statistics for Environmental Toxicity; Low-Dose Extrapolation; Meta-Analysis in Nonclinical Risk Assessment)

It may be necessary in some cases to institute a series of tests in which the system is tested at extreme environmental conditions or with other stress factors accelerated above the normal levels of use. Environmental testing is performed over a wide range of environments such as temperature, pressure, humidity, and outdoor or indoor conditions.

4. Accelerated testing (see Reliability Demonstration; Lifetime Models and Risk Assessment; Mathematics of Risk and Reliability: A Select History; Hazard and Hazard Ratio)

It is used to determine the ability of components, equipment, or systems to withstand stresses much higher than would be expected under actual operating conditions [3]. It may be that the product would not normally fail within the time constraints of the test and, in order to get meaningful data within a reasonable time, the stress factors must be accelerated. Failure rates, mean lives, or reliabilities are obtained at four or five accelerated stress levels, and are then extrapolated to those at actual operating, or derated, stress levels or conditions.

5. Manufacturing testing (see Product Risk Management: Testing and Warranties; Reliability Growth Testing)

This testing takes place after a product's design has been released for production, and generally tends to measure the manufacturing process rather than the product, under the assumption that the released product design is final and good. However, this is not necessarily the case, as postrelease design changes or feature additions are not uncommon. It is still possible to obtain useful reliability information from manufacturing testing without diluting any of the process-oriented information that these tests are designed to produce.

6. Functionality testing (see Use of Decision Support Techniques for Information System Risk Management; Stress Screening)

This type of testing usually falls under the category of operation verification. In these tests, a large proportion, if not all, of the products coming off the assembly line are put to a very short test in order to verify that they are functioning properly. In some situations, they may be run for a predetermined "burn-in" time in order to weed out those units that would have early infantile failures in the field.

7. Burn-in testing (see Burn-in Testing: Its Quantification and Applications)

It is intended to minimize, if not eliminate, defective substandard components from going into the next level of assembly [4]. A combination of high temperature, temperature cycling, and nonthermal stresses such as voltage, wattage, mechanical loads, or stresses, etc., is applied to such units as they come

out of production. Failure analysis needs to be conducted to identify the causes of defects creeping into the production of these components [5].

8. Extended postproduction testing (*see Natural Resource Management*)

This type of testing usually gets implemented toward the end or shortly after the product design is released to production. It is useful to structure these types of tests to be identical to the final reliability verification tests conducted at the end of the design phase. The purpose of these tests is to assess the effects of the production process on the reliability of the product. By replicating these tests with actual production units, potential problems in the manufacturing process can be identified before many units are shipped.

9. Design/process change verification testing (*see Expert Judgment*)

This type of testing is similar to the extended postproduction testing in that it should closely emulate the reliability verification testing that takes place at the end of the design phase. This type of testing should occur at regular intervals during production or immediately following a postrelease design change or a change in the manufacturing process. These changes can have a potentially large effect on the reliability of the product and these tests should be adequate, in terms of duration and sample size, to detect such changes.

Reliability Test Procedures

Reliability tests may cost a lot in terms of test facilities, test fixtures, hardware to be tested, and the number of test engineers and technical personnel, as these resources should be available when needed. There are many different reliability tests for different products. Generic procedures of reliability tests are listed below:

1. Determine test requirements and objectives including detailed description of component or equipment to be tested and necessary test equipment and environment. Review existing data from previous tests and other sources to determine if any test requirements can be met without tests. Review a preliminary list of planned tests to determine whether economies can be realized by combining individual test requirements.

2. Prepare the test plan, including test goals, test types, test methods, test environment, test equipment, test period, number of test samples, test cost estimation, how to collect, record, and analyze the test data, and failure description. The test plan must be discussed and approved both by design engineers and reliability engineers.
3. The choice of the test sample size to be used is a very critical item in testing any product. Unfortunately, usually not enough units of a new product or of a remodeled one are tested. A major effort needs to be expended to obtain the right number of test units. Chapter 13 in [6, pp. 699–777] has been put together to address this subject in detail. Example 3 provides cases for the quantification of test sample sizes.
4. Build the test platform including power supply, test chamber, data recording equipment, test data input, and the devices under test.
5. Start the test, record data such as the time and the phenomena of each failure, and replace the failed units if needed.
6. End the test depending on whether the test is time terminated or failure terminated.
7. Process and analyze the collected data, and decide the best fitted distribution of the data and its parameters. Get the basic reliability information such as MTBFs, failure rate, and reliability. Calculate the reliability of the products for certain operation time and decide whether it fulfills the designed-in reliability goal.
8. Do failure modes effects and criticality analysis (FAMECA), which is a step-by-step procedure for the systematic evaluation of the severity of potential failure modes in a system [6]. It helps to identify potential failure modes for a product or process, assess the risk associated with those failure modes, rank the issues in terms of importance, and identify and carry out corrective actions to address the most serious concerns.
9. Formulate the final reliability test report including failure data analysis, reliability quantification, FAMECA results, and conclusions and recommendations to improve the product's design.

Numerical Examples

Example 1 A company has designed a new type of an engine, and now there are 20 units available for

testing to determine the MTBF and the failure rate of these engines. Design the reliability test and do the data analysis based on the test results. Assume that the times to failure of these engines follow the exponential distribution from previous reliability tests on 10 prototypes. The average life of the 10 prototypes is 500 h.

Solution to Example 1

Step 1. Select the reliability test type for this product. Since the life of these engines could be significantly longer than 500 h, and the test cost will be high if all of the 20 units are tested to failure, it is a better choice to apply a time-terminated, nonreplacement test for these 20 engines.

Step 2. Select the test time. The average life of the 10 prototypes is about 500 h; however, the new design will make the engine life longer. So here we choose 900 h for the test time.

Step 3. Build the test platform and environment. Put the 20 engines into working status at the same time.

Step 4. Test them for 900 h. Record the time to failure of each unit that failed during the test. During the test, five failures occur at the times shown in Table 1.

Step 5. Data analysis. According to [7], the equation to estimate the MTBF for the time-terminated nonreplacement test is

$$MTBF = \frac{\sum_{i=1}^r T_i + (N - r)t_d}{r} \quad (1)$$

where N = total number of units under test, r = number of failures in the test, T_i = time to failure of the i th failed unit, and t_d = chosen test duration.

Table 1 Times to failure for Example 1

i , failure number	T_i , time to failure (h)
1	380
2	418
3	543
4	631
5	852

In the example, $N = 20$, $r = 5$, $t_d = 900$, and the T_i are listed in Table 1. Put these values into equation (1) and get

$$\begin{aligned} MTBF &= \frac{(380 + 418 + 543 + 631 + 852) + (20 - 5) \times 900}{5} \\ &= \frac{1487 + 13500}{5} \\ &= \frac{14987}{5} \\ &= 2997.4 \text{ h} \end{aligned}$$

Since we assume that the times to failure of these engines follow the exponential distribution, the failure rate, λ , is obtained from

$$\lambda = \frac{1}{MTBF} = \frac{1}{2997.4} = 0.000335 \text{ fr/h} \quad (2)$$

From these results we see that the new design has a much higher MTBF than the prototypes.

Example 2 In an accelerated life test of 10 electronic chips, the failure data in Table 2, in hours, were obtained [8]:

The accelerated life test was conducted at 150 °C. The expected use operating temperature is 85 °C. The following steps are performed:

1. Use the Arrhenius model and determine the minimum life at the use temperature.
2. Use the Arrhenius model and determine the mean life at the use temperature.

Solutions to Example 2

1. From the test data, the minimum life at the accelerated temperature is

$$L_{T_A^*} = 2750 \text{ h} \quad (3)$$

The accelerated and use temperatures in Kelvin are

$$T_A^* = 150 + 273 = 423 \text{ K} \quad (4)$$

Table 2 The times to failure of 10 electronic chips of Example 2

2750	3100	3400	3800	4100
4400	4700	5100	5700	6400

and

$$T_U^* = 85 + 273 = 358 \text{ K} \quad (5)$$

respectively.

Assuming the activation energy, $E_A = 0.2 \text{ eV}$, for these units, and Boltzmann's constant $= 8.623 \times 10^{-5} \text{ eVK}^{-1}$. Substituting all of the required quantities into the Arrhenius model, we have

$$\begin{aligned} L_{T_{U-\min}^*} &= L_{T_{A-\min}^*} e^{\frac{E_A}{K} \left(\frac{1}{T_U^*} - \frac{1}{T_A^*} \right)} \\ &= (2750) e^{\frac{0.2}{8.623 \times 10^{-5}} \left(\frac{1}{358} - \frac{1}{423} \right)} \end{aligned}$$

or

$$L_{T_{U-\min}^*} = 7442.06 \text{ h} \quad (6)$$

2. From the test data, the mean life, $L_{T_{A-\min}^*}^*$, at the accelerated temperature is

$$\begin{aligned} L_{T_{A-\min}^*}^* &= (2750 + 3100 + 3400 + 3800 \\ &\quad + 4100 + 4400 + 4700 + 5100 \\ &\quad + 5700 + 6400)/10 \end{aligned}$$

or

$$L_{T_{A-\min}^*}^* = 4345 \text{ h} \quad (7)$$

Using the same $E_A = 0.2 \text{ eV}$ as before, and substituting all of the required quantities into the Arrhenius model, yields the mean life at the use temperature of 85°C , or

$$\begin{aligned} L_{T_{U-\min}^*}^* &= L_{T_{A-\min}^*}^* e^{\frac{E_A}{K} \left(\frac{1}{T_U^*} - \frac{1}{T_A^*} \right)} \\ &= (4345) e^{\frac{0.2}{8.623 \times 10^{-5}} \left(\frac{1}{358} - \frac{1}{423} \right)} \end{aligned}$$

or

$$L_{T_{U-\min}^*}^* = 11\,758.45 \text{ h} \quad (8)$$

Example 3 Let us find the sample sizes required to estimate the achievable MTBF of a product at a desired confidence level, or assurance level, so that the desired MTBF will be attained. Let us choose a confidence level of 90% with the following precisions, or errors:

- (a) $\tau = \pm 10\%$, (b) $\tau = \pm 30\%$, and (c) $\tau = \pm 50\%$

Solutions to Example 3

Here the confidence level (CL) is 90%, or $CL = 1 - \alpha = 0.90$, where α is the risk level or the probability that we will not achieve the 90% assurance that the desired MTBF has been achieved. Then, for a precision of $\alpha = 0.10$, $\alpha/2 = 0.05$, we proceed as follows:

First choose $\tau = \pm 10\%$

Quantify

$$\frac{1 + \tau}{1 - \tau} = \frac{1 + 0.10}{1 - 0.10} = \frac{1.10}{0.90} = 1.2222 \quad (9)$$

and substitute into [9]

$$N = \frac{1}{4} + \left(\frac{1}{2} \right) z_{\alpha/2}^2 \left[\frac{1}{\tau} \left(\frac{1}{\tau} + \sqrt{\frac{1}{\tau^2} - 1} \right) - \frac{1}{2} \right]$$

or

$$\begin{aligned} N &= \frac{1}{4} + \left(\frac{1}{2} \right) (1.645)^2 \\ &\quad \times \left[\frac{1}{0.10} \left(\frac{1}{0.10} + \sqrt{\frac{1}{0.10^2} - 1} \right) - \frac{1}{2} \right] \\ &= 270 \end{aligned} \quad (10)$$

which yields a test sample size of $N = 270$ identical products, which is unbelievably large.

The value of 1.645 is obtained from the area tables of the standardized normal distribution by entering them with the value of $\alpha/2 = 0.05$.

Similarly, for a precision of $\pm 30\%$, or $\tau = 0.30$, the same equation yields a sample size of 29 identical products, and for a precision of $\pm 50\%$, or $\tau = 0.50$, the same equation yields a sample size of 10. It can be seen that the sample sizes that should be used should be much larger than those used today in our industries.

References

- [1] Kececioglu, D.B. (2002). *Reliability & Life Testing Handbook*, DESTech Publications, Lancaster, PA, Vol. 1, 941 pp., pp. 1–9.
- [2] Kececioglu, D.B. (2006). *A Blueprint for Implementing A Comprehensive Reliability Engineering Program*, http://www.weibull.com/Articles/RelIntro/Reliability_Testing.htm, www.weibull.com.
- [3] Hahn, G.J. & Shapiro, S.S. (1967). *Statistical Models in Engineering*, John Wiley & Sons, New York, p. 355.

- [4] Kapur, K.C. & Lamberson, L.R. (1974). *Reliability in Engineering Design*, John Wiley & Sons, New York, p. 564.
- [5] Kececioglu, D.B. (2003). *Burn-In Testing – Its Quantification and Optimization*, DESTech Publications, Lancaster, PA, p. 699.
- [6] Andrews, J.D. & Moss, T.R. (2002). *Reliability and Risk Assessment*, ASME Press, New York, p. 75.
- [7] Kececioglu, D.B. (2002). *Reliability Engineering Handbook*, DESTech Publications, Lancaster, PA, Vol. 1, 720 pp., p. 243.
- [8] Jensen, F. & Peterson, N.E. (1982). *Burn-in, An Engineering Approach to the Design and Analysis of Burn-In Procedures*, John Wiley & Sons, p. 167.
- [9] Kececioglu, D.B. (2002). *Reliability & Life Testing Handbook*, DESTech Publications, Lancaster, PA, Vol. 2, 877 pp., pp. 722–727.

DIMITRI B. KECECIOGLU AND XIAOFANG
CHEN

Detection Limits

Detection limits (DLs) are established within a chemical laboratory to designate low-level data that cannot be distinguished from a zero concentration (*see Low-Dose Extrapolation*). Higher limits also are used to denote data with low, nonzero concentrations that are too imprecise to report as distinct numbers [1]. Data below both types of limits are reported to the customer as a “nondetect”, such as a value somewhere below 10 (<10). The measurement for the <10 could have been anywhere between 0 to just below 10. The true concentration could have been anywhere between 0 to a small distance above 10. The statistical term for data which include nondetects is “censored data”.

This entry discusses the following:

1. definitions and uses of DLs;
2. methods for using nondetect data in exposure assessments.

Definitions and Uses of Detection Limits

Several types of *reporting limits* are established in a chemical laboratory, one of which is a *detection limit*. Two conventions for calculating and naming reporting limits dominate current practice. The convention followed by many analytical laboratories in the United States uses methods and terminology developed at the US Environmental Protection Agency [1]. Limits are calculated by assuming the measurement variance remains constant from low concentrations down to a level of zero concentration. The DL is set as a multiple of several standard deviations, usually three, above zero. Measurements below the DL are considered not significantly different from zero, and reported as a nondetect. Quantitation limits (QLs) are higher reporting limits representing the level at which individual values may first be reliably quantified. Measurements above the QL are reported as individual values. Practices differ on how to report measurements between the two limits. Here, the chemist generally believes that the analyte is present in the sample, but at concentrations that cannot be quantified with full precision.

When either the DL or the QL is consistently used as the reporting limit, no bias occurs. When all

measurements below the QL (including those below the DL) are reported as <QL, no bias occurs. Or, if the DL is consistently used as the reporting limit so that measurements below the DL are reported as <DL, while measurements between the limits are reported as individual values, no bias occurs. Often the individual values are qualified with a “remark” [2, 3], to indicate that the value is estimated. However, when the two types of reporting limits are mixed, an upward bias called *insider censoring* [4] results. This occurs when data measured below the DL are reported as <QL (QL is the reporting limit for lower measurements), while values between the two limits are given individual values (DL is the reporting limit for higher measurements). Effects of this bias, which may be small but are totally avoidable, are outlined in [4].

The second convention for computing and naming reporting limits is based on terms presented by Currie [5] for radiochemical data (*see Radioactive Chemicals/Radioactive Compounds*), and later applied to water and air measurements. The limit above which observations are considered nonzero is called the *critical level* or *decision limit* rather than DL. A second, higher limit that avoids false nondetections as well as false detections is labeled the *DL*. The concept of a QL is similar in both conventions. In this second convention, the measurement variance is not considered constant between zero and the concentration standards [6], and weighted least squares is used to calculate the location of the DL.

As a third alternative, some scientists advocate the numerical reporting of all readings, including those below detection and QL. However, a measure of analytical error, i.e., 0.3 ± 0.78 , must also be included [7, 8]. If the error is ignored, imprecise measurements are given too much credibility. Values of 0.3 are considered higher than 0.2, for example, when they cannot in reality be distinguished. Specialized procedures such as weighting each measurement by the reciprocal of its variance, so that less precise measurements receive lower weight, must be employed to interpret data reported in this way. Rocke and Lorenzato [9] present one method for estimating the variance of low-level analytical measurements, so that the errors of each measurement can be reported and used in subsequent computations.

Methods for Using Nondetect Data in Exposure Assessments

Methods for censored data have been developed and used for decades in the fields of reliability analysis [10] (see **Reliability Data; Systems Reliability**) and survival analysis [11]. However, many risk analysts are not familiar with these methods, and often substitute (fabricate) values for nondetects. Substitution replaces nondetects with an artificial value not measured in the laboratory. For assessments of water chemistry, 1/2 of the DL is typically substituted, while in air chemistry assessments $1/\sqrt{2}$ times the DL is most often used [12]. Substitution produces biased estimates of mean and percentiles, and generally underestimates dispersion measures such as the standard deviation. In numerous simulation studies, substitution performed very poorly [13–15]. In addition, two sources of substitution error are not evaluated in simulations – changing DLs in response to changing laboratory protocols and to characteristics of the media being measured [16]. To illustrate the latter, interferences from other chemicals may increase the DL of the chemical of interest for some observations, but not others. Substituted values then increase for some, though true concentrations do not.

Substitution is best avoided, except for the (perhaps common) substitution of zero and DL to estimate the highest and lowest possible values for the mean. Note that extremes for other statistics, however, such as the standard deviation and *t*-tests cannot be obtained in this way [17].

There are three alternative methods to substitution:

1. Distributional methods

These methods assume that the shape of the data distribution can be modeled by a lognormal, gamma, or other standard distribution. For censored data, statistics (a mean, or regression slope) are computed using maximum likelihood estimation or MLE [12], rather than with the traditional formulae (arithmetic average, least squares, etc.).

2. Nonparametric methods

These methods do not assume a specific distributional shape, but instead use the observed percentiles of the measured data. Kaplan–Meier methods [12] form the basis for most nonparametric censored procedures.

3. Binomial methods

These methods classify data into two categories, below and above a DL. The proportion of values

above the limit are then interpreted, rather than means or medians.

These methods are applied to familiar tasks of chemical risk assessment:

Computing the Mean and Its Confidence Intervals (UCL95)

Exposure concentrations are often estimated by the mean of samples over time or space. Concentration data are most often skewed, in part because they cannot be negative. Estimating the mean for skewed data is difficult to do well. Lognormal distributions are one solution. The logarithms of data are assumed to follow a normal distribution, allowing the original values to be skewed. Frome and Wambach [18] recommend using lognormal maximum likelihood for estimating the mean, as long as data are not highly skewed. Skewness makes estimates of the standard deviation of logarithms unstable, and the standard deviation of logarithms affects estimates of both mean and standard deviation of the original units. When data are highly skewed, the nonparametric Kaplan–Meier method was recommended instead [18].

The UCL95 is a value with only a 5% chance of being less than the true population mean. It is used in probabilistic and deterministic risk assessments as an upper limit of where the population mean is located. Cox's method can estimate the UCL95 of a lognormal distribution [18]. Singh *et al.* [15] performed an exhaustive study of parametric and nonparametric methods for estimating the UCL95 for data with nondetects. Earlier guidance documents recommended substitution of one-half the DL with fewer than 15% nondetects [19]. In contrast, Singh *et al.* [15] found that substitution provides poor coverage (percent of time the statistic equaled or exceeded the population mean) for any distribution, sample size, or percent of censoring. The nonparametric Kaplan–Meier method provided better coverage than MLE, and far better than substitution methods. To compute the UCL95 from the Kaplan–Meier mean, bootstrapping or the familiar *t*-interval formula are used, depending on data characteristics [15].

Computing Percentiles

High percentiles such as the 90th or 95th percentiles of measured concentrations are often used to

estimate the reasonable maximum exposure (RME). Percentiles are computed using measured data alone (nonparametric method) for large data sets, or along with an assumed distribution (parametric method) for more moderate sample sizes. If a single DL is used, and is low in comparison to most of the data, nondetect data will not alter nonparametric estimates of high percentile values – neither substitution nor the other, better methods are necessary. For 35% nondetects below a single DL, for example, any percentile higher than the 35th percentile will be a detected value. In contrast, parametric estimates use every observation to estimate mean and standard deviation as steps to estimating high percentiles. Parametric methods that incorporate nondetects properly (MLE) produce far better estimates than do substitution or deletion of nondetects. For data with multiple DLs, all percentiles below the highest limit are affected by nondetects. Methods such as in reference [12] or [10] should be used.

Correlations

The likelihood r and Kendall's tau correlation coefficients [12] measure the strength of associations between variables for data with one or more DLs. Likelihood r is a parametric coefficient analogous to Pearson's r , but applicable to censored data. Kendall's tau is a nonparametric correlation coefficient commonly used in trend analysis. With both coefficients, nondetects are represented as "below the DL" without requiring a substituted value.

Tau is computed by determining whether y varies randomly or in concert with changes in x [12]. For example, changes in y from <1 to 3 are increases, as are changes from <3 to 10, so that single and multiple DLs can be incorporated by tau. Comparisons between <1 and <3 , or between <3 and 2, are considered ties because the direction of change cannot be determined.

Both of these correlation coefficients vastly outperform Pearson's r , which, in order to be computed, requires a substituted value for nondetects (surely to be incorrect, as is therefore the coefficient) [17]. Tau is invariant to monotonic transformations, while likelihood r is not.

Regression with Covariates

Regression models are used to either estimate a y variable from one or more covariates, or to better

understand the magnitude of the effect of a covariate as expressed by its slope coefficient. Maximum likelihood regression (also called *failure-time regression*) can be computed for data with one or more DLs, with assumptions and tests for significance similar to ordinary regression [12]. Likelihood ratio tests determine whether individual covariates should be included in the regression model. The likelihood r -squared reports a measure of the quality of the regression [11]. One limitation of MLE regression is that no censoring is allowed in the x variables, but only in the y variable [12].

Nonparametric regression is also available when there is only one x variable. The procedure is based on Kendall's tau [20]. Called *ATS* or the *Akritas-Theil-Sen* method by Helsel [12], it allows censoring in both the x and y variables. The ATS slope is like a median line through the data – outliers do not have as strong an influence on the position of this line as they do with parametric regression.

Rather than predicting concentrations, logistic regression can be used to predict the probability of a binary outcome, such as above or below the DL, or above or below the RME. The probability of detecting a pesticide, for example, can be related to one or more x variables describing the intensity of its use, its transport in the environment, etc. The likelihood r -squared again measures the quality of the regression model, while likelihood ratio tests determine which covariates best predict the probability of exceeding the limit/RME. Logistic regression (*see Logistic Regression*) is most applicable when there is a high proportion of nondetects.

Availability of Software

Commercial statistics software generally includes methods for censored data, as well as logistic regression. Nonparametric procedures currently expect censored data to be "greater than" values rather than nondetects, so data with nondetects must be transformed prior to using the nonparametric routines [12]. Maximum likelihood methods can incorporate nondetects without any modification.

The R statistical software system is freely available for Windows, Macintosh and Linux operating systems at the comprehensive R archive network (CRAN, <http://www.r-project.org>). A contributed package named *NADA for R* is available there which

implements the methods for censored data found in Helsel [12].

Macros for Minitab statistical software to implement methods for nondetect data are available at <http://www.practicalstats.com/nada>. ATS regression methods, nonparametric and parametric hypothesis testing procedures, and computing descriptive statistics are among those listed.

SPLIDA for S-Plus implements routines for reliability analysis discussed in reference [10] for S-Plus statistical software. These routines expect censored values to be “greater than” values, and therefore do not explicitly incorporate nondetects. But the methods could easily be adapted by first employing the “flipping” transformation [12].

References

- [1] USEPA (2003). *Technical Support Document for the Assessment of Detection and Quantitation Approaches*, EPA-821-R-03-005, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC.
- [2] Oblinger-Childress, C.J., Foreman, W.T., Connor, B.F. & Maloney, T.J. (1999). *New Reporting Procedures Based on Long-Term Method Detection Levels and Some Considerations for Interpretations of Water-Quality Data Provided by the U.S. Geological Survey National Water Quality Laboratory*, USGS Open-File Report 99–193, U.S. Geological Survey, Reston.
- [3] USEPA (1989). *Risk Assessment Guidance for Superfund (RAGS), Volume I: Human Health Evaluation Manual (Part A)*, EPA/540/1-89/002, Office of Solid Waste, U.S. Environmental Protection Agency, Washington, DC.
- [4] Helsel, D.R. (2005). Insider censoring: distortion of data with nondetects, *Human and Ecological Risk Assessment* **11**, 1127–1137.
- [5] Curie, L.A. (1968). Limits for qualitative detection and quantitative determination, *Analytical Chemistry* **48**, 586–593.
- [6] Gibbons, R.D. & Coleman, D.E. (2001). *Statistical Methods for Detection and Quantification of Environmental Contamination*, John Wiley & Sons, New York.
- [7] Porter, P.S., Ward, R.C. & Bell, H.F. (1988). The detection limit, *Environmental Science and Technology* **22**, 856–861.
- [8] Currie, L.A. (1995). Nomenclature in evaluation of analytical methods including detection and quantification capabilities, *Pure and Applied Chemistry* **67**, 1699–1723.
- [9] Rocke, D.M. & Lorenzato, S. (1995). A two-component model for measurement error in analytical chemistry, *Technometrics* **37**, 176–184.
- [10] Meeker, W.O. & Escobar, L.A. (1998). *Statistical Methods for Reliability Data*, John Wiley & Sons, New York.
- [11] Klein, J.P. & Moeschberger, M.L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, New York.
- [12] Helsel, D.R. (2005). *Nondetects and Data Analysis: Statistics for Censored Environmental Data*, John Wiley & Sons, New York.
- [13] Gilliom, R.J. & Helsel, D.R. (1986). Estimation of distributional parameters for censored trace-level water quality data, *Water Resources Research* **22**, 135–146.
- [14] Lubin, J.H., Colt, J.S., Camann, D., Davis, S., Cernhan, J., Severson, R.K., Bernstein L. & Hartge, P. (2004). Epidemiologic evaluation of measurement data in the presence of detection limits, *Environmental Health Perspectives* **112**, 1691–1696.
- [15] Singh, A., Maichle, R. & Lee, S.E. (2006). *On the Computation of a 95% Upper Confidence Limit of the Unknown Population Mean Based Upon Data Sets with Below Detection Limit Observations*, EPA/600/R-06/022, Office of Research and Development, USEPA, Washington, DC.
- [16] Helsel, D.R. (2005). More than obvious: better methods for interpreting nondetect data, *Environmental Science and Technology* **39**, 419A–423A.
- [17] Helsel, D.R. (2006). Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it, *Chemosphere* **65**, 2434–2439.
- [18] Frome, E.L. & Wambach, P.F. (2005). *Statistical Methods and Software for the Analysis of Occupational Exposure Data with Non-Detectable Values*, Oak Ridge National Laboratory ORNL/TM-2005/52, U.S. Department of Energy, Washington, DC.
- [19] USEPA (2002). *RCRA Waste Sampling Draft Technical Guidance*, EPA-530-D-02-002, Office of Solid Waste, U.S. Environmental Protection Agency, Washington, DC.
- [20] Akritas, M.G., Murphy, S.A. & LaValley, M.P. (1995). The Theil-Sen estimator with doubly-censored data and applications to astronomy, *Journal of the American Statistical Association* **90**, 170–177.

Related Articles

Potency Estimation

Risk from Ionizing Radiation

Risk–Benefit Analysis for Environmental Applications

DENNIS R. HELSEL

Digital Governance, Hotspot Detection, and Homeland Security

In geospatial and spatiotemporal surveillance, it is important to determine whether any variation observed may reasonably be due to chance or not. This can be done using tests for spatial randomness, adjusting for the uneven geographical population density as well as for age and other known risk factors. One such test is the spatial scan statistic (see **Syndromic Surveillance**), which is used for the detection and evaluation of local clusters or hotspot areas. This method is now in common use by various governmental health agencies, including the National Institutes of Health, the Centers for Disease Control and Prevention, and the state health departments in New York, Connecticut, Texas, Washington, Maryland, California, and New Jersey. Other test statistics are more global in nature, evaluating whether there is clustering in general throughout the map, without pinpointing the specific location of high or low incidence or mortality areas.

The declared purpose of digital governance is to empower the public with information access, analysis, and policy to enable transparency, accuracy, and efficiency for societal good at large. Hotspot detection and prioritization become natural undertakings as a result of information access over space and time. Hotspot means spot that is hot, which is of special interest or concern (see **Hazardous Waste Site(s)**). Geoinformatic surveillance for spatial and temporal hotspot detection and prioritization is crucial in the twenty-first century, so also the need for geoinformatic surveillance decision support system equipped with the next generation of geographic and networked hotspot detection, prioritization, and early warning with emerging hotspots.

Scan Statistic Methodology and Technology

Two main problems arise in geographical surveillance for a spatially distributed response variable. These are (a) identification of areas having exceptionally high (or low) response and (b) determination of

whether the elevated response can be attributed to chance variation (false alarm) or is statistically significant. The spatial scan statistic [1] has become a popular method for detection and evaluation of disease clusters. In space–time, the scan statistic can provide early warning of disease outbreaks and can monitor their spatial spread.

Spatial Scan Statistic Background

The spatial scan statistic deals with the following situation. A region R of Euclidian space is tessellated or subdivided into cells labeled by the symbol a . Data is available in the form of a count Y_a (non-negative integer) on each cell a . In addition, a “size” value A_a is associated with each cell a . The cell sizes A_a are regarded as known and fixed, while the cell counts, Y_a , are random variables. In the disease setting, the response Y_a is the number of diseased individuals within the cell and the size A_a is the total number of individuals in the cell. Generally, however, the size variable is adjusted for factors such as age, gender, environmental exposures, etc., that might affect incidence of the disease. The disease rate within the cell is Y_a/A_a . The spatial scan statistic seeks to identify “hotspots” or clusters of cells that have an elevated rate compared with the rest of the region, and to evaluate the statistical significance (p value) of each identified hotspot. These goals are accomplished by setting up a formal hypothesis-testing model for a hotspot. The null hypothesis asserts that there is no hotspot, i.e., that all cells have (statistically) the same rate. The alternative for null hypothesis states that there is a cluster Z such that the rate for cells in Z is higher than for cells outside Z . An essential point is that the cluster Z is an unknown parameter that has to be estimated. Likelihood methods are employed for both estimation and significance testing. Candidate clusters for Z are referred to as zones. Ideally, maximization of the likelihood should search across all possible zones, but their number is generally too large for practical implementation. Various devices (e.g., expanding circles) are employed to reduce the list of candidate zones to manageable proportions (see Figure 1). Significance testing for the spatial scan statistic employs the likelihood ratio test; however, the standard χ -square distribution cannot be used as reference or null distribution – in part because the zonal parameter Z is discrete. Accordingly, Monte

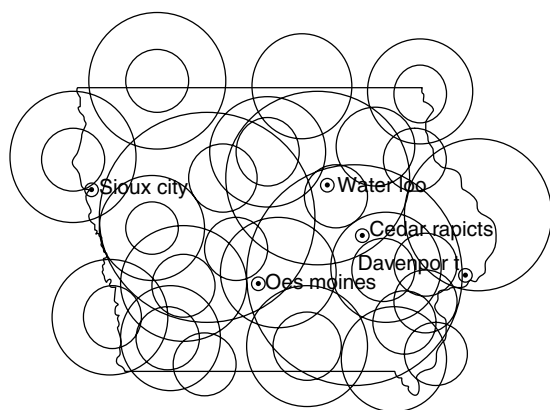


Figure 1 A small sample of the circles used

Carlo simulation is used to determine the needed null distributions.

Explication of a likelihood function requires a distributional model (response distribution) for the response Y_a in cell a . This distribution can vary from cell to cell but in a manner that is regulated by the size variable A_a . Thus, A_a enters into the parametric structure of the response distribution. In disease surveillance, response distributions are generally taken as either binomial or Poisson, leading to comparatively simple likelihood functions.

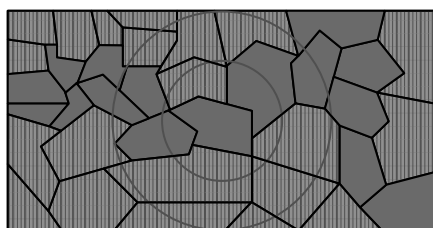
Limitations of Current Scan Statistic Methodology

Available scan statistic software suffers from several limitations (see Figure 2). First, circles have been used for the scanning window, resulting in low

power for detection of irregularly shaped clusters. Secondly, the response variable has been defined on the cells of a tessellated geographic region, preventing application to responses defined on a network (stream network, water distribution system, highway system, etc.). Thirdly, response distributions have been taken as discrete (specifically, binomial or Poisson). Finally, the traditional scan statistic returns only a point estimate for the hotspot and does not attempt to assess estimation uncertainty. All of these limitations are addressed by the innovation proposed next.

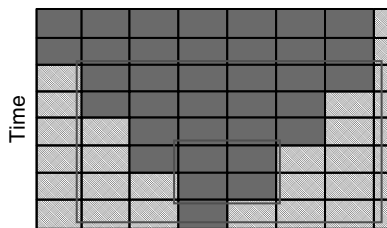
The Proposed Approach

In the approach to the scan statistic, the geometric structure that carries the numerical information is an abstract graph consisting of (a) a finite collection of vertices and (b) a finite set of edges that join certain pairs of distinct vertices. A tessellation determines such a graph: vertices are the cells of the tessellation and a pair of vertices is joined by an edge whenever the corresponding cells are adjacent. A network determines such a graph directly. Each vertex in the graph carries three items of information: (a) a size variable that is treated as known and nonrandom, (b) a response variable whose value is regarded as a realization of some probability distribution, and (c) the probability distribution itself, which is called the response distribution. Parameters of the response distribution may vary from vertex to vertex, but the mean response (i.e., expected value of the response distribution) should be proportional to the value of the size variable for that vertex. The response rate is



(a)

- Cholera outbreak along a river flood-plain
- Small circles miss much of the outbreak
 - Large circles include many unwanted cells



(b)

- Outbreak expanding in time
- Small cylinders miss much of the outbreak
 - Large cylinders include many unwanted cells

Figure 2 Scan statistic zonation for circles (a) and space–time cylinders (b)

the ratio response/size and a hotspot is a collection of vertices for which the overall response rate is unusually large.

ULS Scan Statistic

A new version of the spatial scan statistic is designed for detection of hotspots of arbitrary shapes and for data defined either on a tessellation or a network. This version looks for hotspots from among all connected components of upper level sets of the response rate and is therefore called the ULS scan statistic [2, 3]. The method is adaptive with respect to hotspot shape, since candidate hotspots have their shapes determined by the data rather than by some *a priori* prescription like circles or ellipses. This data dependence will be taken into account in the Monte Carlo simulations (see **Enterprise Risk Management (ERM); Pesticide Risk; Systems Reliability**) used to determine null distributions for hypothesis testing. It will also compare performance of the ULS scanning tool with that of the traditional spatial scan statistic. The key element here is enumeration of a searchable list of candidate zones Z . A zone is, first of all, a collection of vertices from the abstract graph; secondly, those vertices should be connected because a geographically scattered collection of vertices would not be a reasonable candidate for a “hotspot.” Even with this connectedness limitation, the number of candidate zones is too large for a maximum likelihood search in all but the smallest of graphs. The list of zones is reduced to a searchable size in the following way. The response rate at vertex a is $G_a = Y_a/A_a$. These rates determine a function $a \rightarrow G_a$ defined over the vertices in the graph. This function has only finitely many values (called levels) and each level, g , determines an upper level set U_g defined by $\{a : G_a < g\}$. Upper level sets do not have to be connected but each upper level set can be decomposed into the disjoint union of connected components. The list of candidate zones Z for the ULS scan statistic consists of all connected components of all upper level sets. This list of candidate zones is denoted by Ω_{ULS} . The zones in Ω_{ULS} are certainly plausible as potential hotspots, since they are portions of upper level sets. Their number is small enough for practical maximum likelihood search – in fact, the size of Ω_{ULS} does not exceed the number of vertices in the abstract graph (e.g., the number of cells in the tessellation). Finally, Ω_{ULS} becomes a tree under set inclusion, thus facilitating

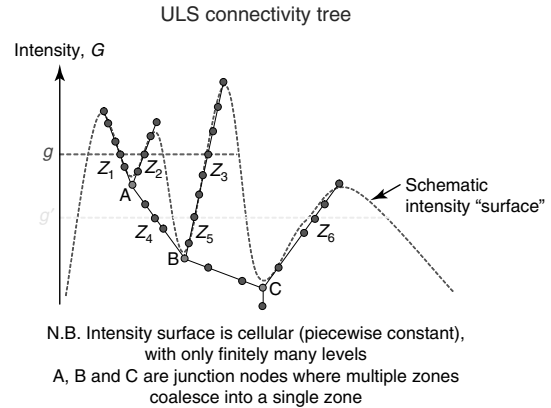


Figure 3 ULS connectivity tree

computer representation. This tree is called the *ULS tree*; its nodes are the zones $Z \in \Omega_{\text{ULS}}$ and are therefore collections of vertices from the abstract graph. Leaf nodes are (typically) singleton vertices at which the response rate is a local maximum; the root node consists of all vertices in the abstract graph (see Figure 3).

Typology of Space–Time Hotspots

Scan statistic methods extend readily to the detection of hotspots in space–time. The space–time version of the circle-based scan employs cylindrical extensions of spatial circles and cannot detect the temporal evolution of a hotspot. The space–time generalization of the ULS scan detects arbitrarily shaped hotspots in space–time. This helps classify space–time hotspots into various evolutionary types – a few of which appear on the left-hand side of Figure 4. The merging hotspot is particularly interesting because, while it comprises a connected zone in space–time, several of its time slices are spatially disconnected.

Key Applications

Broadly speaking, the proposed geosurveillance project and its forum identify several case studies around the world. This section provides a selection of illustrative applications and case studies.

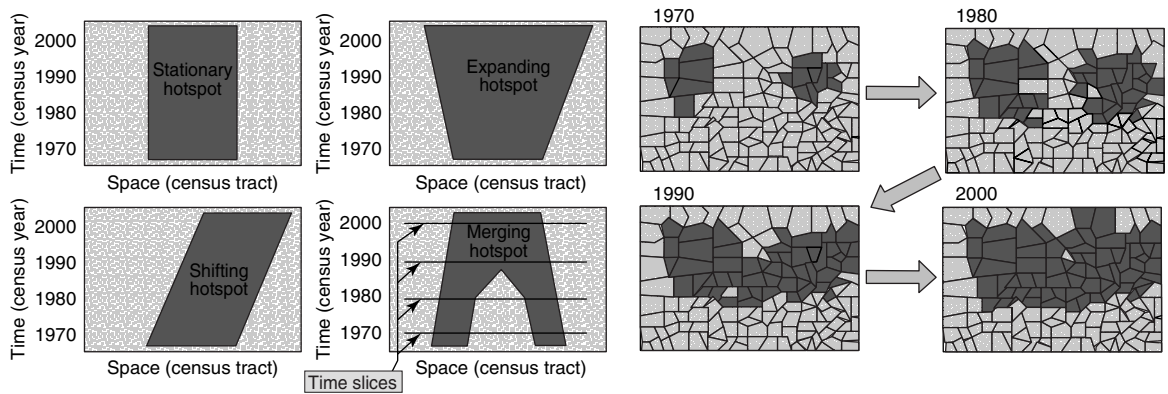


Figure 4 The four diagrams on the left depict different types of space–time hotspots. The spatial dimension is shown schematically on the horizontal axis and time on the vertical axis. The diagrams on the right show the trajectory (sequence of time slices) of a merging hotspot

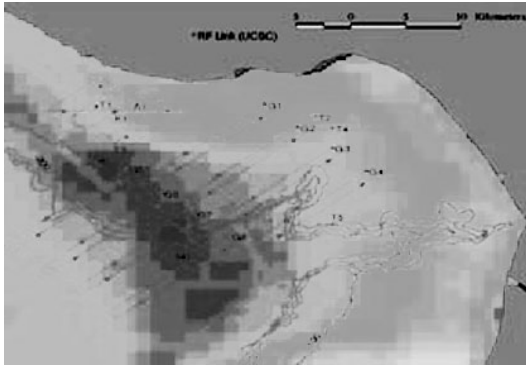


Figure 5 The ocean sampling network

Tasking of a Self-Organizing Oceanic Surveillance Mobile Sensor Network

The autonomous ocean sampling network (AOSN) simulator is used to study coordination and control strategies for high-resolution, spatiotemporally coordinated surveys of oceanographic fields such as bathymetry, temperature, and currents using autonomous, unmanned, undersea vehicles Figure 5. Currently, the network of mobile sensor platforms is autonomous and self-organizing once given high-level tasking from an external tactical coordinator. This case study proposes to use ULS scan statistic theory to identify hotspots in data gathered by the sensor network and use this information to dynamically task mobile sensor platforms so that more data can be gathered in the areas of interest. By detecting

hotspots and tasking accordingly, network resources are not wasted on mapping areas of little change. The ability of the sensor network to dynamically task its own components expands the network’s mission and increases the reactivity to changing conditions in a highly dynamic environment [4–6].

Early Detection of Biological Invasions

Intentional and unintentional introductions of nonnative exotic species have major economic and ecological impacts across the United States. A National Academy of Sciences report [7] estimates the cost of lost crops and containment measures at \$137 billion per year. Early detection of invasive weedy plants is the only cost-effective and tractable option for their containment or eradication. But systems for synthesizing on-the-ground observation, spatial data, and newly acquired remotely sensed data are lacking. We will apply the ULS scan statistic and prioritization tools to obtain more efficient surveys for invasive species and to improve the responsiveness of environmental managers to outbreaks. Japanese stiltgrass has become established in forests and waterways in the eastern United States and threatens to significantly reduce forest and riparian species diversity, and impede water flow in rivers and streams. Often locally established populations have begun to spread before those populations have been detected and likelihood for successful management is severely compromised. Coupling the data resources with the



Figure 6 Biological invasion

scan statistic represents a promising approach to preventing the transition of invasive plants from isolated established populations to spreading ones, like that depicted in Figure 6 [8, 9].

Development of Remote Sensing Methods for Crop Bioterrorism

Testimony to the House National Security Subcommittee this past fall stressed the high probability that

crop terrorism could and would occur. The logistics of ground-based monitoring are daunting, and point to a strong need to monitor US cropland by remote sensing. Remote sensing, to be timely and effective, should resolve zones of infection as small as 5 m in diameter. Improvements in the next generation of sensors will have to be made, as will the techniques for handling the huge volumes of high-resolution data. Another difficulty has been the development of high-quality signatures detectable from airborne or space-borne platforms. To date, reliable detectable signatures have been difficult to obtain. We propose to defeat this impasse by integrating biologists, engineers, and statisticians to generate high-quality hyperspectral signatures, to determine signal strength on unique portions of these signatures, and to develop goals for instrument resolution from various platforms. Specific goals are as follows: (a) develop signatures for two key pathogens using ground-based, portable hyperspectral cameras; (b) determine signal strength *versus* noise of plants infected with single or multiple pathogens and/or insects; (c) enhance hardware and processing algorithms to filter and resolve crop pathogen signals; (d) provide goals for the next generation of air- and space-borne sensors for high-threat pathogens. Figure 7 depicts the process from signature development through identification of threat areas, signal acquisition/processing to the development of an anomaly report [10, 11].

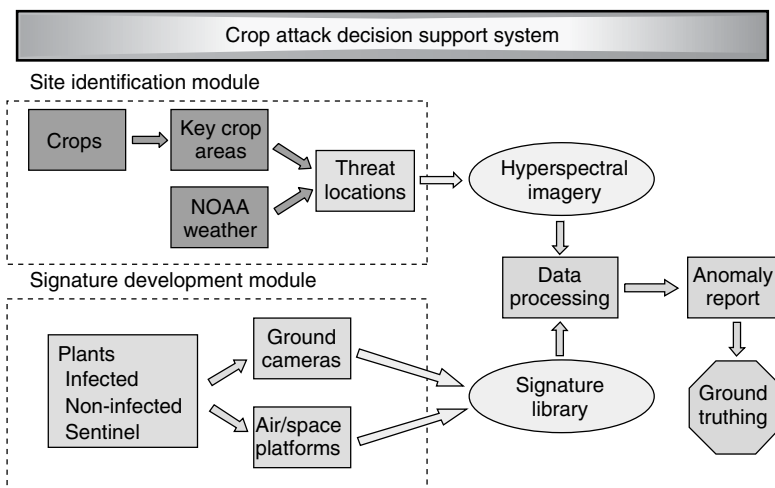


Figure 7 Crop attack decision support system

Cyber Security and Computer Network Diagnostics

Securing the nation's computer networks from cyber attacks is an important aspect of national Homeland Security. Network diagnostic tools aim at detecting security attacks on computer networks. Besides cyber security, these tools can also be used to diagnose other anomalies such as infrastructure failures and operational aberrations. Hotspot detection forms an important and integral part of these diagnostic tools for discovering correlated anomalies. It is constructive to develop a network diagnostic tool at a functional level. The goal of network-state models is to obtain the temporal characteristics of network elements such as routers, typically in terms of their physical connectivity, resource availability, occupancy distribution, blocking probability, etc. There is some prior work [12] in developing network-state models for connectivity and resource availability. Models have also been developed for studying the equilibrium behavior of multidimensional loss systems. The probabilistic finite-state automaton (PFSA) describing a network element can be obtained from the output of these state models. A time-dependent crisis index is determined for each network element, which measures their normal behavior pattern compared to crisis behavior. The crisis index is the numerical distance between the stochastic languages generated by the normal and crisis automata. Use of the variational distance between probability measures seems attractive, although other distances can also be considered. The crisis behavior can be obtained from past experience. The crisis indices over a collection of network elements are then used for hotspot detection using scan statistic methodology. These hotspots help to detect coordinated security attacks geographically spread over a network.

Drinking Water Quality and Water Utility Vulnerability

New York City has installed 892 drinking water sampling stations across the five boroughs. Each 4.5-ft high station is located outdoors and draws water from a nearby water main (Figure 8). The purpose is to monitor general water quality, detect potential health threats, and thwart bioterror activity. Sampling frequency was increased after the 9/11 attacks and, currently, about 47 000 water samples are analyzed



Figure 8 Drinking water sampling station

annually. Parameters analyzed include bacteria, chlorine, pH, inorganic and organic pollutants, color, turbidity, odor, and many others. The network version of the ULS scan statistic can provide a real-time surveillance system for detecting and evaluating water quality hotspots within the distribution system on a parameter-by-parameter basis.

An overall assessment of water quality at each sampling station taking all parameters into account is achieved by employing recent progress in multi-criterion ranking using partially ordered set (poset) prioritization [13].

Syndromic Surveillance Network and Early Warning

Emerging hotspots for disease, biological agents, or medical effects of pollution are identified through modeling events at local hospitals. A time-dependent crisis index is determined for each hospital in a network spread over a city, state or the whole country. This index measures the behavior patterns at each hospital compared to crisis behavior. The behaviors are based on series of hospital admission records containing symptoms and diagnoses. The basic components of behaviors are events, which in this case are hospital admissions. The important attributes of admissions are the information on the admission records and how frequently admissions are

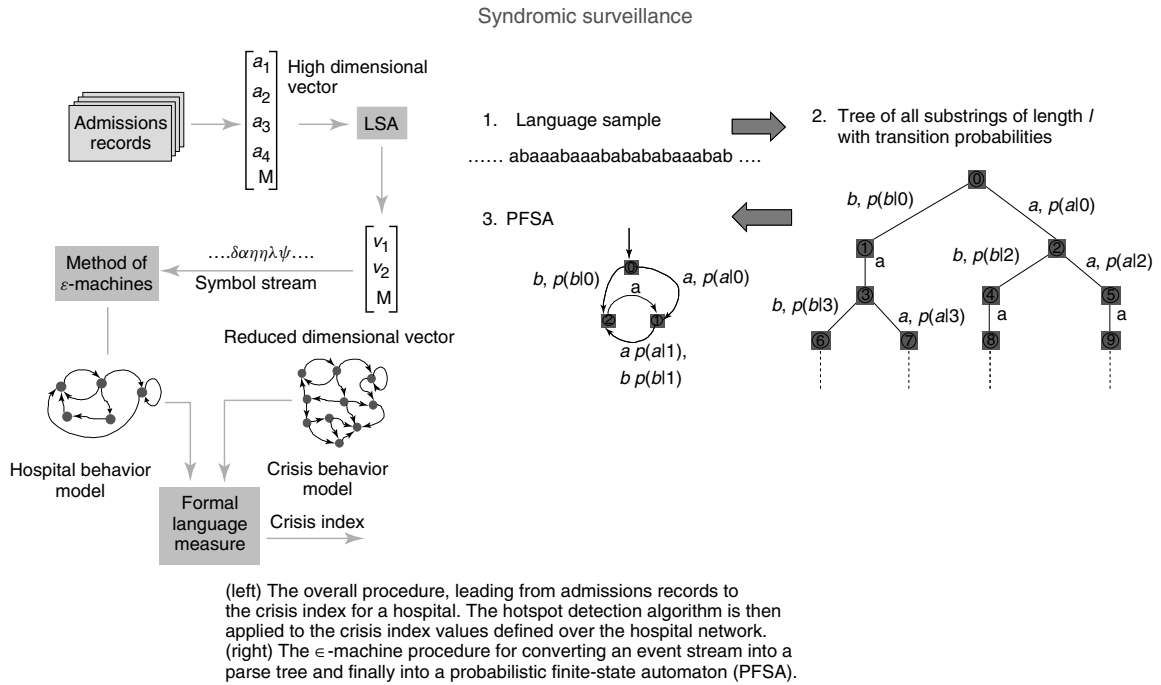


Figure 9 The overall procedure

occurring compared to normal, noncrisis behavior. The behavior stream is represented as PSFA and by the corresponding formal stochastic language. The method of epsilon machines [14, 15] is used to estimate the automaton from the current behavior stream. The variational distance between stochastic languages provides a quantitative measure of how close the current behavior is to that of a crisis. This distance is called the crisis index. The crisis index over the network of hospitals is used for hotspot detection by the ULS scan statistic (see Figure 9).

West Nile Virus: An Illustration of the Early Warning Capability of the Scan Statistic

Since the 1999 West Nile (WN) virus outbreak in New York City, health officials have been searching for an early warning system that could signal increased risk of human WN infection, and provide a basis for targeted public education and increased mosquito control. Birds and mosquitoes with laboratory evidence of WN virus preceded most human infections in 2000, but sample collection and laboratory testing are time consuming and costly. The

cylinder-based space–time scan statistic for detecting small area clustering of dead bird reports has been assessed for its utility in providing an early warning of WN virus activity (Figure 10) [16].

All unique nonpigeon dead bird reports were categorized as “cases” if occurring in the prior 7 days and “controls” if occurring during a historic baseline. The most likely cluster area was determined using the scan statistic and its statistical significance was evaluated using Monte Carlo hypothesis testing. Analyses were performed in a prospective simulation for 2000 and in real time during 2001.

For the 2000 data, dead bird clustering was found in Staten Island over 4 weeks prior to laboratory evidence of WN virus in birds and mosquitoes from this area. Real-time implementation in 2001 led to intensified larval control in eastern Queens, over 3 weeks prior to laboratory confirmation from this cluster area. Dead bird clusters were identified around the residence of five of seven human infections, 0–40 days (median 12) prior to the onset of illness and 12–45 days (median 17) prior to human diagnosis.

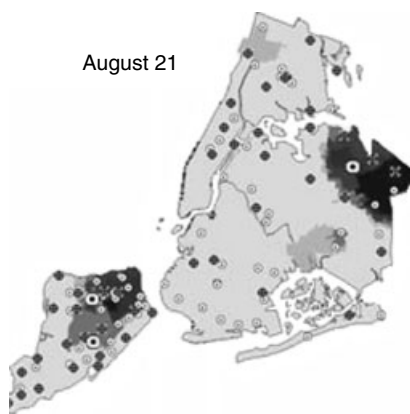


Figure 10 Event map

It has been concluded that scan statistical cluster analysis of dead bird reports may provide early warning of viral activity among birds and of subsequent human infections. Analysis with the ULS space–time scan statistic will be worthwhile. Since the latter allows for arbitrarily shaped clusters in both the spatial and temporal dimensions, there is potential for earlier detection with a reduced false alarm rate.

New York City Subway System Syndromic Surveillance

For certain problems, there is an underlying network structure on which we want to perform the cluster detection and evaluation. For example, the New York City Health Department is monitoring the New York subway system and water distribution networks for bioterrorism attacks. In such a scenario, a circular scan statistic is not useful as two individuals close to each other in Euclidian distance may be very far from each other along the network. However, the ULS methods will be employed for the detection and evaluation of clusters on a predefined network. The essentially linear structure of these networks, compared to tessellation-derived networks, is expected to have a major impact on the form of the null distributions and their parametric approximations.

The New York City Department of Health (DOH) and Metropolitan Transportation Authority (MTA) began monitoring subway worker absenteeism in October 2001 as one of several surveillance systems for the early detection of disease outbreaks. Each day the MTA transmits an electronic line list of workers absent the previous day, including work location and

reason for absence. DOH epidemiologists currently monitor temporal trends in absences in key syndrome categories (e.g., fever-flu or gastrointestinal illness). Analytic techniques are needed for detecting hotspots within the subway network.

Looking Forward

We now live in the age of geospatial technologies. The age old geographic issues of societal importance are now thinkable and analyzable. The geography of disease is now just as doable as genetics of disease, for example. And it is possible to pursue it in an intelligent manner with the rapidly advancing information technology around.

Government agencies continue to require meaningful summaries of georeferenced data to support policies and decisions involving geographic assessments and resource allocations, so also the public with initiatives of digital governance in the country and around the world.

This article briefly describes a prototype hotspot detection system for hotspot delineation and provides a variety of case studies and illustrative applications of societal importance for homeland security.

Surveillance geoinformatics of hotspot detection and prioritization is a critical need of the twenty-first century. Next generation decision support system within this context is crucial. It will be productive to build on the present effort in the directions of prototype and user-friendly methods, tools, and software, and also in the directions of thematic groups, working groups, and case studies important at various scales and levels. The authors have such a continuation effort in progress within the context of digital governance with National Science Foundation (NSF) support and would welcome interested readers to join in this collaborative initiative.

Acknowledgment

This material is based upon work supported by (i) the National Science Foundation under Grant No. 0307010, and (ii) the United States Environmental Protection Agency under Grant Nos CR-83059301 and R-828684-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the agencies.

References

- [1] Kulldorff, M. (1997). A spatial scan statistic, *Communications in Statistics: Theory and Methods* **26**, 1481–1496.
- [2] Patil, G.P. & Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots, *Environmental and Ecological Statistics* **11**, 183–197.
- [3] Patil, G.P., Duczmal, L., Haran, M. & Patankar, P. (2006). On PULSE: the progressive upper level set scan statistic system for geospatial and spatiotemporal hotspot detection, *The 7th Annual International Conference on Digital Government Research*, San Diego, May 21–24.
- [4] Eberbach, E. & Phoha, S. (1999). SAMON: communication cooperation and learning of mobile autonomous robotic agents, *Proceedings of the 11th International Conference on Tools and Artificial Intelligence*, Chicago November 9–11, 1999.
- [5] Phoha, S., Peluso, E. & Culver, R.L. (2002). A high fidelity ocean sampling mobile network (SAMON) simulator, invited paper *IEEE Journal of Oceanic Engineering* **26**(4), 646–653. Special Issue on Autonomous Ocean Sampling Networks.
- [6] Phoha, S., Peluso, E., Stadter, P., Stover, J. & Gibson, R. (1997). A mobile distributed network of autonomous undersea vehicles, *Proceedings of the 24th Annual Symposium and Exhibition of the Association for Unmanned Vehicle Systems International*, Baltimore, June 3–6, 1997.
- [7] National Academy of Sciences (2002). *Predicting Invasions of Nonindigenous Plants and Plant Pests*, National Academy Press, Washington, DC.
- [8] Mortensen, D.A., Bastiaans, L. & Sattin, M. (2000). The role of ecology in developing weed management systems: an outlook, *Weed Research* **40**, 49–62.
- [9] Mortensen, D.A., Dieleman, J.A. & Williams, M.M. (2002). Using remote sensing in integrated weed management: what do we need to see? *Agronomy Journal* (in press).
- [10] Backman, P.A. & Jacobi, J.C. (1997). Developing thresholds for plant disease management, in *Economic Thresholds for Integrated Pest Management*, L. Higley & L. Pedigo, eds, University of Nebraska Press, pp. 114–127.
- [11] Wheelis, M., Casagrande, R. & Madden, L.V. (2002). Biological attack on agriculture: low-tech high-impact bioterrorism, *BioScience* **52**, 569–576.
- [12] Ghosh, D. & Acharya, R. (2001). A probabilistic scheme for hierarchical routing, *Proceedings of Ninth IEEE International Conference on Networks*, IEEE, pp. 416–421.
- [13] Patil, G.P. & Taillie, C. (2004). Multiple indicators, partially ordered sets, and linear extensions: multi-criterion ranking and prioritization, *Environmental and Ecological Statistics* **11**, 199–228.
- [14] Shalizi, C.R., Shalizi, K.L. & Crutchfield, J.P. (2002). Pattern discovery in time series, Part I: theory, algorithm, analysis, and convergence, *Journal of Machine Learning Research* submitted.
- [15] Shalizi, K.L., Shalizi, C.R. & Crutchfield, J.P. (2002). Pattern discovery in time series, Part II: implementation, evaluation, and comparison, *Journal of Machine Learning Research* to be submitted.
- [16] Mostashari, F., Kulldorff, M., Hartman, J., Miller, J. & Kulasekera, V. (2003). Dead bird clusters as an early warning system for West Nile Virus activity, *Emerging Infectious Diseases* **9**(6), 641–646.

GANAPATHI P. PATIL, RAJ ACHARYA AND
SHASHI PHOHA

Dioxin Risk

“Dioxin” refers to a large group of structurally related chemicals called *polychlorinated dibenzo-*p*-dioxins*, although it is sometimes used to refer specifically to one such substance, 2,3,7,8-tetrachloro dibenzo-*p*-dioxin (TCDD). Dioxins comprise two benzene rings connected by two oxygen atoms and contain four to eight chlorine atoms. Dioxins are unintentional byproducts of combustion, both anthropogenic (e.g., municipal waste incineration) and naturally occurring (e.g., forest fires). Dioxins can also form as a result of certain industrial processes, such as bleached paper manufacture. They are persistent in the environment and in the body. They are both ubiquitous and persistent throughout the environment and in all our bodies. Focus on dioxin risks began in the 1940s and 1950s when workers involved in manufacturing 2,4,5-T, the herbicide that was used to make the Vietnam War-era defoliant Agent Orange, developed a serious form of acne. Later studies showed that the workers’ acne and 2,4,5-T’s ability to produce birth defects in rats could be attributed to the presence of dioxin. Many other sources and effects of dioxin have since been identified.

Exposure

The toxicity of different dioxin compounds varies considerably in laboratory animals but generally involves similar biological modes of action. As most human exposure involves complex mixtures of many dioxins, quantitative exposure estimates are referred to in terms of toxic equivalence (TEQ). Toxic equivalency factors (TEFs) for different dioxin compounds have been established by the World Health Organization (WHO) and are calculated relative to TCDD using the results of comparative toxicity studies in laboratory animals [1] (see **Cancer Risk Evaluation from Animal Studies**). So, for example, TCDD is assigned a TEF of 1; hexachlorodibenzo-*p*-dioxins are somewhat less toxic, so they are assigned a TEF of 0.1; and octachlorodibenzo-*p*-dioxin is considerably less toxic, so it is assigned a TEF of 0.0001. Concentrations of different dioxins in a mixture are multiplied by their appropriate TEFs and summed together to estimate a total TEQ for a given exposure situation. Dioxins are stored in body fat and

their half-lives vary considerably, ranging from about 6 months to 20 years [2], so body burdens of dioxins do not tend to reflect exposure quantitatively.

Since risk management aimed at reducing emissions of dioxins was initiated in the 1970s, substantial decreases in environmental levels and in body burdens have been observed. Although somewhat uncertain due to changes in analytical methodologies over time, available data on emissions, environmental and food levels, and tissue levels indicate a several-fold reduction in exposures and body burdens since 1970 [3]. The United States Environmental Protection Agency (USEPA) reported that emissions from quantified sources – such as waste incineration, pesticide manufacture, and chlorine bleaching of pulp and paper – decreased from more than 14 000 g TEQ/year in 1987 to about 1500 g TEQ/year in 2000 – a 90% decrease – and are expected to continue to decline [4]. Human body burdens of TCDD in the United States decreased 10-fold and dioxin TEQ decreased 4- to 5-fold between 1972 and 1999, which, owing to their long elimination half-lives, implies a decrease in exposure of more than 95% [3]. USEPA has estimated that more than 90% of the remaining human exposures to dioxins occur through food consumption, primarily from animal fat [5, 6]. Uncontrolled combustion, including open burning of household trash, agricultural burning, and landfill fires, are now considered to be the largest unaddressed sources of dioxin in the environment [6].

Toxicity

Studies suggest that exposure of the general population to low environmental levels of dioxins does not result in clinical evidence of disease. However, poisoning incidents involving high levels of exposure have produced several effects, especially a persistent, acne-like skin disease called *chloracne*. Studies in workers exposed to dioxins occupationally provide limited evidence of their ability to increase cancer rates. Children exposed accidentally to high levels of dioxins either before or after birth were found to have a number of developmental deficits [7].

There is a large body of literature on the toxicity of dioxins in laboratory animals, although most studies have focused on TCDD. Reviews of that literature are available [5, 8, 9]. Dioxins have been reported to produce effects such as cancer, developmental

toxicity, immunotoxicity, and reproductive toxicity in a variety of species. Developmental effects on the reproductive system are the effects that occur at the lowest doses in many laboratory animals, so those effects have been considered by several organizations to be the most sensitive endpoint of concern for evaluating human risk. The extent to which the laboratory animal data are relevant to humans is controversial, however, in part because human developmental stages do not map well to rodent development [10, 11]. Prenatal exposure of Rhesus monkeys to low doses of TCDD has been reported to induce developmental neurobehavioral effects [12, 13] including cognitive deficits and changes in social interaction. Those effects were considered the most relevant sensitive effect for evaluating human health risk by the US Agency for Toxic Substances and Disease Registry [8].

Safety

Effects of dioxins are generally attributed to their initial interaction at the cellular level with the aryl hydrocarbon hydroxylase (Ah) receptor, which regulates the activity of certain metabolic enzymes. While interaction with the Ah receptor appears to be necessary for most biological effects of dioxins, it is solely insufficient to account for all of those effects. Studies show that the binding affinity of the human Ah receptor resembles that of species that are relatively insensitive to dioxin toxicity, with data suggesting humans are likely to be one to two orders of magnitude less sensitive to the Ah-receptor-related effects of TCDD than rats and monkeys [14].

The fact that laboratory animals are more sensitive than humans and poor surrogates for predicting quantitative risks makes them good models for establishing safe levels of human exposure. Any exposure limit based on a sensitive endpoint in laboratory animals can be expected to be especially protective of human health, including the health of potentially sensitive individuals or life stages. The Joint FAO/WHO Expert Committee on Food Additives has proposed a tolerable monthly intake of 70 pg TEQ/kg body weight as being protective of human health, based on developmental effects observed in male rats exposed to TCDD during gestation [15, 16]. This value translates to a tolerable daily intake (TDI) of 2 pg TEQ/kg body weight. The UK Food Standards

Agency's Committee on Toxicity has recommended a TDI of 2 pg TEQ/kg body weight [17], as has the European Commission's Scientific Committee on Food [18] (although their specific recommendation is 14 pg TEQ/kg body weight/week), consistent with the FAO/WHO TDI. The USEPA has been debating the appropriate exposure limit for dioxins since 1985; no risk estimate or exposure limit is agreed upon. No acceptable daily intake has been established by the US Food and Drug Administration.

References

- [1] Van den Berg, M., Birnbaum, L.S., Denison, M., De Vito, M., Farland, W., Feeley, M., Fiedler, H., Hakanson, H., Hanberg, A., Haws, L., Rose, M., Safe, S., Schrenk, D., Tohyama, C., Tritscher, A., Tuomisto, J., Tysklind, M., Walker, N. & Peterson, R.E. (2006). The 2005 World Health Organization reevaluation of human and mammalian toxic equivalency factors for dioxins and dioxin-like compounds, *Toxicological Sciences* **93**, 223–241.
- [2] Flesch-Janys, D., Becher, H., Gurn, P., Jung, D., Konietzko, J., Manz, A. & Papke, O. (1996). Elimination of polychlorinated dibenzo-*p*-dioxins and dibenzofurans in occupationally exposed persons, *Journal of Toxicology and Environmental Health* **47**, 363–378.
- [3] Hays, S.M. & Aylward, L.L. (2003). Dioxin risks in perspective: past, present, and future, *Regulatory Toxicology and Pharmacology* **37**, 202–217.
- [4] US Environmental Protection Agency (2005). *The Inventory of Sources and Environmental Releases of Dioxin-Like Compounds in the United States: The Year 2000 Update*, External Review Draft. EPA/600/p-03/002A, National Center for Environmental Assessment, Office of Research and Development, Washington, DC, at <http://www.epa.gov/ncea/pdfs/dioxin/2k-update/>.
- [5] US Environmental Protection Agency (2004). *Exposure and Human Health Reassessment of 2,3,7,8-Tetrachlorodibenzo-*p*-dioxin and Related Compounds*, National Academy of Sciences Review Draft, National Center for Environmental Assessment, Office of Research and Development, Washington, DC, <http://www.epa.gov/ncea/pdfs/dioxin/nas-review/>.
- [6] Interagency Working Group on Dioxin (2006). *Questions and Answers About Dioxins*, US Environmental Protection Agency, Department of Health and Human Services, Department of Agriculture, Department of Veterans Affairs, Department of Commerce, Department of State, and the White House Office of Science and Technology Policy, Washington, DC, <http://www.cfsan.fda.gov/~lrd/dioxinqa.html#top>.
- [7] Charnley, G. & Kimbrough, R. (2006). Overview of exposure, toxicity, and risks to children from current levels of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin and related

- compounds in the USA, *Food and Chemical Toxicology* **44**, 601–615.
- [8] Agency for Toxic Substances and Disease Registry (1998). *Toxicological Profile for Chlorinated Dibenzo-*p*-dioxins*, Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta.
 - [9] AEA Technology (1999). *Compilation of EU dioxin Exposure and Health Data*, Prepared for the European Commission DG Environment, Oxfordshire, <http://europa.eu.int/comm/environment/dioxin/download.htm>.
 - [10] World Health Organization (1986). *Environmental Health Criteria 59. Principles for Evaluating Health Risks from Chemicals During Infancy and Early Childhood: The Need for a Special Approach*, Joint Sponsorship of the United Nations Environment Programme, the International Labour Organisation, and the World Health Organization, and on behalf of the Commission of the European Communities, at <http://www.inchem.org/documents/ehc/ehc/ehc59.htm>, Geneva.
 - [11] National Academy of Sciences/National Research Council (2003). *Dioxins and Dioxin-like Compounds in the Food Supply: Strategies to Decrease Exposure*, National Academies Press, Washington, DC.
 - [12] Schantz, S.L. & Bowman, R.E. (1989). Learning in monkeys exposed perinatally to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD), *Neurotoxicology and Teratology* **11**, 13–19.
 - [13] Schantz, S.L., Ferguson, S.A. & Bowman, R.E. (1992). Effects of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin on behavior of monkeys in peer groups, *Neurotoxicology and Teratology* **14**, 433–446.
 - [14] Abraham, K., Geusau, A., Tosun, Y., Helge, H., Bauer, S. & Brockmöller, J. (2002). Severe 2,3,7,8-tetra chlorodibenzo-*p*-dioxin (TCDD) intoxication: insights into the measurement of hepatic cytochrome P450 1A2 induction, *Clinical Pharmacology and Therapeutics* **72**, 163–174.
 - [15] Joint FAO/WHO Expert Committee on Food Additives (2001). 57th meeting of the joint expert committee, Rome, 5–14 June 2001. Summary and Conclusions, <http://www.who.int/pes/jecfa/jecfa.htm>.
 - [16] World Health Organization (2002). Evaluation of certain food additives and contaminants, Fifty-seventh Report of the Joint FAO/WHO Committee on Food Additives, WHO Technical Report Series 909, at <http://whqlibdoc.who.int/trs/WHO-TRS-909.pdf>, Geneva.
 - [17] UK Food Standards Agency Committee on Toxicity of Chemicals in Food, Consumer Products and the Environment (2001). *Statement on the Tolerable Daily Intake for Dioxins and Dioxin-Like PCBs*, COT/2001/07, at <http://www.foodstandards.gov.uk/multimedia/pdfs/cot-diox-full.pdf>.
 - [18] European Commission (2001). *Opinion of the Scientific Committee on Food on the Risk Assessment of Dioxins And Dioxin-like PCBs in Food*, CS/CNTM/DIOXIN/20 final, Health and Consumer Protection Directorate-General, Brussels.

Related Articles

Environmental Health Risk

Environmental Risks

Health Hazards Posed by Dioxin

What are Hazardous Materials?

GAIL CHARNLEY

Disease Mapping

Background

The analysis of epidemiological or public health data is georeferenced. Typically, the data arises in two forms: either (a) the residential address of cases of disease is known or (b) arbitrary small areas such as census tracts, zip codes, or postcodes have counts of disease observed within them. The locational information is used in the analysis usually to make inferences about spatial health effects (*see Geographic Disease Risk*).

Often hypotheses of interest in disease mapping focus on whether the residential address of cases of disease yields insight into etiology of the disease or, in a public health application, whether adverse *environmental health risks* (*see Environmental Health Risk*) exist locally within a region (as exemplified by local increases in disease risk). For example, in a study of the relationship between malaria and diabetes in Sardinia a strong negative relationship has been found. This relation had a spatial expression and the geographical distribution of malaria was important in generating explanatory models for the relation [1]. In public health practice, it is of considerable importance to be able to assess whether localized areas which have larger than expected numbers of cases, of disease, are related to any underlying environmental cause. Here, spatial evidence of a link between cases and a source is fundamental in the analysis. Evidence such as a decline in risk with distance from the putative source of hazard, or elevation of risk in a preferred direction is important in this regard (*see e.g. [2; 3, chapter 7; 4]*).

There are four main areas where statistical methods have seen development: *disease map reconstruction*, *disease clustering*, *ecological analysis*, and *disease map surveillance*. The term *disease mapping* is used in different contexts to either signify the group of methods that are applied to disease maps, or more specifically to the estimation of relative risk from disease maps (*disease map reconstruction*). Hence, there can be some confusion in the use of this term. It is appropriate to consider some common themes or issues, which arise in all areas of the subject.

Basic Definitions and Models

A fundamental feature of data available for analysis is that it is usually discrete (either in the form of a point process or counting process), and the cases of concern arise from within a local human population, which varies in spatial density and in susceptibility to the disease of interest. Hence, any model or test procedure must make allowance for this background (nuisance) population effect. The background population effect can be allowed for in a variety of ways. For count data it is commonplace to obtain expected rates for the disease of interest based on the age–sex structure of the local population, and some crude estimates of local relative risk are often computed from the ratio of observed to expected counts (e.g., SMRs: standardized mortality/incidence ratios). For case event data, expected rates are not available at the resolution of the case locations and the use of the spatial distribution of a control disease has been advocated. In that case, the spatial variation in the case disease is compared to the spatial variation in the control disease. A major issue in this approach is the correct choice of control disease. It is important to choose a control which is matched to the age–sex structure of the case disease but is unaffected by the feature of interest. For example, in the analysis of cases around a putative health hazard, a control disease should not be affected by the health hazard. Counts of control disease cases could also be used instead of expected rates when analyzing count data. Figures 1 and 2 display case event and control data maps for a region of the United Kingdom for a fixed period. Figure 3 displays a typical count data example. For the case event data in Figures 1 and 2 the object was to examine the spatial distribution of excess risk for gastric and esophageal cancer beyond that found in the control disease. Figure 3 is a map of the crude counts of congenital deaths in South Carolina. The spatial analysis of these deaths could be related to adverse environmental risk differences in the state. A geographical analysis could point to areas of adverse risk that require further investigation.

Case Event Data

Case event locations often represent residential addresses of cases and the cases arise from a heterogeneous population that varies both in spatial density

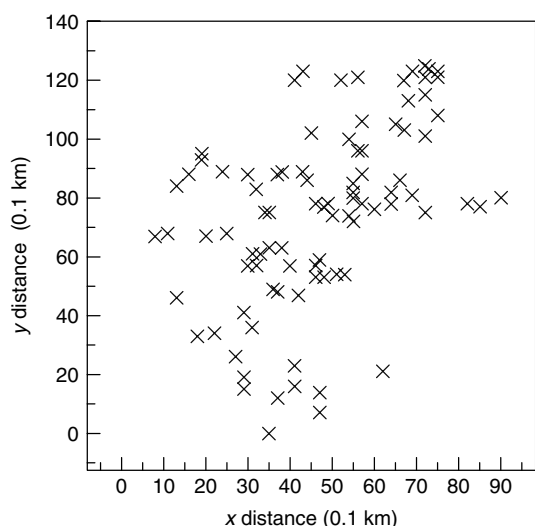


Figure 1 Distribution of death certificate addresses of gastric and esophageal cancer in Arbroath, Scotland, UK (1966–1976)

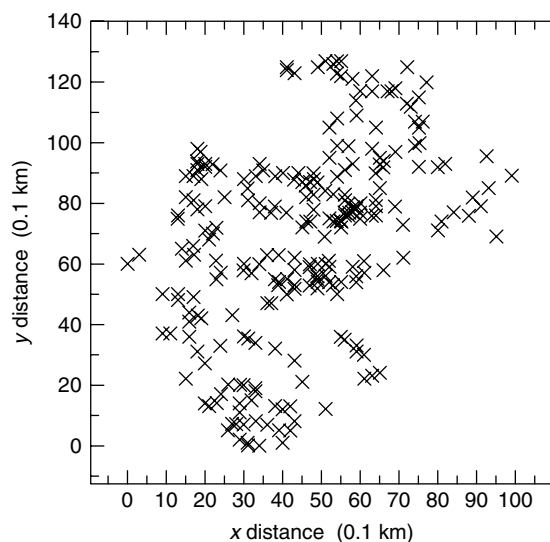


Figure 2 Control distribution: distribution of death certificate addresses for lower body cancers in Arbroath, UK (1966–1976)

and in susceptibility to disease. A heterogeneous Poisson process model is often assumed as a starting point for further analysis. The focus of interest for making inference regarding parameters describing excess risk lies in a relative risk function, which is

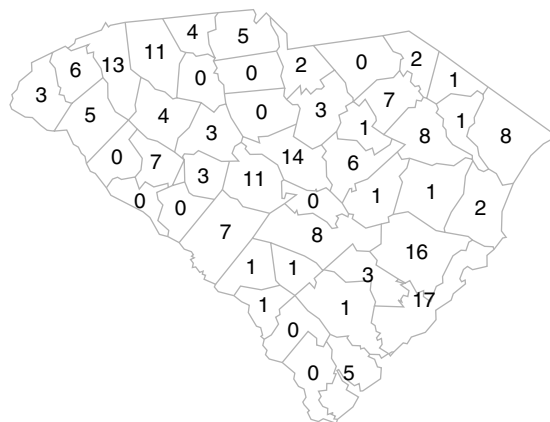


Figure 3 Distribution of counts of congenital deaths within the counties of South Carolina, USA (1990)

included in the first order intensity of the Poisson process.

It is possible that population or environmental heterogeneity may be unobserved in the data set. This could be because either the population background hazard is not directly available, or the disease displays a tendency to cluster (perhaps owing to unmeasured covariates). The heterogeneity could be spatially correlated, or it could lack correlation in which case it could be regarded as a type of “overdispersion”. One can include such unobserved heterogeneity within the framework of conventional models as a random effect.

Count Data

A considerable literature has developed concerning the analysis of count data in spatial analysis of disease (e.g., see reviews in [3] and [4]).

The usual model adopted for the analysis of region counts assumes that the counts are independent Poisson random variables with parameter θ_i in the i th region. This model may be extended to include unobserved heterogeneity between regions by introducing a prior distribution for the log relative risks ($\log \theta_i$). Incorporation of such heterogeneity has become a common approach and the Besag, York and, Mollié (BYM) model is now a standard model. A full Bayesian analysis using this model is available on WinBUGS. A recent overview of Bayesian methods (see **Bayesian Statistics in Quantitative Risk Assessment**) in disease mapping is found in [5].

References

- [1] Bernardinelli, L., Pascutto, C., Montomoli, C., Komakec, J. & Gilks, W.R. (1999). Ecological regression with errors in covariates: an application, in *Disease Mapping and Risk Assessment for Public Health*, A.B. Lawson, A. Biggeri, D. Boehning, E. Lesaffre, J.-F. Viel & R. Bertollini, eds, John Wiley & Sons.
- [2] Lawson, A.B., Bohning, D., Lessafre, E., Biggeri, A., Viel, J.F. & Bertollini, R. (1999). *Disease Mapping and Risk Assessment for Public Health*, John Wiley & Sons.
- [3] Lawson, A.B. (2006). *Statistical Methods in Spatial Epidemiology*, 2nd Edition, John Wiley & Sons, New York.
- [4] Elliott, P., Wakefield, J., Best, N. & Briggs, D. (2000). *Spatial Epidemiology: Methods and Applications*, Oxford University Press, London.
- [5] Lawson, A.B. & Banerjee, S. (2007). Bayesian spatial analysis, in *Handbook of Spatial Analysis*, S. Fotheringham & P. Rogerson, eds, Sage, New York.

ANDREW B. LAWSON

Distributions for Loss Modeling

In this article, we focus on models that can be used for the size and number of losses. For sizes of loss, we restrict ourselves to distributions that do not take on negative values. Similarly, when considering the number of losses occurring in a fixed time period, that number cannot be negative.

When searching for a distribution function to use as a model for a random phenomenon, it can be helpful if the field can be narrowed from this infinitely large set to a small set containing distribution functions with a sufficiently large variety of shapes to capture features that are encountered in practice. We provide a (partially) unified approach to distributions by combining them into nested “families”.

Continuous Size-of-Loss Distributions

We examine a number of distributions that can be used for modeling the amount of loss arising from loss events. By definition, losses can only take on values that are nonnegative. Hence we only look at distributions whose random variable can only take on nonnegative values. For most distributions, the domain runs from zero to infinity. In practice, however, losses are limited by some large amount (such as the total assets of the firm). Often in modeling losses, we ignore this fact because the probability of such an event is extremely small. If the probability of such a loss is sufficiently large to be material, we could apply a limit x to the loss random variable X , and consider the distribution of $X \wedge x = \min(X, x)$. We first introduce a selected set of continuous distributions with support on $(0, \infty)$ and with cumulative distribution function (cdf) denoted by $F(x)$.

In general, using observed data, the principle of parsimony applies. When choosing a model, a simpler model is preferred over a complex one if both models explain a phenomenon adequately. This leads to model selection criteria based on the number of parameters.

Three Important Parametric Families

Distributions can often be organized into related groupings or “families”. Three such families are illustrated in Figures 1–3. Figure 1 shows the four-parameter transformed beta distribution (also known as the *generalized beta* of the second kind, or Pearson type VI distribution)

$$F(x) = \beta\left(\tau, \alpha; \frac{(x/\theta)^\tau}{1 + (x/\theta)^\tau}\right), \quad x > 0 \quad (1)$$

and its special cases where

$$\beta(a, b; x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x t^{a-1}(1-t)^{b-1} dt, \\ a > 0, b > 0, 0 < x < 1 \quad (2)$$

is the incomplete beta function and where

$$\Gamma(\alpha; x) = \frac{1}{\Gamma(\alpha)} \int_0^x t^{\alpha-1} e^{-t} dt, \quad \alpha > 0, x > 0 \quad (3)$$

is the incomplete gamma function with

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt, \quad \alpha > 0 \quad (4)$$

There are eight special cases shown in Figure 1. These special cases arise by setting one or two nonscale parameters to 1 or equating pairs of nonscale parameters. The arrows in Figure 1 show the familial relationships. This structured familial approach to distributions is very useful when selecting from among members of the family. For example, given a set of data, one could compare the importance of the parameter τ when comparing the logistic distribution to the inverse Burr distribution using a likelihood-ratio test. Klugman *et al.* [1] use this approach in model selection. Figure 2 shows the transformed (or generalized) gamma distribution with cdf

$$F(x) = \Gamma(\alpha; (x/\theta)^\tau), \quad x > 0 \quad (5)$$

Its special cases are the gamma, Weibull, and exponential distributions. Figure 2 also shows the inverse transformed gamma distribution (also known as the inverse generalized gamma distribution) with cdf

$$F(x) = 1 - \Gamma(\alpha; (\theta/x)^\tau), \quad x > 0 \quad (6)$$

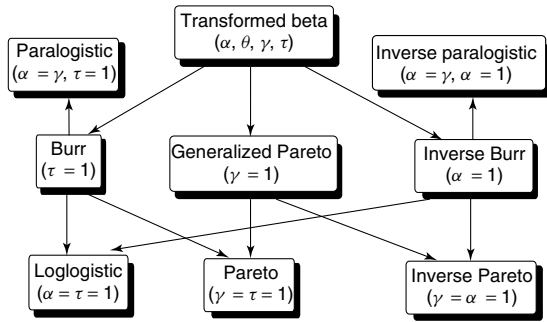


Figure 1 Transformed beta family

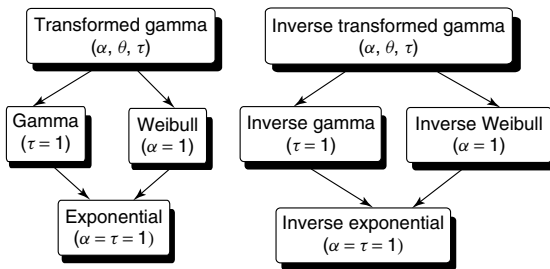


Figure 2 Transformed/inverse transformed gamma family

respectively. Its special cases are the corresponding “inverse” versions of the special cases of the transformed gamma distribution.

Another way to relate distributions is examine what happens as parameters approach the limiting values of zero or infinity. This will allow us to relate all distributions shown in Figures 1 and 2. By letting α go to infinity, the transformed gamma distribution is obtained as a limiting case of the transformed beta family. Similarly, the inverse transformed gamma distribution is obtained by letting τ go to infinity instead of α . Because the Burr distribution is a transformed beta distribution with $\tau = 1$, its limiting case is the transformed gamma with $\tau = 1$, which is the Weibull distribution. Similarly, the inverse Burr has the inverse Weibull (see **Dose-Response Analysis; Reliability Growth Testing**) as a limiting case. Finally, letting $\tau = \gamma = 1$ shows that the limiting case for the Pareto distribution is the exponential (and similarly for their inverse distributions). Figure 3 illustrates the limiting and special case relationships.^a

Tails of Continuous Distributions

The tail of a distribution (more properly, the right tail) is the portion of the distribution corresponding

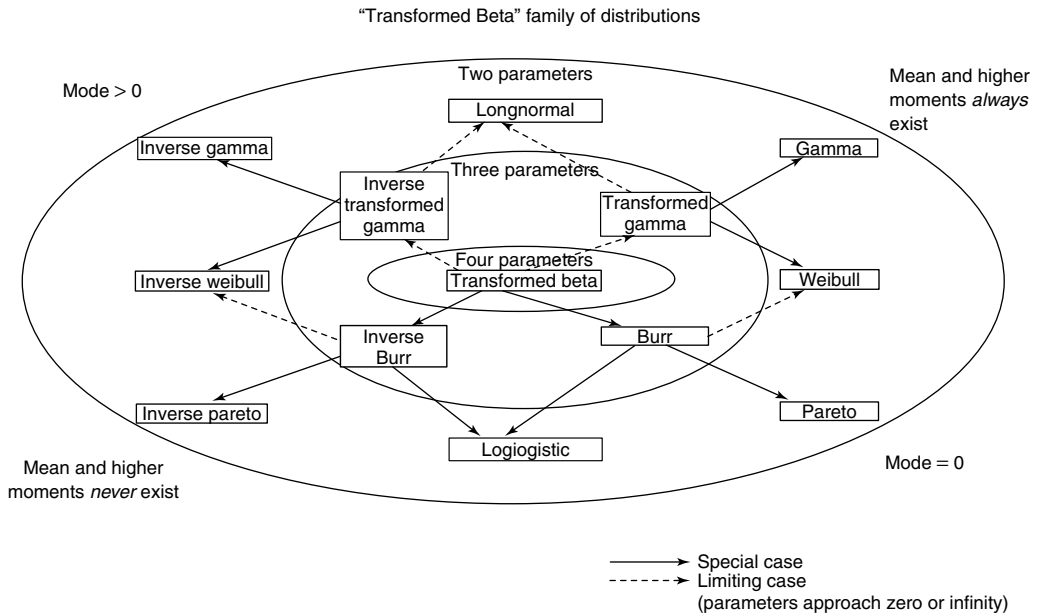


Figure 3 Distributed relationships and characteristics

to large values of the random variable. Understanding large possible loss values is important because these have the greatest impact on the total of risk losses. Distributions that tend to assign higher probabilities to larger values are said to be heavier tailed (*see Default Risk*). The “weight” of the tail can be a relative concept (model A has a heavier tail than model B) or an absolute concept (distributions with a certain property are classified as heavy tailed). When choosing models, tail weight can help narrow the choices or can confirm a choice for a model. Heavy-tailed distributions are particularly important in risk modeling in connection with extreme value theory (*see, for example [2]*).

Classification Based on Moments

The k th raw moment for a continuous random variable that takes on only positive values (such as losses) is given by $\int_0^\infty x^k f(x) dx$. Depending on the probability density function and the value of k , this integral may not exist. One way of classifying distribution is on the basis of how many positive moments exist. It is generally agreed that the existence of all positive moments indicates a light right tail, while the existence of only positive moments up to a certain value (or existence of no positive moments at all) indicates a heavy right tail.

By this classification, the Pareto distribution is said to have a heavy tail and the gamma distribution is said to have a light tail. A look at moment formulas reveals which distributions have heavy tails and which do not, as indicated by the existence of moments. Moment formulas for all distributions considered in this paper are given in Klugman *et al.* [1].

Classification Based on Tail Behavior

One commonly used indication that one distribution has a heavier tail than another distribution with the same mean is that the ratio of the two survival functions diverges to infinity (with the heavier-tailed distribution in the numerator) as the argument becomes large. This classification is based on asymptotic properties of the distributions. The divergence implies that the numerator distribution puts significantly more probability on large values. Note that it is equivalent to examine the ratio of density functions

since

$$\lim_{x \rightarrow \infty} \frac{\bar{F}_1(x)}{\bar{F}_2(x)} = \lim_{x \rightarrow \infty} \frac{\bar{F}_1'(x)}{\bar{F}_2'(x)} = \lim_{x \rightarrow \infty} \frac{f_1(x)}{f_2(x)} \quad (7)$$

where $\bar{F}(x) = 1 - F(x)$ represents the right tail of distribution (*see Operational Risk Modeling; Compliance with Treatment Allocation*).

Classification Based on Hazard Rate Function

The hazard rate function (*see Combining Information; Default Risk; Imprecise Reliability; Hazard and Hazard Ratio; Operational Risk Development*) $h(x) = -\frac{d}{dx} \ln \bar{F}(x)$ also reveals information about the tail of the distribution. Distributions with decreasing hazard are said to have heavy tails and those with increasing hazard rate functions are said to have light tails. The distribution with constant hazard rate, the exponential distribution, has neither an increasing nor a decreasing failure rate. For distributions with (asymptotically) monotone hazard rates, distributions with exponential tails divide the distributions into heavy-tailed and light-tailed distributions.

Comparisons between distributions can be made on the basis of the rate of increase or decrease of the hazard rate function. For example, a distribution has a lighter tail than another if, for large values of the argument, its hazard rate function is increasing at a faster rate.

Classification Based on Mean Excess Function

The mean excess function (*see Potency Estimation; Extreme Values in Reliability*) is given by $e(d) = E[X|X > d]$. It also gives information about tail weight. If the mean excess function is increasing in d , the distribution is considered to have a heavy tail. If the mean excess function is decreasing in d , the distribution is considered to have a light tail. Comparisons between distributions can be made on the basis of the rate of increase or decrease of the mean excess function. For example, a distribution has a heavier tail than another if, for large values of the argument, its mean excess function is increasing at a lower rate.

The mean excess loss function and the hazard rate are closely related. If the hazard rate is a decreasing function, then the mean excess loss function $e(d)$ is

an increasing function of d . Similarly, if the hazard rate is an increasing function, then the mean excess loss function is a decreasing function. It is worth noting (and is perhaps counterintuitive, however), that the converse implication is not true.

There is a second relationship between the mean excess loss function and the hazard rate. The limiting behavior of the mean excess loss function as $d \rightarrow \infty$ may be ascertained using L'Hôpital's rule. We have

$$\lim_{d \rightarrow \infty} e(d) = \lim_{d \rightarrow \infty} \frac{1}{h(d)} \quad (8)$$

as long as the indicated limits exist. These limiting relationships may be useful if the form of $F(x)$ is complicated.

Models for the Number of Losses: Counting Distributions

We now review a class of counting distributions; i.e., discrete distributions with probabilities only at the points $0, 1, 2, 3, 4, \dots$. In a risk context, counting distributions can be used to describe the number of losses or the number of loss-causing events. With an understanding of both the number of losses and the size of losses, we can have a deeper understanding of a variety of issues surrounding risk exposure than if we have only historical information about the total of all losses experienced. Also, the impact of risk mitigation strategies that address either the frequency of losses or the size of losses can be better understood. Finally, models for the number of losses are fairly easy to obtain and experience has shown that the commonly used frequency distributions perform well in modeling the propensity to generate losses. We restrict ourselves to a limited, but quite flexible, family of discrete distributions. Johnson *et al.* [3] provides a comprehensive review of such distributions.

We now formalize some of the notations that will be used for models for discrete phenomena. The probability function (pf) p_k denotes the probability that exactly k events (such as losses) occur. Let N be a random variable representing the number of such events. Then

$$p_k = \Pr(N = k), \quad k = 0, 1, 2, \dots \quad (9)$$

The probability generating function (pgf) of a discrete random variable N with pf p_k is $P(z) = E(z^N) =$

$\sum_{k=0}^{\infty} p_k z^k$. The probabilities are easily obtained from the pgf.

The $(a, b, 0)$ Class

Let p_k be the pf of a discrete random variable. It is a member of the $(a, b, 0)$ class of distributions, provided there exist constants a and b such that

$$p_k = \left(a + \frac{b}{k}\right) p_{k-1}, \quad k = 1, 2, 3, \dots \quad (10)$$

This recursive relation describes the relative size of successive probabilities in the counting distribution. The probability at zero, p_0 , can be obtained from the recursive formula because the probabilities must sum to 1. This boundary condition, together with the recursive formula, will uniquely define the probabilities. The $(a, b, 0)$ class of distributions is a two-parameter class, the two parameters being a and b . By substituting in the pf for each of the Poisson (with pgf $P(z) = e^{\lambda(z-1)}$), binomial (with pgf $P(z) = \{1 + q(z-1)\}^m$), and negative binomial (with pgf $P(z) = \{1 - \beta(z-1)^{-r}\}$) distributions on the left-hand side of the recursion, it can be seen that each of these three distributions satisfies the recursion and that the values of a and b are as given in Table 1. In addition, the table gives the value of p_0 , the starting value for the recursion. In Table 1 the geometric distribution, the one-parameter special case ($r = 1$) of the negative binomial distribution, is also present. These are the only possible distributions satisfying this recursive formula.

The $(a, b, 1)$ Class

Frequently, the four distributions discussed earlier do not adequately describe the characteristics of some data sets encountered in practice, because the distributions in the $(a, b, 0)$ class cannot capture the shape of observed data. For loss count data,

Table 1 The $(a, b, 0)$ class

Distribution	a	b	p_0
Poisson	0	λ	$e^{-\lambda}$
Binomial	$-\frac{q}{1-q}$	$(m+1)\frac{q}{1-q}$	$(1-q)^m$
Negative binomial	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	$(1+\beta)^{-r}$
Geometric	$\frac{\beta}{1+\beta}$	0	$(1+\beta)^{-1}$

the probability at zero is the probability that no losses occur during the period under study. When the probability of occurrence of a loss is very small (as is usually the case), the probability at zero has the largest value. Thus, it is important to pay special attention to the fit at this point. Similarly, it is possible to have situations in which there is less than the expected number, or even zero, occurrences at zero. Any adjustment of the probability at zero is easily handled by modifying the Poisson, binomial, and negative binomial distributions at zero.

A counting distribution is a member of the $(a, b, 1)$ class of distributions provided there exist constants a and b such that

$$p_k = \left(a + \frac{b}{k}\right) p_{k-1}, \quad k = 2, 3, 4, \dots \quad (11)$$

Note that the only difference from the $(a, b, 0)$ class is that the recursion begins at p_1 rather than at p_0 . This forces the distribution from $k = 1$ to $k = \infty$ to be proportional to the corresponding $(a, b, 0)$ distribution. The remaining probability is at $k = 0$ and can take on any value between 0 and 1.

We distinguish between the situations in which $p_0 = 0$ and those where $p_0 > 0$. The first subclass is called the *truncated* (more specifically, zero-truncated (ZT) distributions. The members are the ZT Poisson, ZT binomial, and ZT negative binomial (and its special case, the ZT geometric) distributions. The second subclass will be referred to as the zero-modified (ZM) distributions because the probability is modified from that for the $(a, b, 0)$ class.

The $(a, b, 1)$ class admits additional distributions. The (a, b) parameter space can be expanded to admit an extension of the negative binomial distribution to include cases where $-1 < r \leq 0$. For the $(a, b, 0)$ class, the condition $r > 0$ is required. By adding the additional region to the sample space, the “extended” truncated negative binomial (ETNB) distribution has parameter restrictions $\beta > 0, r > -1, r \neq 0$.

For the ETNB, the parameters and the expanded parameters space are given by

$$\begin{aligned} a &= \frac{\beta}{1 + \beta}, \quad \beta > 0, \\ b &= (r - 1) \frac{\beta}{1 + \beta}, \quad r > -1, \quad r \neq 0 \end{aligned} \quad (12)$$

When $r \rightarrow 0$, the limiting case of the ETNB is the logarithmic distribution with pf

$$p_k = \frac{[\beta/(1 + \beta)]^k}{k \ln(1 + \beta)}, \quad k = 1, 2, 3, \dots \quad (13)$$

and pgf

$$P(z) = 1 - \frac{\ln[1 - \beta(z - 1)]}{\ln(1 + \beta)} \quad (14)$$

The ZM logarithmic distribution is created by assigning an arbitrary probability at zero and adjusting the remaining probabilities proportionately.

It is also interesting that the special extreme case with $-1 < r < 0$ and $\beta \rightarrow \infty$ is a proper distribution, sometimes called the Sibuya distribution. It has pgf $P(z) = 1 - (1 - z)^{-r}$, and no moments exist. Distributions with no moments are not particularly interesting for modeling loss numbers (unless the right tail is subsequently modified) because an infinite number of losses are expected. If this is the case, the risk manager should be fired!

There are no other members of the $(a, b, 1)$ class beyond those discussed above. A summary is given in Table 2.

Compound Frequency Models

A compound frequency larger class of distributions can be created by the processes of compounding any two discrete distributions. The term *compounding* reflects the idea that the pgf of the new distribution $P(z)$ is written as $P(z) = P_N[P_M(z)]$, where $P_N(z)$ and $P_M(z)$ are called the *primary* and *secondary distributions*, respectively.

The compound distributions arise naturally as follows. Let N be a counting random variable with pgf $P_N(z)$. Let $M_1, M_2, \dots$ be identically and independently distributed (i.i.d.) counting random variables with pgf $P_M(z)$. Assuming that the M_j s do not depend on N , the pgf of the random sum $S = M_1 + M_2 + \dots + M_N$ (where $N = 0$ implies that $S = 0$) is $P_S(z) = P_N[P_M(z)]$.

In risk contexts, this distribution can arise naturally. If N represents the number of loss-causing events and $\{M_k; k = 1, 2, \dots, N\}$ represents the number of losses (errors, injuries, failures, etc.) from the individual events, then S represents the total number of losses for all such events. This kind of interpretation is not necessary to justify the use of

Table 2 The $(a, b, 1)$ class

Distribution	$p0$	a	b	Parameter space
Poisson	$e^{-\lambda}$	0	λ	$\lambda > 0$
ZT Poisson	0	0	λ	$\lambda > 0$
ZM Poisson	Arbitrary	0	λ	$\lambda > 0$
Binomial	$(1-q)^m$	$-\frac{q}{1-q}$	$(m+1)\frac{q}{1-q}$	$0 < q < 1$
ZT binomial	0	$-\frac{q}{1-q}$	$(m+1)\frac{q}{1-q}$	$0 < q < 1$
ZM binomial	Arbitrary	$-\frac{q}{1-q}$	$(m+1)\frac{q}{1-q}$	$0 < q < 1$
Negative binomial	$(1+\beta)^{-r}$	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	$r > 0, \beta > 0$
ETNB	0	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	$r > -1, \beta > 0$
ZM ETNB	Arbitrary	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	$r > -1, \beta > 0$
Geometric	$(1+\beta)^{-1}$	$\frac{\beta}{1+\beta}$	0	$\beta > 0$
ZT geometric	0	$\frac{\beta}{1+\beta}$	0	$\beta > 0$
ZM geometric	Arbitrary	$\frac{\beta}{1+\beta}$	0	$\beta > 0$
Logarithmic	0	$\frac{\beta}{1+\beta}$	$-\frac{\beta}{1+\beta}$	$\beta > 0$
ZM logarithmic	Arbitrary	$\frac{\beta}{1+\beta}$	$-\frac{\beta}{1+\beta}$	$\beta > 0$

a compound distribution. If a compound distribution fits data well, that may be enough justification itself.

Consider the case where both M and N have the Poisson distribution. Determine the pgf of this distribution. This distribution is called the *Poisson–Poisson* or *Neyman Type A distribution*. Let $P_N(z) = e^{\lambda_1(z-1)}$ and $P_M(z) = e^{\lambda_2(z-1)}$. Then

$$P_S(z) = e^{\lambda_1[e^{\lambda_2(z-1)}-1]} \quad (15)$$

The Poisson–logarithmic distribution can be shown to be a negative binomial distribution. Thus the Poisson–logarithmic distribution does not create a new distribution beyond the $(a, b, 0)$ and $(a, b, 1)$ classes. Another combination that does not create a new distribution beyond the $(a, b, 1)$ class is the compound geometric distribution where both the primary and secondary distributions are geometric. The resulting distribution is a ZM geometric distribution.

Finally it is easy to show that adding, deleting, or modifying the probability at zero in the secondary distribution does not add a new distribution because it is equivalent to modifying the parameter θ of the primary distribution. This means that, for example, a Poisson primary distribution with a Poisson, ZT Poisson, or ZM Poisson secondary distribution will

still lead to a Neyman Type A (Poisson–Poisson) distribution.

Relationships among discrete distributions are shown in Table 3.

Recursive Calculation of Compound Probabilities

The probability of exactly k losses can be written as

$$\Pr(S = k) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \Pr(M_1 + \cdots + M_n = k) \quad (16)$$

$$\Pr(N = n)$$

Letting $g_n = \Pr(S = n)$, $p_n = \Pr(N = n)$, and $f_n = \Pr(M = n)$, this is rewritten as

$$g_k = \sum_{n=0}^{\infty} p_n f_k^{*n} \quad (17)$$

where f_k^{*n} , $k = 0, 1, \dots$, is the n -fold convolution of the function f_k , $k = 0, 1, \dots$, that is, the probability that the sum of n random variables that are each i.i.d. with probability function f_k will take on value k .

When $P_N(z)$ is chosen to be a member of the $(a, b, 0)$ class, the recursive formula

$$g_k = \frac{1}{1 - af_0} \sum_{j=1}^k \left(a + \frac{bj}{k} \right) f_j g_{k-j},$$

Table 3 Relationships among discrete distributions

Distribution	Is a special case of:	Is a limiting case of:
Poisson	ZM Poisson	Negative binomial, Poisson–binomial, Poisson–inv. Gaussian, Polya–Aeppli, Neyman–A
ZT Poisson	ZM Poisson	ZT negative binomial
ZM Poisson		ZM negative binomial
Geometric	Negative binomial	Geometric–Poisson
	ZM geometric	
ZT geometric	ZT negative binomial	
ZM geometric	ZM negative binomial	
Logarithmic		ZT negative binomial
ZM logarithmic		ZM negative binomial
Binomial	ZM binomial	
Negative binomial	ZM negative binomial	Poisson–ETNB
Poisson–inverse Gaussian	Poisson–ETNB	
Polya–Aeppli	Poisson–ETNB	
Neyman–A		Poisson–ETNB

$$k = 1, 2, 3, \dots \quad (18)$$

End Notes

^a. Thanks to David Clark for creating this picture.

can be used to evaluate probabilities.

This recursion has become known as *the Panjer recursion* (see **Comonotonicity**) after its introduction as a computational tool for obtaining numerical values of the distribution of aggregate losses by Panjer [4]. Its use here is numerically equivalent to its use for aggregate losses where the secondary distribution is the distribution of loss sizes. This result was generalized by Sundt and Jewell [5] to the case of the primary distribution being a member of the $(a, b, 1)$ class:

$$g_k = \frac{[p_1 - (a + b)p_0]f_k + \sum_{j=1}^k \left(a + \frac{bj}{k}\right) f_j g_{k-j}}{1 - af_0}, \quad k = 1, 2, 3, \dots \quad (19)$$

References

- [1] Klugman, S., Panjer, H. & Willmot, G. (2004). *Loss Models From Data to Decisions*, 2nd Edition, John Wiley & Sons, New York.
- [2] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events*, Springer-Verlag, Berlin.
- [3] Johnson, N., Kotz, S. & Kemp, A. (1993). *Univariate Discrete Distributions*, 2nd Edition, John Wiley & Sons, New York.
- [4] Panjer, H. (1981). Recursive evaluation of a family of compound distributions, *ASTIN Bulletin* **12**, 22–26.
- [5] Sundt, B. & Jewell, W. (1981). Further results on recursive evaluation of compound distributions, *ASTIN Bulletin* **12**, 27–39.

HARRY H. PANJER

Dose–Response Analysis

One component of quantitative risk assessment is the search for and the study of the relationship between exposure and effect. Proof of the existence of a *dose–response relationship* is a basic element for the proof of a cause–effect relationship between the exposure to an agent and the occurrence of an effect (see **Causality/Causation**). As a consequence, *dose–response assessment* is a key element in carcinogenic risk assessment. It is part of the hazard characterization step, when the four-step approach to risk assessment is adopted (hazard identification, hazard characterization, exposure assessment, and risk characterization).

The purpose of a dose–response analysis in quantitative risk assessment is the detection, characterization, and quantification of a relationship $D \rightarrow Y(D)$ between the critical effect Y on human health and environment and the dose D of the hazardous exposure. Establishment of a dose–response relationship requires support from life science (biology, medicine, and environmental science), statistical science, and risk management. Life sciences provide both, knowledge for the appropriate choice of the *dose–response model* and the *dose–response data* from appropriate experiments or empirical observations. Statistical science applies methods of statistical inference for estimating effect parameters, characterizing their uncertainty and prediction. Risk assessors and risk managers select in cooperation with the former two disciplines the appropriate risk metric R as function of the basic effect measure $Y: D \rightarrow Y(D) \rightarrow R(D)$.

Dose–Response Data

Dose data are of quantitative nature and describe amounts or concentrations of the agent in the individual after exposure. Therefore dose is a numerical value describing the exact amount of the agent entering the organism or acting at the target site. An agent can enter an organism at various instances (environmental, accidental, by purpose, etc.), through various routes (inhalation, dermal, oral, etc.) before it is deposited at a site of action and before it is metabolized, may induce production of other hazardous substances (see **What are Hazardous Materials?**), and finally may be excreted. As a consequence, one

has to specify nature (applied dose, internal dose, delivered dose, etc.) of the dose. Difficulties in determining the delivered dose, also called *target organ dose*, were identified in [1] by five questions:

- Is dose expressed as an environmental concentration, applied dose, or delivered dose to the target organ?
- Is dose expressed in terms of a parent compound, one or more metabolites, or both?
- Is the impact of dose patterns and timing significant?
- Is there a conversion from animal to human doses, when animal data are used?
- Is a conversion metric between routes of exposure necessary and appropriate?

With respect to the first question one should emphasize that the goal of dose modeling is to estimate, to the extent possible, the delivered dose of the active agent at the target organ or target cell. When direct measurement is not possible, this could be achieved using pharmacokinetic models when sufficient and confident knowledge is available; otherwise, dose–response modeling often proceeds with less specific dose metrics, e.g., the average daily dose of the agent. Similarly, kinetic modeling has been very helpful to determine dose metrics when metabolic pathways have been known as *relevant for the risk assessment*. Interspecies and intraspecies adjustments of the dose and route-to-route extrapolation may use physiologically based models to calculate a dose parameter. For more details on how to deal with these questions, all of which address issues of extrapolations in a more broad sense, see the overview given by Cogliano [2].

Response data characterize the toxic endpoints. Multiplicity of adverse effects may be handled by defining primary and secondary toxic endpoints or through agreement upon the most critical endpoint using information on severity and relevance. Otherwise, statistical methods for multiple testing have to be considered. In cancer risk assessment, tumor incidence data and tumor specific mortality data have been the basic endpoints of response for a dose–response assessment. Response can be analyzed directly by setting the risk measure R equal to the effect measure Y or by defining a risk measure as a function $R(D) = R(Y(D))$ of the observations.

Depending on the scale of the endpoints one can distinguish three response types:

Quantal

$R(D) = P(Y(D) = 1)$: probability of the occurrence of the adverse health event: e.g., tumor, liver damage, renal failure, and death.

Categorical

$R_x(D) = P(Y(D) \geq x)$: probability of the occurrence of an adverse health event at least as serious as a given category, say, x , e.g., adverse event of at least grade 3.

Quantitative

$R(D) = Y(D)$: level of an adverse health outcome observed on a continuous metric scale that is either increasing or decreasing with dose e.g., blood cell counts and weight decay.

Response can be analyzed directly as above or related to a reference response, e.g., the background response $R(0)$. *Additional risk* is defined as $R_{\text{add}}(D) = R(D) - R(0)$. *Extra risk* is defined as $R_{\text{extra}}(D) = [R(D) - R(0)]/[R_{\text{max}} - R(0)]$ where R_{max} is the maximum possible risk level of the study. In case of quantal response $R_{\text{max}} = 1$; in case of quantitative response one may chose $R_{\text{max}} = \max(y_{ij} | i = 1, \dots, I; j = 1, \dots, n_i)$; y_{ij} defined next.

Empirical *dose–response data* are generated in experimental or observational studies. In a controlled experiment, the exposure dose is applied to the individual experimental unit and varies on the basis of the experimental design. The effect size or the occurrence of an effect is recorded. In observational studies, both the individual’s exposure dose and the effect size, or the occurrence of an effect are recorded. It is important to note that a variation of the dose over some dose range takes place. In both cases, data are of the following type:

A total of N individuals is assigned to a set of $I + 1$ groups such that n_i individuals are assigned the dose d_i , respectively; $i = 0, 1, \dots, I$, where $I \geq 2$ doses are arranged in an increasing order, $0 < d_1 < d_2 < \dots < d_I$, and where $d_0 = 0$ denotes the control group, $n_i \geq 0$ and $\sum_{i=0}^I n_i = N$. Each among the n_i individuals exposed to dose d_i exhibits a response y_{ij} , $j = 1, \dots, n_i$, $i = 0, 1, \dots, I$. The dose–response data are then represented as $\{(y_{ij}, d_i), j = 1, \dots, n_i, i = 0, 1, \dots, I\}$. It should be noted that, in a strict sense, dose–response data require at least three groups, two dose groups and a control or three dose groups. The response variable Y , with values y_{ij} for subject j in group i could be an endpoint of continuous, categorical, or dichotomous (presence or absence of a target effect) scale. When considering tumor incidence in a carcinogenesis experiment the response would be either $y_{ij} = 1$ (tumor) or $y_{ij} = 0$ (no tumor). Those data are summarized as $\{d_i, n_i, p_i, i = 0, 1, \dots, I\}$ where $r_i = \sum_{j=1}^{n_i} y_{ij}$ denotes the number of tumors in dose group i and $p_i = r_i/n_i$ is the proportion of tumor-bearing individuals. An example of tumor incidence data is given in Table 1.

Qualitative Dose–Response Analysis

Statistical inference for dose–response data can be obtained through a qualitative or a quantitative methodology depending on whether functional information is used to select the dose–response model or not. If no functional information on dose–response is postulated, one may analyze the dose–response data (d_i, y_{ij}) using multiple hypotheses testing with the null hypothesis

$$H_0: F_0 = F_1 = \dots = F_I \quad (1)$$

where “ $F_i = F_j$ ” describes the fact that the two groups with doses d_i and d_j provide the same effect.

Table 1 Original and fitted data from a dose–response experiment on cancer incidence with quantal response when the Weibull model (see Figure 1) was applied

Dose group	Dose d_i	Sample size n_i	Observed incidence r_i	Observed proportion p_i	Expected incidence	Expected proportion
1	0.0	86	2	0.023	1.67	0.019
2	0.001	50	1	0.02	2.52	0.051
3	0.01	50	9	0.18	6.92	0.138
4	0.1	45	18	0.40	18.78	0.417

A qualitative dose-response model is defined through the alternative H_1 , e.g., a trend alternative

$$H_1 : F_0 \leq F_1 \leq \dots \leq F_I \quad (2)$$

where $F_i \leq F_j$ describes the fact that the response of dose d_j is larger than at dose d_i . If one assumes, e.g., Gaussian normally distributed responses $Y_{ij} \sim N(\mu_i, \sigma_i^2)$, the alternative of monotone increasing trend in the means is given as $H_0 = \mu_0 \leq \mu_1 \leq \dots \leq \mu_I$ with $\mu_0 < \mu_I$. If the responses are proportions p_i and if the alternative $H_1 = p_0 \leq p_1 \leq \dots \leq p_I$ with $p_0 < p_I$ is of interest one usually applies the Cochran-Armitage trend test (see [3]). For situations where a monotone alternative is no more tested, but a downturn of the dose-response relationship is possible, the so-called umbrella alternative is recommended $H_0: \mu_0 \leq \mu_1 \leq \dots \mu_i^* \geq \mu_{i+1}^* \geq \dots \geq \mu_I$, where the * indicates the dose location of downturn.

Dose-Response Models

If functional information on dose-response is available, a dose-response model can be selected from standard models used for quantal dose-response analysis [4]:

Probit

$$R(d) = p + (1 - p) \cdot \Phi(a + b \ln d)$$

Logit

$$R(d) = p + (1 - p) \cdot [1 + \exp(-(a + b \ln d))]^{-1}$$

Weibull

$$R(d) = p + (1 - p) \cdot (1 - \exp(-bd^k))$$

Gamma multihit

$$R(d) = p + (1 - p) \cdot \int_0^d [b^n t^{n-1} / \Gamma(n)] \cdot \exp(-bt) dt$$

One hit

$$R(d) = p + (1 - p) \cdot [1 - \exp(-bd)]$$

Multistage

$$R(d) = p + (1 - p) \cdot \left[1 - \exp\left(-\sum_{i=0}^K a_i d^i\right) \right]$$

when Φ stands for the Gaussian distribution and Γ for the γ function.

For continuous response analysis, a family of nonlinear regression models for the mean response $\mu_i = E[Y_i]$ as a function of dose d_i , has been proposed [5]; also see [6] for a set of models for mutagenic response. For the case of discrete count data (e.g., number of defect cells) one usually applies log-linear regression models, which are part of the larger class of statistical generalized linear models. (For models with a threshold see **Threshold Models**.)

Models of the type described above have been termed *empirical models* or *curve fitting models* because of their origin as purely statistical models (e.g., the probit or logit model), or because of only limited relationship to biological processes (e.g., the multihit model).

The decision on the shape of the dose-response analysis is often based on a judgment of the mechanism for the substance under evaluation. The mechanism of genotoxicity is most prominent in carcinogenic risk assessment. Nongenotoxic carcinogens are dealt with using a threshold type approach where an estimate of the “threshold” is obtained and safety factors are applied for the extrapolation to low doses, whereas genotoxic carcinogens are dealt with using a nonthreshold dose-response relationship and a mathematical dose-response model is used for low dose extrapolation.

These models are in contrast to biologically based models or mechanistic models usually built as stochastic mathematical models based on biological information and knowledge of mode of action of the agent in the organism. Most prominent examples in cancer risk assessment are multistage carcinogenesis models and the two-stage model of clonal expansion; see e.g., the review of [7].

A large family of empirical models can be defined using the “tolerance dose distribution” concept. This is a general statistical approach for establishing dose-response functions $P(d)$. A toxic effect is considered to occur at dose d if the individual’s resistance to the substance is broken at that dose. The excess risk $P(d)$ is then modeled as the probability that the tolerance dose of an individual is less or equal to d :

$$\text{Prob}(\text{tolerance} \leq d) = P(d) \quad (3)$$

where $P(d)$ may be any monotone increasing function with values between 0 and 1. General classes of quantal dose-response models are of the form $P(d) = p + (1 - p)F(g(d, \beta))$, where the “inner”

dose function $g(d, \beta)$ can be taken from a flexible class of transformations and is parameterized by β .

Quantal dose–response models listed above can be extended to address age-dependent incidence $P(t, d)$ by incorporating straightforwardly the time (or age) of the investigated individual into the model which previously had not been time dependent [4]. Such, one can define

Probit

$$P(t, d) = p + (1 - p) \cdot \Phi(a + b \ln d + c \ln t)$$

Logit

$$P(t, d) = p + (1 - p) \cdot [1 + \exp(-(a + b \ln d + c \ln t))]^{-1}$$

Weibull

$$P(t, d) = p + (1 - p) \cdot (1 - \exp(bd^m t^k))$$

Multihit

$$P(t, d) = p + (1 - p) \cdot \int_0^{dt} [b^n u^{n-1} / \Gamma(n)] \exp(-bu) du$$

Multistage

$$P(t, d) = p + (1 - p) \cdot \left\{ 1 - \exp \left[- \sum_{i=0}^K a_i d^i \right] b t^K \right\}$$

More generally, an age-dependent tumor incidence model can be defined using a nonparametric proportional hazard function $\lambda^T(t, d)$ of the time-to-tumor T. The practical estimation of $\lambda^T(t, d)$ from carcinogenesis bioassay data requires statistical methodology of censored survival times and consideration of competing risks since occurrence of the event of interest competes with death, when aging is allowed [8].

Quantitative Dose–Response Analysis

Dose–response analysis aims both at estimating the risk at a particular exposure level and the exposure level associated with a particular risk. A quantitative analysis of the risk of a detrimental effect affecting the individual and denoted R , assumes that exposure can be quantified in terms of a dose $D = d$, and the existence of a functional relationship F between R and D of the general form

$$R = R(D) = F(D; t, x, \beta) \quad (4)$$

where t denotes the time of the observation or the occurrence of the effect, x denotes characteristic(s) of the individual, and β denotes the parameter(s) specifying the functional relationship F . The two aims of a dose–response analysis can then be specified as

1. Estimation of F such that one can calculate the risk $R(d)$ for each given dose $d : \hat{R}(d) = F(d; t, x, \hat{\beta})$.
2. Estimation of F such that one can calculate the dose $D(r)$ for each given risk level $r : \hat{D}(r) = F^{-1}(r; t, x, \hat{\beta})$.

where $\hat{\beta}$ denotes the estimate of the model parameter β in the function F and where F^{-1} denotes the inverse of the function F (existence assumed). Therefore, the analysis of a dose–response relationship proceeds in two steps. First, the dose relationship F has to be established using a mathematical modeling approach. Second, the function F has to be specified completely through statistical estimation. Therefore, usually maximum-likelihood methods are applied to estimate the model parameter β such that $\hat{F}$ as function of $\hat{\beta}$ is also a maximum-likelihood estimate.

Maximum-likelihood estimation for quantal data grounds on the binomial distribution of the number of responses $R_i = r_i$ at the dose level d_i : $R_i \sim \text{Binomial}(n_i, p_i)$ where p_i is defined in the dose–response model. For continuous data a general model is $R(d) = \mu(d) + \varepsilon$, where μ denotes a predictor and $\varepsilon \sim N(0, \sigma^2)$ the error term [3]. Asymptotic theory of inference is well established for the quantal dose–response model and it provides confidence interval estimates, resulting from the asymptotic properties of maximum likelihood estimates which are consistent, unbiased, and efficient (minimal variance) with convergence to a Gaussian distribution. Fisher information theory is used to derive the variance–covariance estimates from which asymptotic confidence intervals can be calculated. Application of Fieller’s theorem then allows the calculation of confidence intervals for functions of model parameters. (For the extrapolation to low doses see **Low-Dose Extrapolation** and **Benchmark Dose Estimation** or [9] and [10].) Recommended steps for the dose–response analysis are as follows:

- identify the kinds of data available on dose and response;

- select the response and dose metric for assessment;
- present and discuss the results of the dose–response assessment; and
- explain the results of the analyses in terms of quality of data available.

Example

A dose–response analysis is illustrated with data on the incidence of hepatocellular carcinoma and adenoma after exposure to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (see **Health Hazards Posed by Dioxin**) from an experiment with one control and three dose groups exposed to doses $d_i, i = 1, \dots, 3$ and quantal responses $p_i = r_i/n_i, i = 1, \dots, 3$ as shown in columns 1–5 of Table 1 (see the lower part of Table 2 in reference [11]). There is a statistically significant trend of increasing tumor incidence from 2% in the control group to 40% in the highest dose group (Cochran–Armitage test for linear trend: $p = 7.6 \times 10^{-10}$). When selecting the Weibull model $R(d) = p + (1 - p)(1 - \exp(-bd^k))$ with the three parameters p = background response, b = slope, and k = shape for a quantitative dose–response analysis one obtains the maximum-likelihood estimates $\hat{p} = 0.0194$ (0.0134), $\hat{b} = 2.0934$ (1.0834), and $\hat{k} = 0.6044$ (0.1589) with standard errors of the parameter estimates in parentheses. The fitted curve $\hat{R}(d) = 0.0194 + 0.9806 [1 - \exp(2.0934 \times d^{0.6044})]$ is shown in Figure 1. This evaluation was

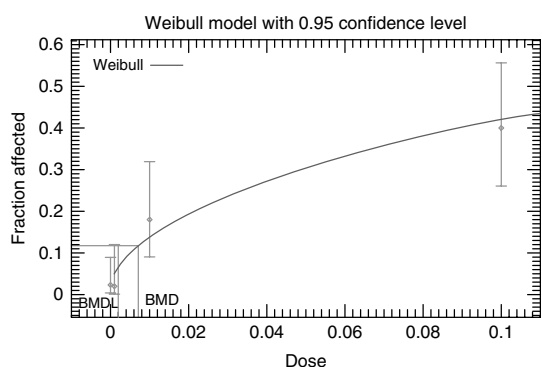


Figure 1 Fit of the Weibull model $R(d) = p + (1 - p) \cdot (1 - \exp(-bd^k))$ to the data of Table 1 using USEPA software *BMDs*. Location of the BMD and 95% – BMDL for the extra risk to specified BMR of 10% are indicated in the graphic. BMD is estimated as 0.00711 and the BMDL as 0.00198

performed using the USEPA software *BMDs*, which allows the estimation of the benchmark dose (BMD) and the 95% lower benchmark dose limit (BMDL) indicated in Figure 1.

References

- [1] US Environmental Protection Agency (1999). *Guidelines for Carcinogenic Risk Assessment*, EPA, Washington, DC.
- [2] Coglianò, V.C. (2005). Principles of cancer risk assessment: the risk assessment paradigm, in *Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment*, Edler L. & Kitsos C.P., eds, John Wiley & Sons, Chichester.
- [3] Piegorsch, W.W. & Bailer, A.J. (1997). *Statistics for Environmental Biology and Toxicology*, Chapman & Hall, London.
- [4] Edler, L., Kopp-Schneider, A. & Heinzl, H. (2005). Dose-response-modeling, in *Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment*, Edler L. & Kitsos C.P., eds, John Wiley & Sons, Chichester, pp. 211–237.
- [5] Slob, W. (2001). Dose-response modeling of continuous endpoints, *Toxicological Sciences* **66**, 298–312.
- [6] Edler, L. (1992). Statistical methods for short-term tests in genetic toxicology – the first fifteen years, *Mutation Research* **277**, 11–33.
- [7] Kopp-Schneider, A. (1997). Carcinogenesis models for risk assessment, *Statistical Methods in Medical Research* **6**, 317–340.
- [8] Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E. & Wahrendorf, J. (1986). *Statistical Methods in Cancer Research, Vol. III – The Design and Analysis of Long-term Animal Experiments*, International Agency for Research on Cancer (IARC), Lyon.
- [9] Edler, L., Poirier, K., Dourson, M., Kleiner, J., Mileson, B., Nordman, H., Renwick, A., Slob, W., Walton, K. & Würtzen, G. (2003). Mathematical modeling and quantitative methods, *Food and Chemical Toxicology* **41**, 283–326.
- [10] Sand, S. (2005). *Dose-Response Modeling: Evaluation, Application, and Development of Procedures for Benchmark Dose Analysis in Health Risk Assessment of Chemical Substances*, Institute of Environmental Medicine, Karolinska Institute Stockholm, Sweden.
- [11] Keenan, R.E., Paustenbach, D.J., Wenning, R.J. & Parsons, A.H. (1991). Pathology reevaluation of the Kociba *et al.* (1978) bioassay of 2,3,7,8-TCDD: implications for risk assessment, *Journal of Toxicology and Environmental Health* **34**, 279–296.

Related Articles

Change Point Analysis

Dynamic Financial Analysis

Dynamic financial analysis (DFA) is *dynamic* in the sense that it incorporates stochastic, or random, elements to reflect a variety of possible outcomes. The term *financial* reflects the fact that DFA incorporates, for insurers, both the underwriting side of the business and the investment side of operations; for any firm it involves combining assets and liabilities to project the aggregate financial position of the organization. The term *analysis* reflects the fact that this approach models the complex interrelationships involved in an organization. Other terms used to describe this process include dynamic financial condition analysis, dynamic capital adequacy testing, or dynamic solvency testing.

A DFA model starts with the current financial position of the organization. The key elements that impact operations are modeled mathematically, incorporating random elements where appropriate. A DFA model simulates the results of hundreds or thousands of iterations, and the financial condition of the firm is determined for each set of results. The relationship between assets and liabilities for each iteration shows whether the firm will be solvent or in financial difficulty in future years. The output from DFA is the distribution of potential financial results for the next few years.

DFA is used for a variety of reasons, including solvency regulation (*see Solvency*), assigning ratings to an organization, evaluating the impact of a change in operations, and determining the value of an organization for a merger or acquisition. When used for solvency regulation, the output from DFA is used to determine how frequently the organization will encounter financial difficulties and how severe those difficulties will be. Regulators utilize DFA to determine which insurers should be put under surveillance, whether additional capital is required for the firm, or if restrictions on operations should be instituted. DFA can alert regulators to potential problems before the firm becomes insolvent. Rating agencies utilize the results of DFA in assigning ratings to organizations, with those least likely to encounter financial difficulties obtaining the highest ratings and those most likely to become impaired receiving lower ratings.

Firms can often respond to potential ratings by changing their operations or capitalization level to reduce the risk of financial impairment and improve their rating. In this event, DFA helps to reduce the number of financial impairments. Organizations also use DFA to evaluate different operational strategies before deciding changes in its operations. DFA is commonly used to evaluate different reinsurance (*see Optimal Risk-Sharing and Deductibles in Insurance; Reinsurance*) contracts, to determine whether to enter a new geographical area or market, or to determine the appropriate growth strategy. When used in merger and acquisition situations, DFA indicates the financial position of the organization under a variety of different economic conditions in order to determine the value of the firm.

Other financial planning approaches include forecasting, sensitivity analysis, and scenario testing. Under forecasting, a firm develops a financial plan based on the most likely conditions to develop. This approach uses a single plan as a benchmark for future performance. Frequently, firms then focus on deviations between the actual results that develop and the financial plan, as if the plan represents what should have happened and any difference resulting from something that went wrong. However, reliance on a single plan does not reflect the uncertainty that is inherent in any business operation. A variation of the single plan approach is to develop three forecasts, representing an optimistic view, a realistic view, or a pessimistic view of future developments. This approach reflects the uncertainty inherent in future developments, but there is no probability estimates associated with each set of results, so there is no way to know how likely each of the forecasts is expected to occur.

Sensitivity analysis also reflects the uncertainty associated with forecasting. Under sensitivity analysis, each important variable that affects operations is varied from the realistic estimate (the base case) to the optimistic and pessimistic values, and the financial forecast is recalculated to reflect this variation. Each variable is changed one at a time, with all other variables being held at the base case level. Sensitivity analysis shows which variables have the greatest impact on operations, so that the firm can focus on generating accurate estimates of these values and can try controlling the variation. However, by restricting the changes to one variable at a time, the range of

uncertainty that is inherent in the forecast is restricted unrealistically.

Scenario testing (*see* **Scenario-Based Risk Management and Simulation Optimization; Actuary**) considers the combined impact of changing a number of variables simultaneously by determining financial forecasts based on a few situations that can reasonably be expected to occur in concert. For example, one scenario could involve interest rates increasing by 100 basis points, equity markets declining by 5%, and unemployment rates increasing by 0.5%. Another scenario could involve interest rates increasing by 200 basis points, equity markets declining by 10%, and unemployment rates increasing by 1.5%. These changes represent realistic relationships among these variables. Several scenarios are developed, and the financial results are determined on the basis of each scenario. A firm's ability to remain financially healthy in each scenario is then evaluated. Although scenario analysis does reflect relationships among variables that are important to a firm's operations, there are several problems with this approach. First, scenario testing considers only a few situations of the many that could occur. This approach gives no indication about the impact of other potential situations. Secondly, there is no information about the likelihood of any particular scenario developing, so it is impossible to evaluate whether a firm's weak financial condition in a particular scenario is a serious problem or not.

DFA, if applied properly, can provide much more valuable information about the financial condition of a firm than forecasting, sensitivity analysis, or scenario testing. The financial condition of the firm under a wide variety of potential developments can be determined, and the likelihood of each set of outcomes can be measured. For example, interest rates could vary over a range of several hundred basis points in the iterations that are run. Each change in interest rates would be associated with a particular change in equity returns, unemployment levels, and any other variable included in the model. The results of the DFA would provide not only the impact of the developments, but the likelihood as well. If the firm is only financially impaired in 1 of the 10 000 iterations, then regulators and the management may not be particularly concerned about that situation. However, if the firm faces financial difficulty in 500 of those 10 000 iterations, and the cause in most cases is a significant increase in interest rates, then steps to reduce interest rate exposure should be considered.

Although DFA is a much more valuable tool than other financial planning techniques, developing a DFA model requires extensive effort. Organizations often spend a year or more, with a number of people working together, to develop a useable DFA model. Consulting firms and large financial services organizations have built these models, but in almost all cases the organization considers the model to be proprietary and will not share the model, or even the details about how the model works, with others. Consulting firms that perform DFA projects for clients will provide the results, but not the model to the customer. One exception to this approach is the public access DFA model known as *Dynamo* developed by Pinnacle Actuaries for property-liability insurers, which is available at no charge on their web site: www.pinnacleactuaries.com. This model is representative of most other DFA models, and can be used as an example of how a DFA model functions.

All DFA models are based on a number of key assumptions that govern how the values of future financial measures are determined. One critical assumption is the interest rate generator (*see* **Asset-Liability Management for Life Insurers**). DFA models tend to use one or more of the popular term structure models in finance (*see* Cairns [1] for a description of these models). Another is the inflation generator (*see* **Asset-Liability Management for Nonlife Insurers**). Inflation can be determined as a function of the nominal interest rate, or as a separately determined value that is then combined with the real interest rate to obtain the nominal rate. Stock returns are modeled on the basis of a single distribution, commonly the lognormal, or on the basis of a regime switching model [2]. The other key assumption for property-liability insurers is catastrophe claims. Some DFA models include an internal catastrophe loss generator (*see* **Insurance Pricing/Nonlife**); the number of catastrophes can be determined on the basis of Poisson distribution and the size of the catastrophe based on a lognormal distribution. More commonly, DFA models rely on one of the catastrophe modeling specialists (such as applied insurance research (AIR), risk management solutions (RMS), or EQECAT) to provide company specific values for catastrophe losses for a large number of simulations that are then incorporated into the DFA model.

The output from a DFA run depends on the parameters for each component. The parameters can

be determined from historical data that is generally updated annually (equity returns) from current values [such as using the contemporary yield curve (*see Solvency; Structured Products and Hybrid Securities*) in an arbitrage interest rate model (*see Multistate Models for Life Insurance Mathematics*)] or based on expert opinion about future conditions (catastrophe models that predict that climate change will affect storm severity). As the objective of a DFA model is to model potential future values, care needs to be taken to assure that the model is not being parameterized to generate historical loss patterns if conditions have changed significantly.

Dynamo has several interrelated modules that interact to generate the results. One module generates the underwriting gains and losses of the insurer. Key inputs to this module are the number of policies in force, premium levels, current loss frequency and severity, loss reserves, expected growth rates, and the current segment of the underwriting cycle the industry is in. Historically, the insurance industry fluctuates between soft markets, where competitive pressure keeps premium levels low (and underwriting losses high) and hard markets, where companies face less competitive pressure on rates. Stochastic variables are used to generate the future loss frequency and inflation, which then impact both loss severity and loss development. The policy growth target is applied to a stochastic model of the underwriting cycle that determines the premium level needed to achieve the planned growth. In soft markets, premium levels need to be restrained or reduced in order to increase the market share. In hard markets, premium levels can be raised in line with policy costs and insurers can still grow in the market share. The interaction of the premium levels charged to achieve growth, loss reserve development, and loss costs based on the projected loss frequency and severity determines the underwriting profitability of the insurer, except for catastrophic losses.

Another module determines the catastrophic losses the insurer incurs each year. In the first step, the number of catastrophes (which is defined as industry wide losses in excess of \$25 million) is determined based on a Poisson distribution. A uniform distribution is used to establish a focal point (by state) for each catastrophe determined from the first step. This allocation is based on historical patterns for natural disasters. Once the state that is determined to be the focal point for a catastrophe is established,

the severity of the catastrophe is generated based on a lognormal distribution, and the allocation among states (the focal point state and neighboring states) is determined based on historical patterns. By repeating the process for each catastrophe determined in the first step, the total catastrophe losses by state for a given year are determined. The insurer is then assigned a share of these losses based on its market share in each state. These catastrophe losses are then incorporated into the underwriting module to determine the total loss experience of the insurer.

Another module reflects reinsurance agreements the insurer has in force. These contracts can be on a quota share basis, excess of loss, or catastrophe contracts. The details of the existing contracts are entered into the model, and the loss experience that is generated by the model is applied to those contracts to determine the loss experience, net of reinsurance. Reinsurance premiums and losses are then included in the underwriting experience of the insurer.

The other major source of income for an insurer is investments. The investment module uses stochastic variables to generate interest rates, inflation rates, and equity returns, which then are used to determine the investment income from interest and dividends, and the changes in asset values. The interest rate model used in Dynamo is a one-factor Cox–Ingersoll–Ross (CIR) term structure model. This generates nominal interest rates each month for the next 20 years, so that both short-term and long-term interest rates can be calculated. Inflation rates are determined from a second stochastic equation that reflects the correlation with interest rate levels. Equity returns are modeled on the basis of a separate stochastic model that is based on the expected risk premium adjusted for changes in interest rates, along with a random factor. An increase in interest rates of 100 basis points reduces the expected risk premium by 400 basis points. The one-factor CIR model is a mean reverting square root model, with the strength of the random factor proportional to the square root of interest rates, so negative interest rates do not occur. The one-factor model is sufficient for most property–liability insurance applications, as results are not heavily dependent on interest rate changes. Catastrophe losses and the underwriting cycle play a larger role than interest rates in the financial results for most property–liability insurers. Life insurers and banks tend to use two-factor term structure models, as interest rates and investment returns are the

most significant sources of uncertainty for many of these organizations. Other DFA models use different interest rate models. Some generate real interest rates by a term structure model and inflation separately, and then combine these to obtain nominal interest rates. Equity returns are also modeled in a variety of ways. One alternative method is to incorporate a regime switching model that shifts the parameters of the model for the risk premium based on Markov process, with one regime having a higher expected return and lower volatility and the second regime having a lower expected return and higher volatility.

The investment module generates investment income that reflects the investment allocation of the insurer, between equities and bonds, and the maturity distribution of the bond portfolio. This module also generates the market and statutory values for the assets. This module, combined with the underwriting, catastrophe, and reinsurance modules, produces aggregate operating results for the insurer. The final module incorporates taxation, based on the current tax rates and the investment allocation between taxable and tax-free bonds. When all the modules are combined, the total after tax operating statements and balance sheet for the insurer are calculated for the next 5 years. Each iteration requires over 1000 random variables, and tens of thousands of calculations. Each run of this program will perform hundreds of iterations in order to determine a probability distribution for future financial positions. This type of simulation could not be done prior to the development of powerful computers that perform these operations.

Sample output from a DFA model is shown below. In Figure 1 the year-end statutory surplus of an insurer is shown based on a run of 1000 iterations.

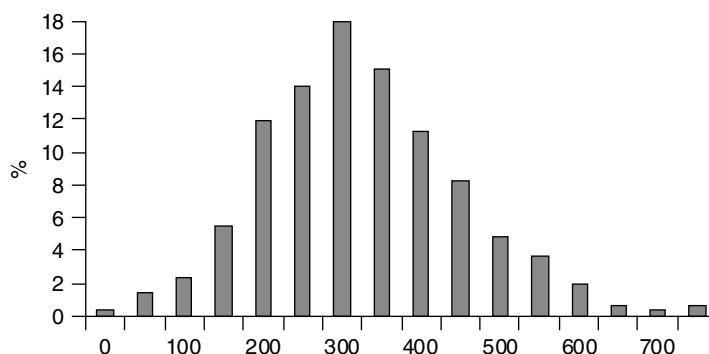


Figure 1 Distribution of year-end surplus

This information would be useful to determine the financial condition of the firm by considering the likelihood that the surplus falls below an acceptable level. In Figure 2, the results of the terminal value of a firm is shown for 500 iterations under different growth rates, ranging from 0 to 10%. Comparison of the distribution of the results can help an insurer determine the optimal growth policy.

Once a DFA model is run, a firm can determine the likelihood of its developing financial difficulties on the basis of its current operating conditions. For each iteration associated with financial difficulties, all the relevant variables can be captured and analyzed. For example, a firm could discover that most of the incidents of financial impairment are associated with unusually large catastrophe losses. This could lead a firm to revise its reinsurance contracts or to change the lines of business it writes, or the geographical areas in which it operates. An insurer may decide to stop writing property coverage in coastal areas based on its DFA model.

Although the results of most DFA applications are proprietary, one published DFA application determined the optimal premium growth rate for an insurer [3]. Rapid growth can impair the value of an insurer, as premium levels may have to be reduced or expenses have to increased to attract significant numbers of new policies. In addition, information asymmetries cause the loss ratio on new business to exceed that on renewal business [4]. Alternatively, low levels of growth do not generate sufficient premium volume to increase the value of the firm. The publicly available DFA model is used to determine the growth rate that optimized firm value.

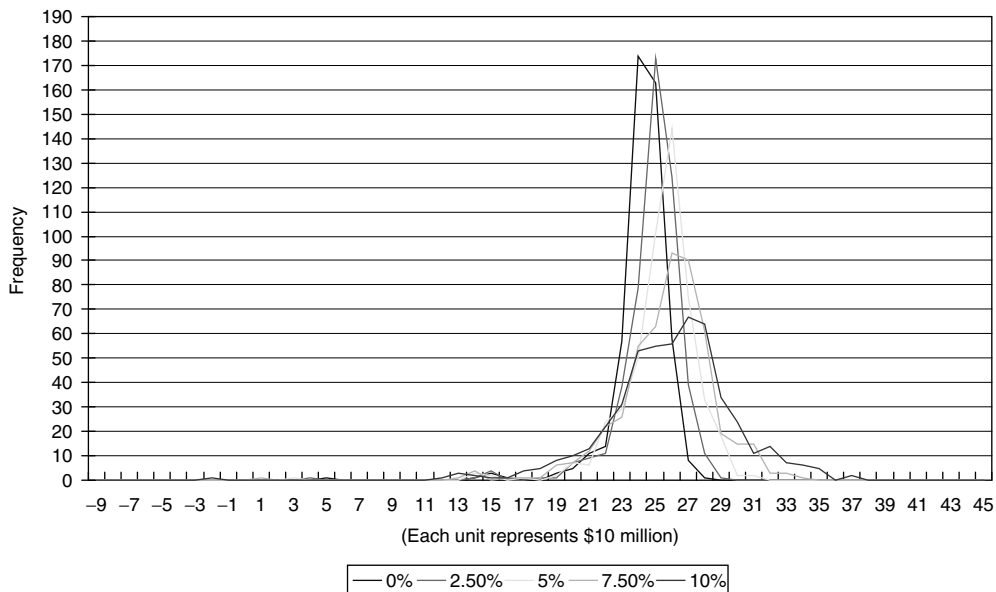


Figure 2 Firm values under different growth rates

Despite the power of an effective DFA model, care needs to be used in using these models. DFA, as any model, is a simplified representation of reality. Many factors that can influence results are not included in the models in order to focus on the primary factors. Thus, all the uncertainty is not reflected in the model; actual results will vary more widely than the model will indicate. Models are built on the basis of what has happened in the past, not on what new conditions can arise. Firms that do not recognize this inherent uncertainty, and accept levels of leverage that are based on the results of a DFA model, are more exposed to risk than they and their regulators realize. Also, situations can change so that factors not included in the model will become so important that they need to be added in the future. DFA is dynamic in the sense that it will continually evolve. Finally, some factors are simply too difficult to quantify, and therefore are not included in the model. An example would be the chance of management defrauding the firm, which is a significant factor in insurance insolvency, but would be extremely difficult to quantify accurately. What is the likelihood that a firm would be willing to assign to the chance that its management is committing fraud? This value would be quite low in most well run firms. In those more exposed to this risk, it is unlikely that the firm

would recognize or be willing to place a realistic level on this risk as a form of self-denial. In general, DFA can be useful as a guide for firms, but they must be used with a healthy dose of skepticism.

References

- [1] Cairns, A.J.G. (2004). *Interest Rate Models: An Introduction*, Princeton University Press.
- [2] Hardy, M. (2001). A regime-switching model of long-term stock returns, *North American Actuarial Journal* **5**(2), 41–53.
- [3] D'Arcy, S.P. & Gorvett, R.W. (2004). The use of dynamic financial analysis to determine whether an optimal growth rate exists for a property-liability insurer, *Journal of Risk and Insurance* **71**(4), 583–615.
- [4] Cohen, A. (2005). Asymmetric information and learning: evidence from the automobile insurance market, *Review of Economics and Statistics* **87**(2), 197–207.

Further Reading

- Babcock, N. & Issac, D.B. (2001). Beyond the frontier: using a DFA model to derive the cost of capital, *ASTIN Colloquium International Actuarial Association*, Washington, at http://www.actuaries.org/ASTIN/Colloquia/Washington/Isaac_Babcock.pdf.
- Canadian Institute of Actuaries (1999). *Dynamic Capital Adequacy Testing – Life And Property And Casualty*, Educational

- Note, Committee on Solvency Standards for Financial Institutions <http://www.actuaries.ca/publications/1999/9930e.pdf>.
- Casualty Actuarial Society Valuation and Financial Analysis Committee, Subcommittee on dynamic financial models (1995). *Dynamic Financial Models Of Property-Casualty Insurers*, Casualty Actuarial Society Forum.
- Casualty Actuarial Society Valuation and Financial Analysis Committee, Subcommittee on the DFA Handbook (1996). *CAS Dynamic Financial Analysis Handbook*, Casualty Actuarial Society Forum.
- Correnti, S., Sonlin, S.M. & Issac, D.B. (1998). Applying a DFA model to improve strategic business decisions, *CAS Dynamic Financial Analysis Call Paper Program*, Landover, Maryland, pp. 15–51.
- D’Arcy, S.P., Gorvett, R.W., Herbers, J.A. & Hettinger, T.E. (1997a). Building a dynamic financial analysis model that flies, *Contingencies* 9(6), 40–45.
- D’Arcy, S.P., Gorvett, R.W., Herbers, J.A., Hettinger, T.E., Lehmann, S.G. & Miller, M.J. (1997b). Building a public access PC-based DFA model, *CAS Dynamic Financial Analysis Call Paper Program*, Landover, Maryland, pp. 1–40.
- D’Arcy, S.P., Gorvett, R.W., Hettinger, T.E. & Walling, R.J. (1998). Using the public access dynamic financial analysis model: a case study, *CAS Dynamic Financial Analysis Call Paper Program*, Landover, Maryland, pp. 53–118.
- Hodes, T., Cummins, J.D., Phillips, R. & Feldblum, S. (1996). The financial modeling of property/casualty insurance companies, *CAS Dynamic Financial Analysis Call Paper Program*, Landover, Maryland, pp. 3–88.
- Walling, R.J., Hettinger, T.E., Emma, C.C. & Ackerman, S. (1999). Customizing the public access model using publicly available data, *CAS Dynamic Financial Analysis Call Paper Program*, Landover, Maryland, pp. 239–266.
- Wilkie, A.D. (1995). More on a stochastic model for actuarial use, *British Actuarial Journal* 1, 777–964.

STEPHEN P. D’ARCY

Early Warning Systems (EWSs) for Predicting Financial Crisis

Scientists have been working for decades on early warning systems (EWSs) to predict the occurrence of natural disasters; however, research on the application of EWSs in financial markets has gathered momentum only in the past decade. A commonly known triggering point that aroused such an interest was the financial crisis in Mexico, aka the Tequila crisis, in 1994. Research in this area was given a further boost during the aftermath of the Asian financial crises in 1997/1998. Natural disasters such as tsunami, can cause the loss of thousands of human lives and the financial damage caused cannot be underestimated. During the Asian financial crises, for example, the disruption to economic activities and hence the quality of life of those affected was substantial. This is why there has been such an immense interest in the development of EWSs to predict a financial crisis, so that the government can implement appropriate policies to prevent it, and financial market participants can take proper measures to minimize their losses.

The conceptual framework of EWSs for financial crises is quite simple. It builds on the premise that an economy and its financial markets would behave differently prior to an imminent financial crisis. The “abnormal” behavior has a systematic and recurrent pattern, which is discernible. Therefore, one can judge whether a crisis is about to occur from the movements of relevant economic and financial indicators.

Definition of a Financial Crisis

No EWS will work effectively without a clear definition of what a financial crisis is. Banking crises and currency crises (or exchange rate crises) are generally regarded as financial crises. These crises are not necessarily independent of each other. More often than not, they are interrelated. A banking crisis can easily lead to pressure on exchange rates. Recent research efforts on EWSs have been focusing on currency crises. A possible explanation for this phenomenon is

that banking crises are relatively easier to detect from prudential supervisory information. Banking regulators usually have ready access to financial information and prudential indicators, such as the balance sheet and liquidity ratio of individual banks. In addition, the consequence of a banking crisis is arguably easier to contain with the provision of deposit insurance and lender of last resort by central banks.

To begin with, every EWS requires a quantitative definition of a currency crisis. Currency crises are characterized by abrupt and intensive depreciation pressure, highly volatile interest rate movements, rapid depletion of foreign reserves, or a combination of these phenomena. Abdul Abiad [1] provides a comprehensive survey on how crisis is defined in EWSs. The paper outlines 30 different definitions adopted in recent research studies for a currency crisis. The definitions can be broadly divided into a few groups. First, some EWSs identify currency crises by looking at the extent and speed of depreciation. For example, the Emerging Markets Risk Indicator (EMRI) model developed by Credit Suisse First Boston defines a currency crisis as a depreciation exceeding 5% (or double that of the preceding month) within 1 month.

Another group of models looks at such indicators as exchange rate, interest, and the pace of reserve depletion. The models compare the current value of these variables with their respective value at tranquil periods in order to identify a currency crisis. For example, Kaminsky *et al.* [2] defined a currency crisis as the weighted average of 1-month changes in exchange rates and reserves more than three standard deviations above the country average during tranquil periods. The weighted average is converted into an exchange market pressure index or a speculative pressure index, which takes the form of a binary crisis variable. The index will take the value of 1 if the weighted average exceeds a certain threshold, and 0 otherwise. There are a number of modifications to the approach, such as incorporating “expert judgment” (*see Imprecise Reliability; Expert Judgment; Uncertainty Analysis and Dependence Modeling*) in the construction, focusing only on extreme crises that cause severe economic recession. There are weaknesses in the binary index approach to define a crisis. The main concerns are the loss of information during the transformation of a continuous variable into a binary one. The binary index cannot distinguish between cases with index values below the threshold.

In order to address these concerns, some studies use a continuous index or transform the index into a crisis score within a bounded range.

Choice of Indicators

Like most economic phenomena, financial crisis itself is a concept that is difficult to observe and measure directly. The selection of appropriate indicators is imperative to the success of developing useful EWSs. In general, high-frequency financial market data such as exchange rates and interest rates, provide up-to-date information on a real-time basis but noises are involved. On the other hand, macroeconomic data, such as growth, exports, and capital flows, contain less noise, but are less frequent and only available after a certain degree of time lag. Most studies on EWSs employed both financial market and economic data to construct their models. There are various means to determine the threshold levels for the indicators; see Chan and Wong [3] for choices relating to classification trees. A common and convenient way to do so is to use three standard deviations from the mean value during tranquil periods. Commonly used indicators can be classified into the following three categories.

Macroeconomic Indicator

Macroeconomic indicators are used to measure the robustness of economic activities, degree of external imbalance, adequacy of foreign exchange reserves, and ease of monetary and credit condition. Common indicators under this categories include deviation of exchange rate from trend; current account balance (usually expressed as a ratio of GDP); export growth, ratio of money supply to official foreign exchange reserves; growth of domestic credit and real interest rates.

Capital Flow Indicator

Capital flow indicators measure the degree of lending boom, extent of short-term debt, and composition of capital flows. Common indicators in this category include growth of assets of the banking sector, ratio of short-term external debt to official foreign exchange reserves, amount of portfolio flows, and changes in international investment position.

Financial Market Data

This category of data focuses mainly on the soundness of the banking system. Indicators commonly used are: capital adequacy ratio, loan to deposit ratio, as well as growth in bank deposits and money supply.

Key Modeling Approaches to Early Warning Systems

Since the outbreak of the Asian financial crises in 1997, the International Monetary Fund (IMF) has implemented several models of EWSs to monitor the occurrence of financial crisis. The two most extensive and well-documented approaches are the signal approach and the limited dependent regression approach. These models are outlined below.

The signal approach compares a set of financial market and economic variables, in the period prior to the crises, with their respective values during tranquil periods. A warning signal will be issued whenever there is a large deviation between the extreme values. Under the signal approach, it is necessary to develop objectives and systematic means to identify extreme values (*see Large Insurance Losses Distributions; Extreme Value Theory in Finance; Mathematics of Risk and Reliability: A Select History*) and to decide how many extreme values amongst the set of variables would be needed in order to infer that a crisis may be emerging. The first issue can be addressed by setting the threshold level for each of the variables that gives an optimal balance between the probability of missing a crisis and that of giving out false signals. In practice, this can be done by choosing the thresholds that minimize the noise-to-single ratio. The second issue can be addressed by the construction of an index that takes the weighted average of individual signals.

Kaminsky *et al.* [2] provide a well-documented study on EWSs using the signal approach. The study, commonly known as the (Kaminsky, Lizondo, and Reinhart) *KLR model*, employed data of 20 countries, between 1975 and 1990. Both developing and industrial countries were included in the sample. In the study, a total of 15 economic and financial variables were studied. Values of these variables were compared to their respective levels in a 24-month period prior to the crises in tranquil periods. The study suggested that the top five variables that have

the strongest explanatory power are: (a) deviation of real exchange rate from a deterministic trend; (b) the occurrence of a banking crisis; (c) rate of export growth; (d) equity price movement; and (e) the ratio of money supply (M2) to reserve.

The second approach to EWSs is the use of limited dependent regression models. Similar to the signal approach, this approach generally models the financial crisis as a binary variable, taking a value of either 0 (noncrisis period) or 1 (crisis period). Explanatory variables are selected to undergo a multivariate probit or logit regression (*see Dose-Response Analysis; Credit Scoring via Altman Z-Score; Logistic Regression*) to provide a probability of the occurrence of a crisis within a specific time window. Advocates of this approach argue that it has several advantages over the signal approach. First, it is easy to interpret the model results because the prediction results are always expressed as a simple probability. Secondly, it takes into account the importance of all explanatory variables simultaneously and the usefulness and relevance of any additional explanatory variable can be easily assessed.

Berg *et al.* [4] used the probit model to develop an EWS using the same crisis definition and prediction horizon as the KLR model. The model is commonly known as the *Developing Country Studies Division (DCSD) model*. It is named after the DCSD of the IMF, in which the model was first formulated. The explanatory variables included in the model were deviation of real exchange rate from trend, the ratio of current account balance to GDP, export growth, growth of foreign exchange reserves, and the ratio of short-term debt to foreign exchange reserves.

Mulder *et al.* [5] introduced corporate sector balance sheet variables to the models developed by Berg and Pattillo and found that variables, such as degree of leveraged financing, ratio of short-term debt to working capital, balance sheet indicators of banks, corporate debt to foreign banks as a ratio of exports, and some measures of shareholders rights are useful in the prediction of financial crisis. However, since many corporate sector data are only available once a year and with considerable time lag, this approach may not be able to capture changes in the highly volatile financial markets in a timely manner.

There are many modifications to the limited dependent regression approach. The modifications range from using different geographical coverage,

time periods, and frequency of data and the introduction of new explanatory variables. For example, Weller [6] considered the degree of financial liberalization; Grier and Geier [7] considered exchange rate regime; and Eliasson and Kreuter [8] developed a continuous, rather than a binary measure for financial crisis.

A third approach to EWS is the Markov-switching models (*see Dynamic Financial Analysis; Bayesian Statistics in Quantitative Risk Assessment*). This approach assumes that an economy has two states, namely, tranquil periods and crisis periods. The two states can only be indirectly observed through certain financial market variables, such as exchange rates, that display very distinct behavior in these two states. The approach tries to predict the probability of an economy moving from the tranquil state to the crisis state at a given period of time on the basis of the strength or weakness of an economy's fundamentals. Abdul Abiad [1] developed such a model and found that it correctly predicted two-thirds of crises in sample.

Hawkins and Klau [9] offered another approach to EWSs. Instead of building econometric models to predict financial crisis, they developed the vulnerability indicator approach. They systematically presented various indicators of external and banking sector vulnerability by transforming the value of each indicator into a discrete score within a specific range. An indicator with a higher score suggests greater vulnerability. The scores of individual indicators are summed together with equal weights to achieve an overall vulnerability score that ranges between -10 and 10.

How Well Do Early Warning Systems Work?

To be sure, none of the EWSs developed so far gives overwhelmingly accurate crisis prediction. Crises were missed and false alarms were issued. Berg *et al.* [4] provided a systematic assessment on the effectiveness of a variety of early warning models developed prior to the Asian financial crisis in 1997. In particular, the study found that the KLR model gave fairly good forecasts. Many of the hardest hit economies during 1997/1998, such as Korea and Thailand were successfully identified by the model, although economies that were not so badly hit also appeared at the top of the prediction. The study

concluded that the model was statistically significant and informatively superior to random guessing.

Berg *et al.* [10] revisited the issue and compared the predictive power of the KLR and DCSD models with various nonmodel-based indicators, such as yield spreads and agency rating, between January 1999 and December 2000. The forecasts of the models in question were all “out-of-sample” forecasts. The results of the study were mixed. While the KLR model provided good out-of-sample prediction during the period, the accuracy of the out-of-sample forecast produced by the DCSD model deteriorated noticeably compared with that during the in-sample period. One possible explanation for the mixed results is that the occurrence of financial crises has been limited to a few cases. For example, during the period under assessment, there were only 8 crises out of a sample of over 500 observations. Despite the findings, Berg concluded that EWSs do provide a systematic, consistent, and unbiased means to process and analyze information pertinent to financial crises. While EWSs are not sufficiently robust to be relied upon as the sole tool to predict financial crises, the study affirmed that the best EWSs performed substantially better than non-model-based forecasts over the Asian financial crisis period.

Alternative Approaches to Early Warning System

Given the limitation to the predictive power of EWSs, there are other new frameworks developed to supplement such systems. One of these is the resilient framework that measures the resilience level of an economy. The framework does not attempt to predict financial crisis. It is an assessment of the soundness of economic and financial systems at a particular point of time. Chan and Wong [3] developed such a framework using a two-stage data feedback data mining approach involving the combination of a fuzzy logic framework (*see Imprecise Reliability; Expert Elicitation for Risk Assessment*) and the classification and regression tree (CART) technique to allow users to examine the resilience level of an economy. In their study, economies are classified into five different resilient levels between 1 and 5. The fuzzy logic framework evaluates expert opinions on a set of economic and financial market variables that are similar to those employed in other EWSs, and generates a

fuzzy logic score (from 1 to 5) for each observation. The CART technique then generates decision trees to maximize the likelihood that observations were classified into groups that were originally assigned by the fuzzy scoring system. The end product of the CART process is a decision tree that can be used to generate resilience scores for future observations. Back testing of the resilience framework suggested that it successfully detects substantial deterioration of resilience level of crisis-hit economies.

Conclusion

EWSs provide systematic and objective analyses of macroeconomic and financial market data in predicting financial crisis. This paper reviewed various approaches to the modeling of EWSs for financial crisis and discussed the effectiveness of these systems. Notwithstanding the research efforts and the noticeable progress in this area, there is still yet a sufficiently robust EWS that can be relied upon as the sole system for crisis prediction. Development in other research frameworks that are complementary to EWS such as the resilience framework, is encouraging.

References

- [1] Abiad, A. (2003). *Early Warning Systems: A Survey and a Regime-Switching Approach*, International Monetary Fund Working Paper No. 03/32.
- [2] Kaminsky, G.L., Lizondo, S. & Reinhart, C.M. (1998). *Leading Indicators of Currency Crises*, International Monetary Fund Staff Paper, Vol. 45, Issue 1.
- [3] Chan, N.H. & Wong, H.Y. (2007). Data mining of resilience indicators, *IIE Transactions* **39**, 617–627.
- [4] Berg, A. & Pattillo, C. (1999). *Are Currency Crises Predictable? A Test*, International Monetary Fund Staff Paper, Vol. 46, Issue 2.
- [5] Mulder, C., Perelli, R. & Rocha, M. (2002). *The Role of Corporate, Legal and Macroeconomic Balance Sheet Indicators in Crisis Detection and Prevention*, International Monetary Fund Working Paper WP/02/59.
- [6] Weller, C.E. (2001). *Currency Crises Models for Emerging Markets*, De Nederlandsche Bank Staff Report No. 45.
- [7] Grier, K.B. & Robin, M.G. (2001). Exchange rate regimes and the cross-country distribution of the 1997 financial crisis, *Economic Inquiry* **39**(1), 139–148.
- [8] Eliasson, A.-C. & Kreuter, C. (2001). *On Currency Crisis Models: A Continuous Crisis Definition*, Deutsche

-
- Bank Research Quantitative Analysis Unpublished Working Paper.
- [9] Hawkins, J. & Klau, M. (2000). *Measuring Potential Vulnerabilities in Emerging Market Economics*, Bank for International Settlements Working Paper No. 91, October.
- [10] Berg, A., Borensztein, E. & Pattillo, C. (2004). *Assessing Early Warning Systems: How Have They Worked in Practice?* International Monetary Fund Working Paper No. 04/52.

SUNNY W.S. YUNG

Ecological Risk Assessment

Ecological risk assessment (ERA) is a generic term used to describe any formal process whereby ecological threats are identified, their likelihood of occurrence estimated or guessed, and their consequences articulated. ERA is a subset of environmental risk assessment (*see* **Environmental Hazard; Environmental Health Risk; Environmental Risks**). It focuses specifically on the elicitation, quantification, communication, and management of risks to the *biotic* environment. While environmental risk assessment dates back to the 1930s [1], ERA is relatively new and commenced as a United States Environmental Protection Agency (USEPA) project in the 1980s to develop tools for environmental regulation and management [2]. Since that time there have been rapid advances in the sophistication and complexity of ERA “tools” although, as noted by Kookana *et al.* [3], ERA currently suffers from a poor understanding of processes governing ecological risks and a paucity of appropriate data. Other challenges exist as well, including the following:

- difficulties in identifying pathways for chemicals in the environment;
- lack of understanding of fate and effect of pollutants in the receiving environment;
- high levels of uncertainty in models and processes underpinning the ERA process;
- unquantified errors in model outputs;
- no agreed approach to combining uncertainties of different kinds comprehensively in a single analysis;
- lack of standard approaches to ERA;
- difficulty in developing and applying a “system-wide” ERA – that is quantifying the overall risk associated with multiple stressors and threats;
- diffuse linkages between the outcomes of an ERA and management responses;
- difficulties in assessing the utility of an individual ERA.

With respect to the last point, Suter [2] notes that “assessors have developed methods for determining the likelihood that a safe exposure level will be

exceeded, but have seldom specified the benefits of avoiding that exceedence”. Thus, a certain degree of “evaluation bias” tends to characterize ERAs where the risk of an undesirable ecological outcome is rarely evaluated against the benefit of a desirable ecological outcome. In many instances, the framing and the contextual setting of the problem at hand influence the outcomes of the ERA [4]. To a large extent, they determine the set of solutions considered, as well as the focus of the assessment, the kinds of endpoints used, the time frames considered, the data collected, and who is considered to be a “stakeholder”.

Components of an Ecological Risk Assessment

As with environmental risk assessment, ERAs have been dogged with multiple definitions of risk (*see* **Absolute Risk Reduction**), a confused lexicon, and a lack of a transparent and consistent framework to guide the ERA process [5]. Most environmental decisions are set in socially charged contexts. People stand to gain or lose substantially. Arguments are clouded by linguistic ambiguity, vagueness, and underspecificity to which analysts themselves are susceptible. Prejudice gets in the way of constructive discussion. A transparent framework helps to relieve these impediments.

We use “risk” here to denote the chance, within a prescribed time frame, of an adverse event with specific (usually negative) consequences. Other terms commonly used in ERAs are hazard and stressor. A *hazard* (*see* **Hazard and Hazard Ratio**) is a situation or event that could lead to harm [6]. Ecological hazards can be natural (e.g., cyclones, earthquakes, fires) or related to human activities (e.g., destruction of a habitat). Hazards are possibilities, without probabilities. They are all those things that might happen, without saying how likely they are to happen [4]. Stressors are the elements of the system that precipitate an unwanted outcome (for example, low dissolved oxygen in a river is a stressor that ultimately results in the death of aquatic life). Suter [7] created a system of thinking to help people to define environmental hazards and their consequences. He defined endpoints as an expression of the values that we want to protect. There are three broad kinds:

1. *Management goals* are statements that embody broad objectives, things such as clean water or a healthy ecosystem. They are defined in terms of goals that are both ambiguous and vague but they carry with them a clear social mandate.
2. *Assessment endpoints* translate the management goals into a conceptual model, and satisfy social objectives. Clean water may be water that can be consumed and bathed in by people. A healthy ecosystem may be one in which all ecological stages are represented, all natural ecological processes continue to operate, and populations of important plants and animals persist. But assessment endpoints cannot be measured.
3. *Measurement endpoints* are things that we can actually measure. They are operational definitions of assessment endpoints that are in turn, conceptual representations of management goals. Thus, measurement endpoints for freshwater may include counts of *Escherichia coli* or the concentration of salt. Measurement endpoints of a healthy ecosystem may be the abundance of several important species (threatened species or game species), and the prevalence of diseases and invasive species.

We have defined risk to be the chance or likelihood of an adverse outcome. Probability is a mathematical construct (a metric) that quantifies the likelihood of uncertain events. In this context, the duality between risk and probability is apparent and these terms are often used interchangeably in ERAs. The “frequentist” definition of probability is based on the statistical frequency (or relative frequency) with which an event is expected to occur. The term *subjective probability* also has two meanings. The first meaning is a lack of knowledge about a process or bias. The second meaning is that it indicates purely personal degrees of belief. Personal beliefs are unknown only insofar as a person does not know his/her own mind [8] (see **Bayes’ Theorem and Updating of Belief; Bayesian Statistics in Quantitative Risk Assessment**).

Standard approaches to risk analysis (see **Volatility Modeling**) may be particularly vulnerable to psychological frailties including insensitivity to sample size, overconfidence, judgment bias, anchoring (the tendency to provide subjective assessments similar to those already proposed, or proposed by a

dominant individual in a group), and arbitrary risk tolerance (see reviews by Fischhoff [9], Morgan *et al.* [10], and Freudenburg *et al.* [11]). In addition, scientific training fails to acknowledge the pervasive presence and role of linguistic uncertainty [12], particularly the vague, underspecified, and ambiguous language that characterizes many risk assessments.

Challenges for Ecological Risk Assessment

It is challenging to consider the full extent of uncertainty present in any analysis, to characterize it fully and carry the uncertainties through chains of calculations and logic. A host of new methods offer prospects for solutions; in addition to standard treatments such as probability trees (see **Decision Trees**) and *Monte Carlo*, emerging approaches include fuzzy numbers, rough sets, evidence theory, imprecise probabilities, probability bounds, game theory, and *decision analysis* (see **Decision Analysis**). The task lies ahead to evaluate these methods and develop experience in their use so that they can be applied effectively and routinely.

Adams [13] argued that risk assessments always involve decisions about values and preferences, and are colored by the personal experiences and prospects of the individuals conducting the assessments. He argued that, in general, we get by with crude abstractions shaped by belief. This view objects to the artificial separation of stakeholder, risk analyst, and manager/decision maker [14]. Instead of using technical analysis, risk assessments could be conducted through stakeholder engagement, elicitation of preferences and values, and consensus building. Adams may be right. Certainly, the importance of psychology and context provide strong support. The answers generated by quantitative risk analysts may be little more than smoke and mirrors, reflecting the personal prejudices and stakes of those conducting the analysis. It is likely that at least some of the problems alluded to by Adams will affect all risk analyses. The extent to which they are felt will depend on the nature of the problem, the amount and quality of data and understanding, the personal outcomes for those involved in the analysis, and the degree to which their predispositions can be made apparent.

Many disagreements among stakeholders are resolved by seeing clearly what the other participants want, and why they want it. Risk assessments that

combine social preferences with formal analytical tools can have their greatest utility in meeting these challenges. They work when they are logically robust and relatively free from linguistic ambiguity. They are not necessarily any closer to the truth than purely subjective evaluations. But they have the potential, if properly managed, to communicate all the dimensions of an ecological problem to all participants. They may do so in a way that is internally consistent and transparent, serving the needs of communication (assuming appropriate skills in the analyst).

ERA is relatively new and as such is still “finding its feet”. While it is acknowledged that more work needs to be done in harmonizing different quantitative approaches, developing a consistent lexicon, and producing robust guidelines, natural resource managers and environmental stakeholders have much to gain from the formalized approach to environmental decision making under uncertainty.

References

- [1] Eduljee, G.H. (2000). Trends in risk assessment and risk management, *Science of the Total Environment* **249**, 13–23.
- [2] Suter, G.W. (2006). Ecological risk assessment and ecological epidemiology for contaminated sites, *Human and Ecological Risk Assessment* **12**, 31–38.
- [3] Kookana, R., Correll, R. & Barnes, M. (2006). Ecological risk assessment for terrestrial ecosystems: the summary of discussions and recommendations from the Adelaide workshop (April 2004), *Human and Ecological Risk Assessment* **12**, 130–138.
- [4] Burgman, M.A. (2005). *Risks and Decisions for Conservation and Environmental Management*, Cambridge University Press, Cambridge.
- [5] Fox, D.R. (2006). Statistical issues in ecological risk assessment, *Human and Ecological Risk Assessment* **12**, 120–129.
- [6] The Royal Society (1983). *Risk Assessment: Report of a Royal Society Study Group*, London.
- [7] Suter, G.W. (1993). *Ecological Risk Assessment*, Lewis, Boca Raton.
- [8] Hacking, I. (1975). *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*, Cambridge University Press, London.
- [9] Fischhoff, B. (1995). Risk perception and communication unplugged: twenty years of progress, *Risk Analysis* **15**, 137–145.
- [10] Morgan, M.G., Fischhoff, B., Lave, L. & Fischbeck, P. (1996). A proposal for ranking risk within Federal agencies, in *Comparing Environmental Risks*, J.C. Davies, ed, Resources for the Future, Washington, DC, pp. 111–147.
- [11] Freudenburg, W.R., Coleman, C.-L., Gonzales, J. & Helgeland, C. (1996). Media coverage of hazard events: analyzing the assumptions, *Risk Analysis* **16**, 31–42.
- [12] Regan, H.M., Colyvan, M. & Burgman, M.A. (2002). A taxonomy and treatment of uncertainty for ecology and conservation biology, *Ecological Applications* **12**, 618–628.
- [13] Adams, J. (1995). *Risk*, UCL Press, London.
- [14] Kammen, D.M. & Hassenzahl, D.M. (1999). *Should We Risk It? Exploring Environmental, Health, and Technological Problem Solving*, Princeton University Press, Princeton.

Related Articles

Axiomatic Models of Perceived Risk

Evaluation of Risk Communication Efforts

History and Examples of Environmental Justice Risk and the Media

DAVID R. FOX AND MARK BURGMAN

Economic Criteria for Setting Environmental Standards

The development and choice of standards or guidelines to protect the environment and human health from pollution and other hazards can be achieved by directly controlling emissions of pollutants, by allowing some economic optimization based on legal standards, or a combination of both. As always, the justification for a choice of control system depends on the regulatory context. That is, specific statutes will direct an agency to perform some type of economic analysis, but not permit another to justify its choice of technology and then impose it on the producers of the hazard. For example, a statute may explicitly ask that a standard be set on the balancing of the risk, costs, and benefit; another may limit those analyses to risks only.

Economic costs include the value of all impacts that can be stated in monetary units, that is, these factors are *monetized*. However, some costs and benefits cannot easily be monetized because they are intangible, or not priced by the market, directly or indirectly. For example, the value of improving visibility by reducing air pollution (*see* **Air Pollution Risk**) is an environmental service not generally priced by the market, unlike the value of reductions in morbidity or mortality from the same type of pollution. Nonetheless, economic methods can be used to approximate those intangible costs (e.g., *via* methods such as hedonic pricing) so that the price paid by the consumer correctly *internalizes* (incorporates) all costs, for given levels of social benefit, resulting from a reduction in pollution.

As always, regulatory laws must consider economics as well as the full aspect of regulation: statutes, secondary legislation, and judicial review *via* case law. In the United States, costs have been found by the courts to be important, but not likely to limit the use of expensive technology, unless those costs were disproportionate to the benefits achieved by the selected control technology. The US Federal Water Pollution Control Act imposes a limited form of cost-benefit analysis (CBA) for water pollution control (*see* **Water Pollution Risk**) depending on the level of effluent control (33 U.S.C. Section 1251

et seq.). Marginal cost (the cost per unit of mass of pollution removed) analysis of control technology plays an important role in determining the level of effluent discharge permitted. The case *Chemical Manufacturers Association v. EPA* (870 F.2d 177 (5th Cir. 1989)), which dealt with changes in the marginal cost of controlling effluent discharges, exemplifies how cost-effectiveness can be used in risk-based regulations. The US EPA can impose a standard, unless the reduction in effluent is *wholly out of proportion* to the cost of achieving it. In the US case *Portland Cement Assoc. v. Ruckelshaus*, the court held that Section 111 of the Clean Air Act (CAA) explicitly requires taking into consideration the *cost of achieving emission reductions* with the *nonair quality health and environmental impacts and energy requirements* (486 F.2d 375, cert. den'd, 417 U.S. 921 (1973)). The court then found that the US EPA met the statutorily required consideration of the cost of its regulatory acts (*Sierra Club v. Costle*, 657 F.2d 298 (D. C. Cir. 1981) explains these issues).

We focus on three conceptually different ways to balance the cost and benefits associated with risky choices:

- *cost-effectiveness analysis* (CEA) and methods such as ALARA (as low as reasonably achievable) and ALARP (as low as reasonably practicable);
- *cost-benefit analysis*, CBA;
- *legal feasibility analysis*.

Cost-Effectiveness

A concern of CEA is the efficiency of removal of pollution (measured by appropriate physical units, such as micrograms of particles per cubic meter of air, $\mu\text{g m}^{-3}$) of alternative pollution control technologies that are added to physical facilities such as power plants, to maximize the reduction in the emissions of pollutants. A narrow view of economic efficiency – specific to pollution removal – drives this form of balancing. It consists of calculating the efficiency (in percentage) in pollution removal, capital costs, and variable costs such as those associated with operation and maintenance, for alternative pollution control technologies designed to achieve the same technical objective (such as removing a specific quantity of particulate matter, solids, from a process).

CEA is also used to study the economic gains made by substituting fuels that have a higher content of pollutants by those having a lower content (e.g., high versus low sulfur fuel oil). For example, the European Union (EU) REACH Program looks at the reductions in risk (the focus of Regulation, Evaluation and Authorisation of Chemicals (REACH)) on a per chemical basis with changes in cost because of changes at the manufacturing level. CEA can also be coupled to life cycle, and cradle-to-grave risk analysis. Specifically, CEA can be a basis for setting exposure limit values (e.g., in the EU Directive 2004/40/EC); these are as follows:

limits on exposure to electromagnetic fields which are based directly on established health effects and biological considerations ... that will ensure that workers exposed ... are protected against all known (short-term) adverse health effects.

The reduction in risk should be based on the general principle of prevention (Directive 89/391/EC).

In many risk assessments, the alternative choices of environmental control technologies to reduce pollution, and thus reduce risk, can be represented as depicted in Figure 1. Note that the efficiency of pollution removal falls between 0 and 100%. Figure 1 depicts three hypothetical technologies that can be used to reduce the emission of pollutants; each technology is not divisible and is mutually exclusive to the other. This economic evaluation is limited to engineering factors, such as thermodynamic efficiency, derating owing to use of energy for operating the pollution control technology (if the technology were not used more energy would be available for sale to the consumer), and so on. The change in the cost of a technological choice can be understood as follows. The change (generally,

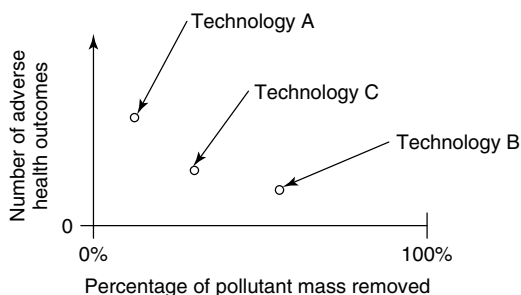


Figure 1 Three hypothetical, not divisible alternative environmental pollution control technologies

an increase) in capital and operating costs in going from the *best practicable control technology* (Technology A) to the *best available control technology* (Technology C), which may involve an intermediate technology, the *best conventional control technology* (Technology B), produces social benefits by reducing adverse outcomes or risk (however measured, possibly as number of deaths or any other factor for the various outcomes) by being more efficient in reducing the emission of pollutants.

Risk assessment yields an evaluation of the probability and magnitude of the consequences from exposure to one or more hazards and their severity. Severity of the consequences suggests that length of stay in hospital or at home from an injury or a disease, permanent damage, or death are some of the attributes that should be considered in measuring it. Accounting for severity results in an indicator composed of two components. The first is the number of life years lost (LYL), calculated by addition, over all relevant health endpoints, of the product of the number of deaths caused by a specific disease and life expectancy, at the age of death because of that disease EC [1]. The second is the years lived with disability (YLD), which is the sum of the product of the number of persons affected by a nonlethal disease, the duration of this disease, and a weight to account for its severity. The disability adjusted life years (DALY) is the summation of LYL and YLD. For example, the United Kingdom, UK, Treasury guidance on CEA [2], states that:

Cost-effectiveness analysis compares the costs of alternative ways of producing the same or similar outputs. One relevant form of CEA uses the “quality adjusted life year (QALY)” output measure.

The QALYs, according to the HM Treasury [2]

are estimated by assigning every life-year a weight on a scale where one represents full health and zero represents death. The most common methods of determining the health related utility values to weight QALYs are:

- the standard gamble: an individual is asked to consider two alternatives, full health, with a risk of death, and an imperfect state of health, with no risk of death;
- the time trade off: an individual is asked to consider how many years in full health would be equivalent to, say, five years in a given health state;
- the visual analogue scale: a thermometer type scale where full health is shown at the highest

point, and worst possible health state is shown at the lowest point. Individuals simply indicate where on the scale they feel that the health state is located; and,

- the person trade-off: individuals are asked what number of people being cured from one particular state is equal to, say, ten people being saved from death.

To clarify, we provide two of the examples developed by the UK HM Treasury [2] using QALYs and the *EQ-5D* (European Quality of Life 5 Dimensions) scale.

Example 1 Scores for the EQ-5D are given across 5 different dimensions: mobility; pain/discomfort; self-care; anxiety/depression; and the ability to carry out usual activities such as work, study, housework, and leisure pursuits. Each of these is scored out of 3 alternatives (e.g., for mobility – no problems; some problems walking about; confined to bed), giving a total possible combination of 243 health states (in addition, unconsciousness and death are included). Using the techniques above, patients are asked to self-score their health state and their views about their state of health relative to normal health (a score of 1) and death (a score of 0). Some examples of the health states (and their QALY values in brackets) are as follows:

- 11111 – no problems (QALY in such a state: value = 1.0);
- 11221 – no problems walking about; no problems with self-care; some problems with performing usual activities; moderate pain or discomfort; moderately anxious or depressed (0.760);
- 12321 – no problems walking about; some problems washing or dressing self; unable to perform usual activities; some pain or discomfort; not anxious or depressed (0.516).

Example 2 A safety measure might prevent approximately 100 people per year from suffering an average loss of 5 QALYs each, from chronic health effects. Thus: QALYs gained = $100 \times 5 = 500$ person-QALY/year. Using current benchmarks for QALYs at 18, an acceptable maximum cost of preventing this loss is $500 \times £30\,000 = £1.5$ million. Alternatively, this resource might be spent to prevent a single expected fatality per year, as it is close to the upper benchmark.

As Low as Reasonably Achievable, ALARA, or Practicable, ALARP

These specialized approaches relate economic cost to reductions in exposure either in occupational or

environmental settings. For example, the ALARA analysis “should be used in determining whether the use of respiratory protection is advisable” (US Nuclear Regulatory Commission [3]). The US Department of Energy (DOE), also bases some of their policies and procedures for safely dealing with exposure to ionizing radiation on ALARA (10 CFR 835.101). The ALARA principle is based on the (in some instances, assumed) proportionality between the cost and the effectiveness of safety measures being studied – costs are incurred to decrease the risk, given the alternatives being studied. Specifically, an ALARA for radioactive substances is \$1000/person-rem *averted*, which is taken to be the limit of cost-effective reduction in radiological exposure for the public [4]. It is *an upper limit on an individual’s permissible radiation dose*; here, *reasonably achievable* means *to the extent practicable* [3]. Importantly, there is no requirement that *radiation exposure must be kept to an absolute minimum*. Thus, the driving force is the combination of practicability of safety, which is a function of *the purpose of the job, the state of technology, the cost of averting dose, and the benefits* [3].

The United Kingdom has health and safety legislation requiring that risks be reduced “as low as is reasonably practicable” (ALARP). The

legal interpretation ... is that a risk reduction measure should be implemented, unless there is a ‘gross disproportion’ between the cost of a control measure and the benefits of the risk reduction that will be achieved.

In the United Kingdom, a legal definition of *ALARP* is given in *Edwards v National Coal Board* (1949):

‘Reasonably practicable’ is a narrower term than ‘physically possible’ ... a computation must be made by the owner in which the quantum of risk is placed on one scale and the sacrifice involved in the measures necessary for averting the risk (whether in money, time or trouble) is placed on the other, and that, if it be shown that there is a gross disproportion between them – the risk being insignificant in relation to the sacrifice – the defendants discharge the onus on them.

The ALARP ensures that the exposure of workers and the general public to risk is *as low as reasonably practicable*. One framework (below) has been devised to interpret this requirement [2]:

Furthermore, this test of reasonableness applies to all risk levels, although in practice regulators may develop decision boundaries, as Table 1 shows.

The EU has *proportionality* as one of its main guiding principles. This principle has affected UK and European laws by setting out a number of criteria for CEA: ALARA, BPM (best practicable means), BPEO (best practicable environmental option), and BAT (best available techniques). Each of these has its own characteristics, but all are based on the trade-off between risk and the cost of reducing it [2].

Viscusi [5] exemplifies ways in which a cost-effectiveness assessment can put into perspective government interventions using a common metric – the costs of saving lives. An abstract of these numbers (in 1984, \$US) is given in Table 2.

The National Highways Traffic Safety Administration (NHTSA), and the Occupational Safety and Health Administration (OSHA), save statistical lives at the costs depicted in Table 2. The range of *costs per life saved* between mandating passive restraints and formaldehyde exposure in the workplace is approximately \$72 000 million. Abelson

[6] summarized the cost of regulating pollution as approximately averaging \$115 billion (in \$1990). Using 1984 dollars means that all values expressed in dollars are transformed into constant 1984 dollars. 1984 is the *base-year*. Values in dollars in other periods, say 1986, are *nominal*. This transformation uses an index such as the *price index*.

Because CEA does not include all of the positive and negative social impacts of a technological choice, it is limited in scope, as Figure 1 depicts in terms of three hypothetical technologies and Figure 2 depicts in terms of a smooth cost function. To overcome this limitation, and to account for change in the social costs because of diverting societal-scarce resources while reducing risks, it is preferable to use CBA fully to assess the effect of changes in reducing the burden of pollution. The net result – addressed by CBA – involves the assessment of the damage from pollution balanced against the tangible and intangible costs to society because of that reduction, possibly accounting for uncertainty. Clearly, there is a trade-off in going from CEA to CBA – more costs and more time consuming, with the possible introduction of spurious relationships.

Table 1 Relationships between tolerability of risks and criterion for action

Category	Action required	Criteria
Intolerable	Extremely reluctant to accept any arguments for not doing more to reduce the risk	If for workers, a risk of death of 1 in 1000 per annum
Tolerable if as low as reasonably practicable	A case specific “ALARP” demonstration is required. The extent of demonstration should be proportionate to the level of risk	Risk levels broadly between “intolerable” and “broadly acceptable”
Broadly acceptable	No case specific demonstration is required. The “ALARP” demonstration should be fairly straightforward, assuming existing codes of practice etc. are up to date	For workers and the general public, if it is a risk of death of 1 in 1 000 000 per annum

Table 2 Selected cost per life saved (millions 1984 \$US, Viscusi, 1993) by technological choice^(a)

Action	Agency, year of regulation	Annual risk	Expected number of annual lives saved by regulation	Cost per life saved (millions 1984 \$)
Passive restraints/belts	NHTSA, 1984	9.1×10^{-5}	1850	0.30
Occupational benzene exposure	OSHA, 1987	8.8×10^{-4}	3.8	17.10
Occupational formaldehyde exposure	OSHA, 1987	6.8×10^{-4}	0.01	72000

^(a) Reproduced from [5]. © American Economic Association, 1996

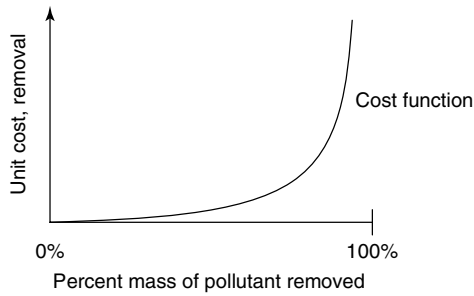


Figure 2 Hypothetical marginal cost function

Cost-Benefit Analysis (CBA)

CBA is the complete economic analysis of the impact of a standard (*see Cost-Effectiveness Analysis*). It thus subsumes risk analysis in the sense that risks – a form of social costs – are one of the many costs that society incurs when benefiting from development. CBA benefits from a number of economic subdisciplines that include microeconomics, public finance, financial management, welfare, labor, and environmental economics. Simply put, it is summarized by calculating the present discounted value (PDV), of the difference between the total *social* costs and the total *social* benefits of one or more actions, discounted at the appropriate social discount (or interest) rate. Each action has the same objective. The final choice of the best options is guided by the maximization of the net discounted benefits or by the minimization of the net social costs. A deterministic expression for CBA is derived from the formula for the future simple discounted value of an amount of money, A , which is held today, calculated for t uniform periods of time at a (constant) interest rate r (dimensionless) is given by

$$\text{Future Compounded Value, } FCV = A(1 + r)^t \quad (1)$$

The PDV is (A is changed to $(B_t - C_t)$) given by

$$PDV = \sum_t (B_t - C_t) / (1 + r)^t \quad (2)$$

The PDV formula provides today's ($t = 0$) monetary value of a discrete stream of net benefits, calculated as the difference between costs C , and benefits B , discounted at a rate r , over a number of periods of time (symbolized by $t, t = 1, t = 2, \dots$), beginning at time $t = 0$. If a single, lump sum value (a single

sum without prior payments) of the net benefits were sought, the formula simplifies to

$$PDV = [(B_t - C_t) / (1 + r)^t] \quad (3)$$

Example 3 Let the value of the benefit equal \$10 000.00 and the cost equal \$5 000.00, the number of yearly time periods is 10. Assume a constant interest rate of 5%. Today's value of the net benefit that will be generated only at the end of year 10 is $PDV = 5000 / (1 + 0.05)^{10} = \3069.57 . In this calculation the interest rate is *nominal*, which means that inflation and the risk of default is included in the 5%. The *real* interest rate might be about 2%.

The formula for rank-ordering potential actions through the CBA is as helpful as it can be deceptive, because of the way certain components in the formula can be manipulated to emphasize its results. Although most formulae can be manipulated, a clear vulnerability exists in that many environmental intangibles can only be approximated because they fall outside the more precise operations of the market. A second aspect concerns the choice of the discount rate. The decision rule is to choose the alternative that yields the largest positive PDV. The formula is not sensitive to monetary unit – using the Euro, dollar or pound does not affect the rank order of a choice. However, the choice of the base year for the monetary unit (e.g., 1980 \$ rather than current \$) can affect the rank order of the alternatives. Therefore, a constant dollar should be used in the calculations. A net *disbenefit* occurs when the difference between the costs and benefits is negative. If compounding occurred more frequently within a year, for example quarterly, the formula would have to be changed slightly to reflect that situation.

Example 4 The cumulated discounted value of an initial amount $(B - C)$, in which the interest rate per year is r , and k is the frequency of compounding per year for n years. A sum can be compounded k/r times per year as follows: $(\text{initial amount}) \times [(1 + r/k)^k]^n$. Consider the simple interest payments for \$1000 for 2 years at 10% per year: $1000(1 + 0.10)^2 = \$1210$. The effect of compounding interest four times over 2 years yields \$1216: $1000(1 + 0.05)^4 = \$1216.00$.

The process for obtaining a CBA can be summarized as a sequence of key steps (nothing

is said about the political viability of the preferred alternative):

Define the Objective → Identify Mutually Exhaustive and Complete Set of Alternatives (Options or Actions) to Reach that Objective → For Each Alternative, Measure Social Risk, Costs and Benefits → Calculate the Net Present Discounted Value (using the same discount rate) of Each Alternative → Rank Each Alternative According to the Magnitude of Its Net Present Discounted Value (if the technologies are divisible, the ranking can be based on the benefit to cost ratio (B/C)) → Performing Uncertainty Analysis to Assess the Dominance of the Preferred Alternative, Relative to other Alternatives → Recommend to Decision-Makers (and stakeholders) the Choice of the Preferred Option, Basing such Choice on the Maximum Net Present Discounted Value (or the B/C ratio with the highest value, when the projects are divisible).

Discounting occurs over the economic lifetime of the alternatives considered; it is the means to account for the effect of time on the value of money. The alternative yielding the highest positive difference ($B - C$) is optimal. The calculations include direct costs and benefits, such as employment and employment effects, as well as indirect costs and benefits, whether tangible or not. This type of analysis also includes anticompetitive effects, if the alternatives being studied cause such effects. Hence, distortions to the labor market, price-induced effects on wages, and so on, are part of CBA, if relevant to the specific application. For example, the US National Environmental Policy Act (NEPA, 42 U.S.C. Sections 431–4347), one of the earliest modern environmental protection acts (enacted in 1972), requires the balancing of the environmental costs of a project against its economic and environmental benefits. Although a numerical balancing is often involved, NEPA does not require the calculation of a numerical cost-benefit ratio or the computation of the net cost-benefit number for each alternative taken to mitigate an environmental impact. It should come as no surprise that CBA has generated many disputes between stakeholders such as environmental organizations, industry, and the US EPA.

A recent case exemplifies the issues. In 2001, the US Supreme Court affirmed that, under Section 109 of the CAA (amended, 42 U.S.C. Section 7409(a)), federal standard setting (for the National Ambient

Air Quality Standards (NAAQS)) for particulate matter and tropospheric ozone must be set *without* consideration of costs and benefits. The new ozone NAAQS was being lowered to 0.08 ppm, and for the first time the EPA was regulating 2.5 μm particulate matter (lowered from 10 μm). The sources of particulate matter being regulated were principally trucks and other diesel engines, and mining. The rationale for lowering the particulate matter standard was that it would reduce the number of premature deaths by 15 000 per year and serious respiratory problems in children by 250 000 per year. According to the US Chamber of Commerce, to implement these regulations it would cost industry approximately 46 billion US dollars per year. This case, decided by the US Supreme Court as *Whitman, Admin. of EPA, et al. v American Trucking Associations, Inc., et al.* (No. 99–1257, February 27, 2000) addresses four issues, two of which are critical to understanding setting of ambient environmental standards under Section 109(d)(1). This legislative mandate is the *attainment and maintenance ... requisite to protect the public health with an adequate margin of safety*. It is explicit. The Supreme Court held that *section 109(b) does not permit the Administrator to consider implementation costs in setting NAAQS*. *Whitman* stands for the proposition that the requirement by Congress must be clearly stated – if costs and benefits are considered, the CAA must specifically allow it. If it does not, it is most likely an agency cannot supplant what the Congress did not wish.

The legislative *adequate margin of safety* does not trigger CBA by itself. It only commands the Administrator of the EPA to adopt an unspecified degree of conservatism, as can be represented and achieved using factors of safety (numbers between 1 and 10, multiplied, and then applied to scale down the exposure to make it smaller and thus, in principle safer, for those at risk). However, not all sections of the CAA, are so stringent as to the level of conservatism (errring on the side of safety, when there are remarkable scientific uncertainties, as was held in *Union Electric Co., v EPA*, a US Supreme Court decision (427 U.S. 246 (1976))). This case illustrates the issue of having to consider economic factors (e.g., costs of compliance) when attempting to safeguard human health. Specifically, costs *are* excluded under Section 109, the section governing NAAQS, but risks are not (established by Congress under Section 109(b)(1)); as held in *Lead Industries Assoc., v EPA* decided

in a lower court (647 F.2d 1130, 1148 (CA DC, 1980). Even though the CAA seems to be very clear, nonetheless, the issue of costs was litigated all the way up to the US Supreme Court: the *Whitman* decision appears to settle this issue – the clear absence of a command to account for costs was not an omission by Congress. Section 109 is protective of human health and welfare. The American Trucking Association argued that imposing stringent air pollution standards would increase the risk to health by increasing unemployment because industry would have to lay off workers. The residual risks can be the increased number of suicides, violence at home, and other social problems such as those caused by unemployment. After reviewing the record from the early days of the CAA, from 1967 to February 27, 2001, the Court determined that Congress was quite explicit when it *wanted* to include any consideration of cost. It had done so in other cases involving water and air pollution. With regard to Section 109, the Court found that Congress (*Senate Report* No. 91–1196, pp. 2–3 (1970)) clearly meant to be protective and thus was only concerned with risks.

The health of the people is more important than the question of whether the early achievement of ambient air quality standards protective of health is technically feasible. ... (Thus) existing sources of pollutants either should meet the standard of the law or be closed down ...

The Supreme Court rejected the American Trucking Association's contention that the phrase *adequate margin of safety* does not provide an *intelligible principle* of protection, therefore being vague and thus unconstitutional. The CAA requirement is that the EPA must establish that

(for the relatively few pollutants regulated under Sect. 108) uniform national standards at a level that is requisite to protect public health from the adverse effects of pollutants in the ambient air.

The term *requisite* legally means *sufficient, but not lower or higher than necessary*. This is a key constitutional constraint that bounds the US EPA's ability to set standards: if it is exceeded, then the EPA has violated the Constitution because it now has *legislated* change. The Court held that Section 109(b)(1) of the CAA does not grant the US EPA legislative powers: an agency cannot use its own interpretative discretion to somehow gain what it cannot have. Moreover, the US EPA cannot arbitrarily construe

statutory language to nullify *textually applicable provisions* (of a statute) *to limit its discretion*. The US Constitution prohibits the Executive branch of the government from legislating. Legislation is the prerogative of the Congress of the United States. Justice Breyer agreed that the CAA is unambiguous on this point: costs are not included in the determination of the NAAQS. He concluded that *a rule likely to cause more harm to health than it prevents is not a rule that is requisite to protect human health*. Section 109(a) does not preclude comparative risk assessments. An even more complex analysis extends the domain of CBA to include industry-wide effects: it was used by the US Supreme Court in a case concerned with occupational exposure to airborne benzene (a carcinogen, discussed in the next section).

Legal Feasibility Analysis

Legal feasibility analysis should not be confused with the more traditional feasibility analysis that is an initial stage in the investigation of the feasibility of a project and thus entails a bounding of costs before a more comprehensive engineering analysis is undertaken. *Feasibility analysis*, in the present context, determines the extent to which some firms in an industry can fail, when that industry, as a whole, is not threatened. The phrase *to the extent feasible* was held to mean *capable of being done* by the US Supreme Court, when it found that Congress had carried out the appropriate cost-benefit balancing. This fact excluded an agency of the government from doing such balancing, particularly when Congress intended to place *the benefit of the workers' health above all other considerations, save those making the attainment of the benefit unachievable* (*American Textile Manufacturers Assoc. v Donovan, U. S. L. Week* 49: 4720 (1981)). The federal agency involved in this dispute was the federal OSHA, which had estimated the cost of compliance with the new standard, determined that such costs would not seriously burden the textile industry and would maintain *long-term profitability and competitiveness* of the industry [7]. The relevant section of the Occupational Health and Safety Act (29 U.S.C. Sections 655(b) </(5) and 652(8)), states as follows:

The Secretary ... shall set the standard which most adequately assures, to the extent feasible, on the basis of the best available evidence, that no

employee will suffer material impairment of health or functional capacity even if such employee has regular exposure to the hazard dealt with by such standard for the period of his working life.

Although not called *feasibility analysis*, there are parallels to this line of legal reasoning in the EU, where an important limitation to regulations is illustrated by the annulment of the EU's first comprehensive ban on direct and indirect tobacco advertising and sponsorship (*via* Directive 98/43/EC) by the European Court of Justice (ECJ) in 2000, in a case that saw the German Government and four British tobacco companies sue the European Commission. The judgment by the Court held that the Directive 98/43/EC was disproportionate to its means, principally because it did not harmonize competition but tended to destroy it. The restrictions the Directive placed on trade in tobacco were seen as disproportionate to those needed to ensure the proper functioning of the internal EU market [8].

Bubble, Trading of Emissions, and Other Instruments

Traditionally, environmental pollution standards are imposed by law and are monitored according to specific protocols that are also legally enacted and enforced. This form of environmental protection has been called *command and control*. Typically, an environmental statute sets principles, legal tests, and other forms of guidance that are then used by a regulatory agency, or an authority of the government, to set pollution standards. Standards are then set through those regulations (secondary legislation) uniformly for a country or region, unless there are specific regional requirements that result in special standards for that region. Although this is a simplification, as those familiar with special provisions of the law for state implementation plans know, it is a basis for command and control regulatory standard setting. An example is the US National Primary (protecting human health) and Secondary (protecting other welfare aspects such as visibility, vegetation, and so on) Air Quality Standards for sulfur dioxide, particulate matter, carbon monoxide, ozone, nitrogen dioxide, and lead that were promulgated, under authority of the CAA, by the US EPA. These standards apply to specific *stationary* sources of pollution, such as fossil-fuels burning power plants; they are numerical

and include suitable averaging times, (e.g., for SO₂, the US federal primary standard is 80 µg m⁻³ of air, annual arithmetic mean). The standards have been set at the level appropriate to protect sensitive individuals and not just the average member of the US population. Primary Standards were set without the CAA explicitly demanding that the US EPA should account for the costs of regulation before issuing the standards. Command and control regulatory framework is often associated with forcing new pollution technology, such as *the best available control technology* to achieve requisite level of emissions. Achieving the concentration of pollutants set forth in those standards has resulted in complex relations between the states and the federal government. These, in turn, have resulted in the creation of *State Implementation Plans, New Source Performance Standards, Prevention of Significant Deterioration, Lowest Achievable Emission Rates, Reasonably Available Control Technology*, and others [9].

Legislators have come to realize that profits are either preferable to command and control regulatory standards alone or can help minimize costs by using economic instruments in a portfolio of options for the producers of pollution. For example, the US EPA has been working with economic incentives, such as the offset program, since the 1970s. The *offset program* requires that a new stationary source operating in a region that had not met the NAAQS would have to purchase a sufficient amount of emission reduction instruments to offset its contributions of pollution to the region. The *bubble* policy, an early approach in the late 1970s to regulating air emissions, consists of an emission bubble that is placed over several sources of air pollution; the contributions from individual sources can be modulated to meet permitted overall emission levels that can be achieved by accounting for different efficiencies of the individual contributors to that overall bubble. In the EU, the use of economic instruments can complement standard-based command and control methods. A Member State can set targets, but emitters can allocate pollution emissions among themselves and decide on how to meet those targets. In principle, the correct price of noncompliance equates with the government's desired level abatement. The obvious advantage is that sources with lower compliance costs can overcomply and receive economic benefits from those with higher compliance costs; the net result is a

lower net cost to society, compared to command and control measures.

Emission trading and banking is another relatively recent addition to the portfolio of economic instruments used to reduce the emission of pollutants. Emission banking produces pollution credits. These form when there is a gap between the environmental standard and the emissions of pollutants from a firm, if those emissions are below the regulatory standard. Firms can use these credits as if they were currency – they can be banked, transferred, sold, or auctioned under the supervision of the environmental agencies. The environmental standards cannot be exceeded; consequently, the gain, relative to the more uniform command and control system, is that economic efficiency is allowed to work at the level of the firm, stack, or other levels, to minimize pollution and costs and thus generate aggregate societal benefits. The reason for allowing pollution trading through a government-regulated permit system is that those permitted to do so are driven by the market to reduce pollution, increase the efficiency of production, and opt for local or regional flexible control strategies. Specifically, market-based instruments such as pollution permits can accurately account for the different marginal costs of production, technological innovations, and fuel use. The trading system is as follows: a company reduces its emissions below a legal level and thus accrues emission credits that can be used elsewhere or be traded. The economic incentives generated by the trading system contribute to minimizing pollution and increasing the efficiency of production.

Trading is generally based on the sale of pollution permits. This is an activity that, in the United States, is controlled at the state level and can then be left to the market operation, provided that reporting and legally binding standards are met. For example, the CAA, Title I, Section 110(a)(20(A)) allows the states to use such economic instruments as *fees, marketable permits, and the auction of emission rights* to provide incentives, while maintaining competition and economic efficiency. The CAA's Title IV allows electric utilities specifically to trade sulfur dioxide emission allowances, under an emission cap, if emission is below that cap. If the cap is exceeded, the utility doing so loses the allowances and can be fined at levels that are several-fold larger than the

marginal cost of compliance. The statute places monitoring and reporting requirements under the jurisdiction of the US EPA. Under the CAA, economic instruments designed to provide incentives to reduce emissions can be combined. This has been done for acid rain resulting from emission from power generation by burning fossil fuels, principally coal, at the national level. The cap on the total emissions was such that it would result in sulfur dioxide reductions of 10 million tons from the 1980 levels, in 2010. The allowances are based on one ton of sulfur emitted, with the permit system envisioned to operate till the year 2030. The trading of acid rain allowances began in 1993; it included auctions, spot, and advance allowances. The proceeds from auctions are distributed among those from whom allowances were withheld. Trading and caps have also been combined at the state level. In California, the South Coast Air Quality Management District of Los Angeles administers the Regional Clean Air Incentives Market (RECLAIM) under the CAA. It developed a trading system for stationary sources emitting NO_x and SO_x . RECLAIM covers approximately four hundred stationary sources and it is open to sources that generate more than four tons of either of these pollutants per year. This program also uses Best Available Control Technology (BACT) (Rule 2005, Amended May 6, 2005), for new or relocated facilities. In Arizona (Environmental Services Department, 2000), BACT is air pollutant-specific and applies to any new or modified stationary source. For a new source, for instance, BACT is triggered if a new facility can emit more than 150 lbs/day or 25 tons/year of VOCs, NO_x , SO_x or PM; more than 85 lbs day⁻¹ or 15 tons/year or more than 550 lbs/day or 100 tons/year of CO, and so on.

Best Pollution Control Technologies

EU's Council Directive 96/61/EC (1996, integrated pollution prevention and control) gives the following definition for BAT:

'best available techniques' shall mean the most effective and advanced stage in the development of activities and their methods of operation which indicate the practical suitability of particular techniques for providing in principle the basis for emission limit values designed to prevent and, where that is not practicable, generally reduce emissions and the impact on the environment as a whole:

‘techniques’ shall include both the technology used and the way the installation is designed, built, maintained, operated and decommissioned,

– ‘available techniques’ shall mean those developed on a scale which allows implementation in the relevant industrial sector, under economically and technically viable conditions taking into consideration the costs and advantages, ... as long as they are reasonably accessible to the operator,

– ‘best’ shall mean most effective in achieving a high general level of protection of the environment as a whole.

In this context, *emission limit values* are in mass or concentration units. Under the European Parliament Directive 2001/80/EC (limitations of emissions from the air from large combustion plants),

... Member States (of the EU) may require compliance with emission limit values and time limits for implementation which are more stringent than those set out ... They may include other pollutants, and they may impose additional requirements or adaptation of plant to technical progress ... Member States may, without prejudice to this Directive and Directive 96/61/EC, and taking into consideration the costs and benefits as well as their obligations under Directive 2001/81/EC of the European Parliament and of the Council of 23 October 2001 on national emission ceilings for certain atmospheric pollutants (13) and Directive 96/62/EC, define and implement a national emission reduction plan for existing plants, ...

EU Directive 96/61/EC (national emission ceilings for certain atmospheric pollutants), states in its Article 9, as follows:

... the Commission shall report to the European Parliament and the Council ... on the extent to which the interim environmental objectives ... are likely to be met by 2010 and on the extent to which the long-term objectives ... could be met by 2020. The reports shall include an economic assessment, including cost-effectiveness, benefits, an assessment of marginal costs and benefits and the socioeconomic impact of the implementation of the national emission ceilings on particular Member States and sectors. ... They shall take into account the reports made by Member States pursuant to Article 8(1) and (2), as well as, *inter alia*:

- (a) ...
- (b) developments of best available techniques in the framework of the exchange of information under Article 16 of Directive 96/61/EC;
- ...
- (h) any major changes in the energy supply market within a Member State and new forecasts

reflecting the actions taken by Member States to comply with their international obligations in relation to climate change;

...

- (k) new technical and scientific data including an assessment of the uncertainties in:
 - (i) national emission inventories;
 - (ii) input reference data;
 - (iii) knowledge of the transboundary transport and deposition of pollutants;
 - (iv) critical loads and levels;
 - (v) the model used; and an assessment of the resulting uncertainty in the national emission ceilings required to meet the interim environmental objectives mentioned in Article 5.
- ...
- (l) whether there is a need to avoid excessive costs for any individual Member State;
- (m) a comparison of model calculations with observations of acidification, eutrophication and ground-level ozone with a view to improving models;
- (n) the possible use, where appropriate, of relevant economic instruments.”

Article 10 of this Directive also states as under:

All reviews shall include a further investigation of the estimated costs and benefits of national emission ceilings, computed with state-of-the-art models and making use of the best available data to achieve the least possible uncertainty and taking also into account progress in the enlargement of the European Union, and of the merits of alternative methodologies, in the light of the factors listed in Article 9.

According to Pearce [10],

... European studies suggest that benefits exceed costs even for scenarios defined in terms of maximum technologically feasible reduction (MFR) of pollutants, i.e. scenarios in which the most pollutant-reducing technologies are used. Such scenarios should be characterised by very high marginal abatement costs at very high levels of pollution reduction, precisely the context where one would expect incremental benefits to be less than incremental costs. While the benefit cost ratio does appear to fall for such scenarios relative to other more modest abatement targets, the reduction is not dramatic and benefits continue to exceed costs.

In the United States, the Clean Water Act (33 U.S.C. Sections 1251–1387) is intended to assist in the development and implementation of waste treatment management plans and practices to achieve the goals

of the Act. These must provide for waste treatment using the *best practicable technology* before discharging of pollutants to water. To the extent practicable, waste treatment management is area-wide and provides for the control or treatment of all point and nonpoint sources of pollution (Section 1281).

Under the Framework Directive (2000/60/EC), the European Commission develops a list of *priority substances* using the criteria stated in Article 16(2). Specifically, the Commission should prioritize substances:

on the basis of risk to or via the aquatic environment, identified by: (a) risk assessment carried out under Council Regulation n. 793/93, Council Directive 91/414/EEC, and Directive 98/8/EC of the European Parliament and Council, or (b) target risk-based assessment (Council Regulation 793/93), focusing solely on aquatic ecotoxicity and on human toxicity via the aquatic environment.

Moreover, the Commission must take into account the risk assessment performed under other relevant EU legislation. For example, pesticide may be listed as a priority substance under the Framework Directive, but only if the risk assessment found that the pesticide presented a *significant risk to or via the aquatic environment* under the Plant Protection Products Directive [11].

The EU (COM(2006) 70 final 2005 Environment Policy Review (Enhancing New Policy Development)) states as follows in regard to air pollution:

... Thematic Strategy on Air Pollution ... built on an impact assessment by leading experts, 100 stakeholder meetings and a public consultation that drew 11 000 responses. A peer-reviewed, state-of-the-art modelling framework was developed to examine the economic, social and environmental impacts of different potential policies. It examined links between air pollution and other domains: human health, soil acidification, water eutrophication and further action to reduce carbon emissions. Both the thematic strategies and other environmental policies used detailed impact analysis to select the best policy instrument, and included consideration of market-based alternatives to direct regulation. Administrative burden and impacts on competitiveness were analysed, allowing synergies to be identified. For instance, studies for the Waste thematic strategy indicated that best-practice waste minimisation could save manufacturers €2.9-4.2 bn annually. ... Meeting existing objectives at less cost is the rationale behind streamlining (or simplifying) existing legislation.

Conclusions

The alternatives to control exposure to pollution and risks range from command and control measures to economic instruments, and include risk-based strategies. We find that most jurisdictions adopted a risk-cost-benefit rationale to justify how to set standards to eliminate or reduce to an acceptable level, the impact of hazardous conditions. In some cases, the sole expressed legislative concern is with risk elimination or reduction, independent of costs. Yet, even in those situations, the inescapable fact is that it costs time and resources to reduce risks. The statute does not address with respect to costs, that silence is not golden: public money must be spent, others will necessarily be facing less to meet their objectives. We conclude that the preferable way to assess societal choices – for a given objective – is through risk-cost-benefit analysis coupled with the minimization of its expected value. This precept does not exclude alternative ways to rank those choices and adopt other criteria to justify them. The precept provides a common and well-understood method for comparison against other methods and solutions. The ultimate decision is political; the risk-cost-benefit framework is informative and essential, but does not provide the complete answer.

References

- [1] European Commission (2000). *First Report on the Harmonisation of Risk Assessment Procedures (Part 1)*, European Commission, Brussels.
- [2] HM Treasury (2005). *Managing Risk to the Public: Appraisal Guidance*, Her Majesty's Treasury, London.
- [3] US Nuclear Regulatory Commission, *Regulatory Guide* (Rev. 1, 1/1996), Reg. Guide 8.29, US Nuclear Regulatory Commission, Washington, DC.
- [4] Kastenbergh, W.E. & Frank, M.V. (2005). Rational for a set of risk-based safety goals for launching space nuclear power plants, *Proceedings Space Nuclear Conference 2005, San Diego. June 5–9 2005*.
- [5] Viscusi, W.K. (1993). The Value of Risks to Life and Health, *Journal of Economic Literature* **31**, 1912–1946.
- [6] Abelson, P. (1993). Regulatory Cost, *Science* **259**, 159.
- [7] Ricci, P.F. & Molton, L. (1981). Risk and benefits in environmental law, *Science* **214**, 1096.
- [8] McKee, M., MacLehose, L. & Nolte, E. (2004). *Health Policy and European Union Enlargement, European Observatory on Health Systems and Policy Series*, Open University Press, London.
- [9] Tietenberg, T. (2001). *Environmental Economics and Policy*, 3rd Edition, Addison-Wesley, Boston.

- [10] Pearce, D. (2000). *Valuing Risks to Life and Death: Towards Consistent Transfer Estimates in the European Union and Accession States*, European Commission, Prepared for EC DG XI. (Nov. 13, 2000, rev. Dec. (2000)).
- [11] Mereu, C.A. (2002). The interplay between the EU Water Framework Directive, The List of Priority Substances and Sector-Specific Legislation, *International Environmental Law Committee Newsletter Archives* 4(2), 1.

Related Articles

Cost-Effectiveness Analysis Environmental Health Risk

PAOLO F. RICCI

Effect Modification and Interaction

The term *effect modification* has been applied to two distinct phenomena. For the first phenomenon, effect modification simply means that some chosen measure of effect varies across levels of background variables. This phenomenon is thus more precisely termed *effect-measure modification*, and in the statistics literature is more often termed *heterogeneity* or *interaction* [1]. For the second phenomenon, effect modification means that the mechanism of effect differs with background variables, which is known in the biomedical literature as *dependent action* or (again) *interaction*. The two phenomena are often confused, as reflected by the use of the same terms (*effect modification* and *interaction*) for both. In fact they have only limited points of contact.

Effect-Measure Modification (Heterogeneity of Effect)

To make the concepts and distinctions precise, suppose we are studying the effects that change in a variable X will have on a subsequent variable Y , in the presence of a background variable Z that precedes X and Y . For example, X might be treatment level such as dose or treatment arm, Y might be a health outcome variable such as life expectancy following treatment, and Z might be sex (1 = female, 0 = male). To measure effects, write Y_x for the outcome one would have if administered treatment level x of X ; for example, if $X = 1$ for active treatment, $X = 0$ for placebo, then Y_1 is the outcome a subject will have if $X = 1$ is administered, and Y_0 is the outcome a subject will have if $X = 0$ is administered. The Y_x are often called *potential outcomes* (see **Causality/Causation**).

One measure of the effect of changing X from 0 to 1 on the outcome is the difference $Y_1 - Y_0$; for example, if Y were life expectancy, $Y_1 - Y_0$ would be the change in life expectancy. If this difference varied with sex in a systematic fashion, one could say that the difference was modified by sex, or that there was heterogeneity of the difference across sex. Another common measure of effect is the ratio Y_1/Y_0 ;

if this ratio varied with sex in a systematic fashion, one could say that the ratio was modified by sex.

For purely algebraic reasons, two measures may be modified in very different ways by the same variable. Furthermore, if both X and Z affect Y , absence of modification of the difference implies modification of the ratio, and *vice versa*. For example, suppose for the subjects under study $Y_1 = 20$ and $Y_0 = 10$ for all the males, but $Y_1 = 30$ and $Y_0 = 15$ for all the females. Then $Y_1 - Y_0 = 10$ for males but $Y_1 - Y_0 = 15$ for females, so there is a 5-year modification of the difference measure by sex. However, suppose we measured the effects by expectancy ratios Y_1/Y_0 , instead of differences. Then $Y_1/Y_0 = 20/10 = 2$ for males and $Y_1/Y_0 = 30/15 = 2$ for females as well, so there is no modification of the ratio measure by sex.

Consider next an example in which $Y_1 = 20$ and $Y_0 = 10$ for males, and $Y_1 = 30$ and $Y_0 = 20$ for females. Then $Y_1 - Y_0 = 10$ for both males and females, so there is no modification of the difference by sex. However, $Y_1/Y_0 = 20/10 = 2$ for males and $Y_1/Y_0 = 30/20 = 1.5$ for females, so there is modification of the ratio by sex.

Finally, suppose $Y_1 = 20$ and $Y_0 = 10$ for males, and $Y_1 = 60$ and $Y_0 = 40$ for females. Then $Y_1 - Y_0 = 10$ for males and $Y_1 - Y_0 = 20$ for females, so the Y difference is smaller among males than among females. However, $Y_1/Y_0 = 20/10 = 2$ for males and $Y_1/Y_0 = 30/20 = 1.5$ for females, so the Y ratio is larger among males than among females. Thus, modification can be in the opposite direction for different measures of effect.

Biologic Interaction

The preceding examples show that one should not, in general, equate the presence or absence of effect-measure modification to the presence or absence of interactions in the biologic (mechanistic) sense, because effect-measure modification depends entirely on what measure one chooses to examine, whereas the mechanism is the same regardless of that choice. Nonetheless, it is possible to formulate mechanisms of action that imply homogeneity (no modification) of a particular measure. For such a mechanism, the observation of heterogeneity in that measure can be taken as evidence against the mechanism (assuming of course that the observations are valid). It would be fallacious, however, to infer the mechanism is correct

if homogeneity was observed, for the usual reason that many other mechanisms (some unimagined) would imply the observation.

A classic example is the simple “independent-action” model for the effect of X and Z on Y , in which subjects affected by changes in X are disjoint from subjects affected by changes in Z [2, 3]. This model implies homogeneity (*absence* of modification by Z) of the average X effect on Y when that effect is measured by the difference in the average Y . In particular, suppose Y is a disease indicator (1 if disease occurs, 0 if not). Then the average of Y is the proportion getting disease (the incidence proportion, often called the *risk*) and the average Y difference is the risk difference. Thus, in this context, the independent-action model implies that the risk difference for the effect of X on Y will be constant across levels of Z ; in other words, the risk difference will be homogeneous across Z , or unmodified by Z .

If X and Z both have effects, this homogeneity of the difference forces ratio measures of the effect of X on Y to be heterogeneous across Z . When additional factors are present in the model (such as confounders) homogeneity of the risk differences can also lead to heterogeneity of the excess risk ratios [4]. Thus, under the simple independent-action model, the independence of the X and Z effects will cause the measures other than the risk difference to be heterogeneous, or modified, across Z .

Biologic models for the mechanism of X and Z interactions can lead to other patterns. For example, certain multistage models in which X and Z act at completely separate stages of a multistage mechanism can lead to homogeneity of ratios rather than differences, as well as particular dose–response patterns. Special caution is needed in interpreting observed patterns, however, because converse relations do not hold. Many different plausible biologic models will imply identical patterns in the effect measures [5].

Synergism and Antagonism

Taking the simple independent-action model as a baseline, one may offer the following *dependent-action* definitions for an outcome indicator Y as a function of the causal antecedents X and Z . *Synergism* of $X = 1$ and $Z = 1$ in causing $Y = 1$ is defined as necessity and sufficiency of $X = 1$ and $Z = 1$ for causing $Y = 1$, i.e., $Y = 1$ if and only if $X = 1$ and $Z = 1$. We may also say that $Y = 1$ in

a given individual would be a synergistic response to $X = 1$ and $Z = 1$ if $Y = 0$ would have occurred instead if either $X = 0$ or $Z = 0$ or both. In potential-outcome notation where Y_{xz} is the outcome when $X = x$ and $Z = z$, this definition says synergistic responders have $Y_{11} = 1$ and $Y_{10} = Y_{01} = Y_{00} = 0$. *Antagonism* of $X = 1$ by $Z = 1$ in causing $Y = 1$ is defined as necessity and sufficiency of $Z = 0$ in order for $X = 1$ to cause $Y = 1$. This definition says antagonistic responders to X or Z have $Y_{10} = 1$ or $Y_{01} = 1$ or both, and $Y_{11} = Y_{00} = 0$.

With these definitions, synergism and antagonism are not logically distinct concepts, but depend on the coding of X and Z . For example, switching the labels of “exposed” and “unexposed” for one factor can change apparent synergy to apparent antagonism, and *vice versa* [1, 2]. The only label-invariant property is whether the effect of X on a given person is altered by the level of Z , i.e., the action of X is dependent on Z . If so, by definition we have biologic interaction.

Absence of any synergistic or antagonistic interaction among levels of X and Z implies homogeneity (*absence* of modification by Z) of the average X effect across levels of Z when the X effect is measured by the differences in Y across levels of X [2, 6]. The converse is false, however. Homogeneity of the difference measures (e.g., lack of modification of the risk difference) does not imply absence of synergy or antagonism, because such homogeneity can arise through other means (e.g., averaging out of the synergistic and antagonistic effects across the population being examined).

A more restrictive set of definitions is based on the sufficient-component cause model of causation [1, 7]. Here, *synergism* of the indicators X and Z is defined as the presence of $X = 1$ and $Z = 1$ in the same sufficient cause of $Y = 1$, i.e., the sufficient cause cannot act without both $X = 1$ and $Z = 1$. Similarly, *antagonism* of $X = 1$ by $Z = 1$ is defined as the presence of $X = 1$ and $Z = 0$ in the same sufficient cause of $Y = 1$. These definitions are also coding dependent.

Extensions to Continuous Outcomes

The use of indicators in the above definitions may appear restrictive but is not. For example, to subsume a continuous outcome T such as death time, we may define Y_t as the indicator for $T \leq t$ and apply the above definitions to each Y_t . Similar devices can be

applied to incorporate continuous exposure variables [6]. The resulting set of indicators is of course unwieldy, and in application has to be simplified by modeling constraints (e.g., proportional hazards for T).

Noncausal (Statistical) “Interaction”

Both the preceding usages of “effect modification” and “interaction” refer to causal phenomena (see **Causality/Causation**). In the statistics literature, “interaction” is often used without explicit reference to causality. For example, in the context of regression modeling, an “interaction term” is usually nothing more than a term involving the product of two or more variables. Consider a *logistic regression* (see **Logistic Regression**) to predict a man’s actual sexual preference A ($A = 1$ for men, 0 for women) from his self-reported preference R and the interviewer’s gender G ($G = 1$ for male, 0 for female),

$$P(A = 1|R, G) = \text{expit}(\alpha + \beta R + \gamma G + \delta \cdot R \cdot G) \quad (1)$$

where $\text{expit}(x) = e^x/(1 + e^x)$ is the logistic function. Such a model can be useful in correcting for misreporting. The term $\delta \cdot R \cdot G$ (or sometimes just $R \cdot G$ or just δ) is often called an *interaction term*. It is, however, more accurately called a *product term*, for presumably, neither self-report nor interviewer status has any causal effect on actual preference, and thus cannot interact causally or modify each other’s effect (since there is no effect to modify).

If $\delta \neq 0$, the product term implies that the regression of A on R depends on G : for male interviewers the regression of A on R is

$$P(A = 1|R, G = 1) = \text{expit}(\alpha + \beta R + \gamma \cdot 1 + \delta \cdot R \cdot 1) = \text{expit}(\alpha + \gamma + (\beta + \delta)R) \quad (2)$$

whereas for female interviewers the regression of A on R is

$$P(A = 1|R, G = 0) = \text{expit}(\alpha + \beta R + \gamma \cdot 0 + \delta \cdot R \cdot 0) = \text{expit}(\alpha + \beta R) \quad (3)$$

Thus we can say that the gender of the interviewer affects or modifies the logistic regression of actual preference on self-report. Nonetheless, since

neither interviewer gender nor self-report affect actual preference (biologically or otherwise), they have no biologic interaction.

When both the factors in the regression do causally affect the outcome, it is common to take the presence of a product term in a model as implying biologic interaction, and conversely to take absence of a product term as implying no biologic interaction. Neither inference is correct. The size and even direction of the product term can change with choice regression model (e.g., linear *versus* logistic), whereas biologic interaction is a natural phenomenon oblivious to our choice of model for analysis [1, 8]. Assuming no bias is present, however, a product term in a linear statistical model for a causal dependency can only arise from the presence of biologic interaction in the dependent-action sense [1, 2].

References

- [1] Greenland, S., Lash, T.L. & Rothman, K.J. (2008). In *Modern Epidemiology*, 3rd Edition, K.J. Rothman, S. Greenland & T.L. Lash, eds, Lippincott, Philadelphia, Chapter 5.
- [2] Greenland, S. & Poole, C. (1988). Invariants and noninvariants in the concept of interdependent effects, *Scandinavian Journal of Work, Environment, and Health* **14**, 125–129.
- [3] Weinberg, C.R. (1986). Applicability of simple independent-action model to epidemiologic studies involving two factors and a dichotomous outcome, *American Journal of Epidemiology* **123**, 162–173.
- [4] Greenland, S. (1993). Additive-risk versus additive relative-risk models, *Epidemiology* **4**, 32–36.
- [5] Moolgavkar, S. (1986). Carcinogenesis modeling: from molecular biology to epidemiology, *Annual Review of Public Health* **7**, 151–169.
- [6] Greenland, S. (1993). Basic problems in interaction assessment, *Environmental Health Perspectives* **101**, (Suppl. 4), 59–66.
- [7] VanderWeele, T.J. & Robins, J.M. (2007). The identification of synergism in the sufficient-component cause framework, *Epidemiology* **18**, 329–339.
- [8] Rothman, K.J. (1976). Causes, *American Journal of Epidemiology* **104**, 587–592.

Related Articles

Compliance with Treatment Allocation

SANDER GREENLAND

Efficacy

The primary objective of a clinical trial (*see* **Randomized Controlled Trials**) is generally to establish the efficacy, or true biologic effect, of a new treatment or intervention [1]. The randomized controlled trial in which the new treatment or intervention is compared with a standard treatment or placebo is the gold standard to accomplish this objective. Efficacy is distinct from effectiveness (i.e., the effect of a treatment or intervention when widely used in practice) [1].

A key consideration in the design of comparative efficacy or phase III trials is the selection of the primary efficacy endpoint or outcome for evaluation [2]. This choice depends on the nature of the disease, the nature of the treatment, the duration of the treatment, the expected effects of the treatment, and the time frame for the expected effect. The hypothesis to be tested, the study design, and the sample size are all based on the measurement of the specified endpoint and on the difference between the new treatment group and the standard or placebo group that is considered meaningful.

These endpoints can include response rates to treatments such as cure rates, mortality or survival rates, and continuous variables such as changes from baseline in blood pressure or cholesterol levels. The efficacy evaluation is based on the comparison of these outcomes in the treatment and control groups. For example, the difference in the complete response rates after 4 months of chemotherapy on a new treatment and the standard treatment provides an estimate of the efficacy of the new treatment; the difference between the new and standard treatments in change in diastolic blood pressure from baseline levels after 8 weeks of therapy provides an estimate of the efficacy of the new agent. The difference in mortality or survival rates between the two groups can be used to estimate efficacy. In the presence of censored data, median survival times or the relative risk of death in the two groups can be used to estimate the efficacy of the new treatment compared to the standard one.

Composite or combined endpoints are also often used to evaluate efficacy. For example, we might consider the outcome of a long-term trial to be time to the first occurrence of a nonfatal myocardial infarction, coronary bypass surgery or stent placement for

unstable angina, or mortality from cardiovascular disease.

Ideally, the endpoints to assess efficacy should be obtained for all randomized patients or subjects, and should be unambiguous and verifiable. These measures should be validated and meet expected criteria for accuracy, reproducibility, reliability, and consistency over time [3]. In some circumstances, the assessment of efficacy requires long-term follow-up, and surrogate endpoints are considered for the efficacy evaluation. The criteria for the use of a surrogate endpoint require that it be correlated with the clinical endpoint of interest (e.g., mortality) and capture the effect of the treatment or intervention on the clinical endpoint so that it can be reliably used as a replacement [4]. In early clinical trials to evaluate HIV therapies, CD4+ counts were used as surrogate markers for mortality; subsequently, these counts were found not to be associated with mortality [5].

The evaluation of the efficacy of a new vaccine that is administered to healthy individuals is associated with additional complexities. Preventive vaccine trials are large and the event rates are generally relatively low even in a placebo group. The usual measure of vaccine efficacy (or effectiveness) is estimated as the percentage reduction in disease incidence attributable to the new vaccine, i.e.,

$$\left[1 - RR \left(\frac{\text{vaccinated individuals}}{\text{placebo controls}} \right) \right] \times 100\%$$

where RR is the relative risk. For a discussion of the complexities of intent-to-treat effectiveness studies and per-protocol efficacy studies, see Horne *et al.* [6].

The interpretation of the efficacy results from any trial depends on the conduct of the trial as well as the analysis. The risks of reaching incorrect conclusions regarding the efficacy of a new treatment are increased as the deviations from the planned design and randomization increase. When the intent to treat and the per-protocol or as-treated populations do not differ in any major ways, then the risks of incorrectly attributing efficacy (or lack of efficacy) to a new treatment are reduced.

Summary

We have reviewed the general concepts involved in the evaluation of efficacy from randomized controlled clinical trials. The criteria for the use of surrogate

endpoints and the special issues in vaccine efficacy are noted.

References

- [1] Piantadosi, S. (1997). *Clinical Trials: A Methodologic Perspective*, Wiley-Interscience, New York.
- [2] Friedman, L.M., Furberg, C.D. & DeMets, D.L. (1998). *Fundamentals of Clinical Trials*, 3rd Edition, Springer-Verlag, New York.
- [3] International Conference on Harmonization (1998). E9: guidance on statistical principles for clinical trials, *Federal Register* **63**(179), 49583–49598.
- [4] Prentice, R.I. (1994). Surrogate endpoints in clinical trials: definition and operational criteria, *Statistics in Medicine* **8**, 431–440.
- [5] Fleming, T.R. & DeMets, D.K. (1996). Surrogate endpoints in clinical trials: Are we being misled? *Annals of Internal Medicine* **7**, 605–613.
- [6] Horne, A.D., Lachenbruch, P.A. & Goldenthal, K.L. (2001). Intent-to-treat analysis and preventive vaccine efficacy, *Vaccine* **19**, 319–326.

Related Articles

Compliance with Treatment Allocation

Intent-to-Treat Principle

JUDITH D. GOLDBERG AND ILANA
BELITSKAYA-LEVY

Engineered Nanomaterials

Nanotechnology is a generic term encompassing technologies that depend on manipulating matter at the nanoscale, to produce new materials, devices, and products. It bridges many scientific disciplines, often leading to new and innovative applications. Engineered nanomaterials are a product of nanotechnology, and have an intentionally produced structure at the scale of approximately 1–100 nm. The simplest engineered nanomaterials include compact particles within this size range, formed from single elements or compounds. More complex nanomaterials include nonspherical particles such as carbon nanotubes, and heterogeneous particles such as silver-coated titanium dioxide. Engineered nanomaterials also include aggregated forms of these particles, composites incorporating nanoscale materials, and materials with engineered nanoscale voids. These materials are designed to demonstrate properties that depend on a carefully controlled and implemented nanostructure.

Risk Assessment Challenges

The close dependency between structure and functionality challenges how quantitative risk assessment is applied to these new materials. Nanostructure-dependent properties may lead to health risks that not only depend on chemistry- and mass-based dose, but also on structure-related properties that include particle size, surface area, and shape. Studies have shown that poorly soluble nanostructured materials in rodent lungs can elicit a response that is associated with surface area and surface chemistry, and there is evidence that the size, surface chemistry, and possibly shape of nanomaterials may determine transport through the body and even into cells [1]. Quantitative risk assessment of engineered nanomaterials will depend on determining when they behave differently from conventional materials, what properties are associated with their potential impact, and how exposure, dose and response can be most appropriately evaluated.

Not every engineered nanomaterial will present new challenges. Maynard and Kuempel [2] have proposed that an emphasis should be placed on nanomaterials that can enter the body in some form and

have the potential to exhibit nanostructure-dependent biological behavior. These include powders, suspensions, and slurries of nanometer-scale particles, as well as nanostructured aggregates of these particles if they are sufficiently small to be inhaled or ingested. Other engineered nanomaterials, such as those where nanostructured particles that could potentially enter the body might be released, may require special attention. For instance, machining engineered nanomaterial composites may release airborne nanostructured particles that can be inhaled; and abrasion of engineered nanomaterial surface coatings may release nanostructured materials that are biologically available.

Research is ongoing to support the quantitative risk assessments of engineered nanomaterials. Where these new materials behave in unexpected ways, new information is needed on hazard assessment, dose evaluation, and exposure characterization [3]. In many cases, it is likely that structure-related attributes such as particle size, shape, surface area, and surface chemistry will be more important than mass and chemical composition in determining potential risk [4].

Summary

The challenges of quantifying risks presented by some engineered nanomaterials will most likely continue for many years, particularly as these materials become increasingly sophisticated. Simple solutions are unlikely, given the number of interrelated parameters that may influence potential health impact. Until more comprehensive information is available, it will be necessary to identify those engineered nanomaterials that present unique challenges, and to use existing information together with expert extrapolation as a basis for qualitative and semi-quantitative risk assessment [5, 6].

References

- [1] Oberdörster, G., Oberdörster, E. & Oberdörster, J. (2005). Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles, *Environmental Health Perspectives* **113**(117), 823–840.
- [2] Maynard, A.D. & Kuempel, E.D. (2005). Airborne nanostructured particles and occupational health, *Journal of Nanoparticle Research* **7**(6), 587–614.

- [3] Maynard, A.D., Aitken, R.J., Butz, T., Colvin, V., Donaldson, K., Oberdörster, G., Philbert, M.A., Ryan, J., Seaton, A., Stone, V., Tinkle, S.S., Tran, L., Walker, N.J. & Warheit, D.B. (2006). Safe handling of nanotechnology, *Nature* **444**(16), 267–269.
- [4] Oberdörster, G., Maynard, A., Donaldson, K., Castranova, V., Fitzpatrick, J., Ausman, K., Carter, J., Karn, B., Kreyling, W., Lai, D., Olin, S., Monteiro-Riviere, N., Warheit, D. & Yang, H. (2005). Principles for characterizing the potential human health effects from exposure to nanomaterials: elements of a screening strategy, *Particle and Fibre Toxicology* **2**(8), Doi: 10. 1186/1743-8977-2-8.
- [5] Nel, A., Xia, T., Mädler, L. & Li, N. (2006). Toxic potential of materials at the nanolevel, *Science* **311**, 622–627.
- [6] Maynard, A.D. (2007). Nanotechnology: the next big thing, or much ado about nothing? *Annals of Occupational Hygiene* **51**, 1–12.

Related Articles

Detection Limits

Environmental Hazard

Mobile Phones and Health Risks

ANDREW D. MAYNARD

Enterprise Risk Management (ERM)

Definition

Enterprise risk management (ERM) is a recent risk-management technique, practiced increasingly by large corporations in industries throughout the world. It was listed as one of the 20 breakthrough ideas for 2004 in Harvard Business Review [1]. ERM reflects the change of mind-set in risk management over the past decades. Business leaders realize that certain risks are inevitable in order to create value through operations, and some risks are indeed valuable opportunities if exploited and managed effectively. Sensible risk management flows from the recognition that a dollar spent on managing risk is a dollar cost to the firm, regardless of whether this risk arises in the finance arena or in the context of a physical calamity such as fire. ERM thus proposes that the firm addresses these risks in a unified manner, consistent with its strategic objectives and risk appetite.

Most corporations adopt the definition of ERM proposed by the Committee of Sponsoring Organizations of the Treadway Commission (COSO) in their 2004 ERM framework [2]. It intended to establish key concepts, principles, and techniques for ERM. In this framework, ERM is defined as “a process, affected by an entity’s board of directors, management and other personnel, applied in strategy setting and across the enterprise, designed to identify potential events that may affect the entity, and manage risk to be within its risk appetite, to provide reasonable assurance regarding the achievement of entity objectives”. This definition highlights that ERM reaches the highest level of the organizational structure and is directed by corporations’ business strategies. The concept of risk appetite is a crucial component of the definition. Risk appetite reflects a firm’s willingness and ability to take on risks in order to achieve its objectives. Once it is established, all subsequent risk-management decisions will be made within the firm’s risk appetite. Thus, the articulation of risk appetite greatly affects the robustness and success of an ERM process. Different themes of business objectives are applied to determine risk appetite. Among the most common ones are solvency concerns, ratings concerns, and earnings volatility concerns [3]. The

themes directing the risk appetite process should be consistent with the corporation’s risk culture and overall strategies.

Despite its wide acceptance, the COSO definition is not the only available definition. For example, the casualty actuarial society (CAS) offered an alternative definition in its 2003 overview of ERM [4]. In CAS’s definition, “ERM is the discipline by which an organization in any industry assesses, controls, exploits, finances, and monitors risks from all sources for the purpose of increasing the organization’s short- and long-term value to its stakeholders”. Individual corporations may also define ERM uniquely according to their own understanding and objectives. Creating a clear, firm-tailored definition is an important precursor to the firm implementing a successful ERM framework. In fact, a 2006 survey of US corporations identified that lack of an unambiguous understanding of ERM is the one obstacle preventing companies from putting ERM in place [5].

Current Development of ERM

As a rising management discipline, ERM varies across industries and corporations. The insurance industry, financial institutions, and the energy industry are among the industry sectors where ERM has seen relatively advanced development in a broad range of corporations [6]. The enforcement of ERM in these industries was originally stimulated by regulatory requirements. Recently, more corporations in other industries, and even the public sector, are becoming aware of the potential value of ERM and risk managers are increasingly bringing it to top executives’ agendas. According to a 2006 survey of US corporations, over two-thirds of the surveyed companies either have an ERM program in place or are seriously considering adopting one [5]. An earlier survey of Canadian companies obtained similar results. It found that over a third of the sample companies were practicing ERM in 2003 and an even larger portion of the sample companies were moving in that direction [7]. For a detailed case study in the corporate use of ERM, see Harrington and Niehaus [8] on United Grain Growers.

ERM Implementation

Notwithstanding the attractiveness of ERM conceptually, corporations are often challenged to put it

into effect. One of the main challenges is to manage the totality of corporation risks as a portfolio in the operational decision process, rather than as individual silos, as is traditionally done. Several specific aspects of ERM implementation together with present challenges are considered below.

Determinants of ERM

Although ERM is largely considered as an advanced risk-management concept and toolkit, it is carried out at different paces by corporations. Studies have examined corporate characteristics that appear to be determinants of ERM adoption. For example, Liebenberg and Hoyt [9] found that firms with greater financial leverage are more likely to appoint a chief risk officer (CRO), a signal for their adoption of ERM. In another study, factors including presence of CRO, board independence, chief executive officer (CEO) and chief financial officer (CFO) support for ERM, use of Big Four auditors, and entity size were found to be positively related to the stage of ERM adoption [6]. These factors reflect ERM's role in corporate governance: launch and pursuit of the ERM process lead to better corporate governance, which is desired by both external and internal constituencies.

Operationalization of ERM

The core of the challenge lies in operationalizing ERM in practice. Integration of risks is not merely a procedure of stacking all risks together, but rather a procedure of fully recognizing the interrelations among risks and prioritizing risks to create true economic value. Important components of this procedure include risk identification, risk measurement, risk aggregation/other modeling approaches, risk prioritization, and risk communication.

Risk Identification. The four major categories of risks considered under an ERM framework are hazard risk, financial risk, operational risk (*see Operational Risk Modeling*), and strategic risk [4]. Hazard risk refers to physical risks whose financial consequences are traditionally mitigated by purchasing insurance policies. Examples of hazard risk include fire, theft, business interruption, liability claims, etc. Financial risk refers to those risks involving capital and financial market. Market risk (interest rate risk, commodity risk, and foreign exchange risk) and

credit risk (default risk) are among the most important financial risks. This type of risk is usually hedged by financial instruments, such as derivatives. Operational risk (in Basel II, operational risk is defined as “the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events”) is a nascent risk category and has inspired increasing interest. Operational risk includes internal fraud, external fraud, employment practices and workplace safety, clients, products and business practices, damage to physical assets, business disruption and system failures, execution, delivery, and process management [10]. The newly released Basel Capital Accord II [10] first drew attention to operational risk in the banking industry. The impact soon spread to other industries and now operational risk is ranked as the most important risk domain by US corporate executives [5]. However, given the complex and dynamic nature of operational risk, there is no easy solution: effective risk management of operational risks requires sophisticated and innovative risk-management techniques. Finally, strategic risk is more directly related to the overall strategies of corporations. It includes reputation risk, competition risk, regulatory risk, etc. The management of strategic risk does not automatically fall into standard categories of risk-management techniques. Specific risks perceived by each corporation need to be identified and managed customarily.

Under ERM, the identification of individual risks in different categories should facilitate successive prioritization and integration of risks to best achieve business objectives within the corporation's risk appetite. Moreover, not all risks that are likely to face the corporation fall into one of the above major categories. Any event that may adversely affect the corporation's achievement of its objectives is considered a risk under ERM. Therefore, proper objective identification is a prerequisite for risk identification. For example, business objectives can be described by certain key performance indicators (KPIs), which are usually financial measures such as return on equity (ROE), operating income, earnings per share (EPS), and other metrics for specific industries, e.g., risk adjusted return on capital (RAROC) and risk-based capital (RBC) for financial and insurance industries [4]. Risks are then recognized by means of these company performance metrics. This is the first step to implement a sound ERM process.

Risk Aggregation and Risk Measures. A central step toward operationalizing ERM is to establish a modeling approach for risk integration. The study by Holmer and Zenios [11] is among the earliest studies that shed light on value creation by process integration/holistic management. They proposed an approach that integrates different parts of the production process (designing, pricing, and manufacturing) to improve productivity of financial intermediaries. Although risk management was rarely involved in that work, the underlying rationale is essentially the same.

A sensible way to unify and integrate different types of risks is to derive the total risk (loss) distribution. The process starts with individual risks, which, as random variables, are usually represented by certain probability distribution functions. An aggregated risk distribution for the entire corporation can be derived from these individual risk distributions. A risk measure is then developed to reflect the risk level, e.g., the risk measure can be denoted in dollar terms, in the form of capital requirements. This risk aggregation approach is especially useful for financial firms and insurers, where the goal is to calculate regulatory or economic capital based on the aggregated risk distribution.

Aggregated risk distribution functions are essentially described by the marginal distributions of individual risks and the interrelations (or dependence structure) among the risks. Marginal distributions are found for each identified individual risk through parametric models, nonparametric models, or stochastic simulations [12]. Parametric models fit certain predetermined distribution functions to data using techniques such as maximum-likelihood estimation (MLE) or minimum mean squared error (MSE) estimation. Nonparametric models rely on histogram or smoother-based (e.g., kernel density) estimation from historical data. Stochastic simulation methods repeatedly sample from the distributions of random input variables to build up joint distributions of output quantities. Among these methods, stochastic simulation methods have become increasingly popular in both academia and practice.

There are also multiple ways to capture the interrelations among risks. A simple approach is through variance–covariance matrices. Correlations among different risks are either calculated on the basis of historical data or conjectured by domain experts. Alternatively, structure simulation models

can be employed to link possibly correlated risks to common factors [4]. For example, certain macroeconomic conditions may drive multiple market risks and thus result in interactions among them. Interrelations among risks can be exploited to take advantage of natural hedges and to place early warnings on catastrophic events where different types of risks strike together, which may lead to real economic benefits created by ERM.

At a slightly more sophisticated level, dependence structures among risks can be modeled using *copulas* (see **Copulas and Other Measures of Dependency**). Suppose we have two risks X and Y with marginal distribution functions $F_X(x)$ and $F_Y(y)$. Denote their joint distribution function by $F_{X,Y}(x, y)$. Then a copula [13] is obtained as

$$C(u, v) = F_{X,Y}(F_X^{-1}(u), F_Y^{-1}(v)) \quad (1)$$

Various types of copulas (for example, normal copula or student- t copula) can be employed together with different choices of marginal distributions to derive joint distribution functions.

Among the available risk measures, quantile-based measures are perhaps the most prevalent currently. This class of risk measures focus on the tail area of the distribution functions, i.e., those events occurring with low probabilities but are associated with large losses should they occur, which reflect an intention to protect shareholders' value in the event of default or insolvency. The well-known *value-at-risk* (VaR) (see **Value at Risk (VaR) and Risk Measures**) measure is of this type. VaR is the maximum loss that might be suffered at a given confidence level (e.g., 95%) over a certain period of time (e.g., one trading day). Mathematically, VaR at the α confidence level is the α quantile of the loss distribution function $F_X(x)$, or

$$\text{VaR}_\alpha = F_X^{-1}(\alpha) \quad (2)$$

Although VaR measures are extensively employed, especially in financial risk management, doubts have been raised on its ability to depict a complete risk picture as a “coherent” risk measure,^a meaning a risk measure satisfying several normative axioms [12]. One of the most important concerns is that VaR, in general, fails to satisfy the subadditivity property^b for coherent risk measures, except under special distributional assumptions such as the multivariate normal distribution [14].

A closely related alternative measure, expected shortfall (or loosely, tail-VaR), is proposed to make up for the possible shortcomings of VaR. Expected shortfall takes into account not only the probability of adverse events, as VaR does, but also the average magnitude of these events. Mathematically,

$$ES_\alpha = \frac{1}{1-\alpha} \int_\alpha^1 F_x^{-1}(p) dp \quad (3)$$

where α is the confidence level.

Further considerations lead to other classes of risk measures. For example, the so-called spectral risk measures [15] incorporate a weighting function to describe different degrees of risk aversions on quantiles. In this sense, expected shortfall is imperfect since it assigns equal weight ($\frac{1}{1-\alpha}$) to the entire $(1-\alpha)$ region (and a weight of zero outside the region), implying risk neutrality rather than risk aversion in the region. Moreover, an important risk measure based on the distorted distribution functions was developed by Wang [16, 17]. The distorted survival function of the loss $S^*(x)$ is produced by applying a distortion function $g(\cdot)$ to the original survival function $S(x)$:

$$S(x) = 1 - F(x) \quad (4)$$

$$S^*(x) = g[S(x)] \quad (5)$$

where $F(x)$ is the distribution function and g is an increasing function with $g(0) = 0$ and $g(1) = 1$.

Wang [16, 17] suggests specific choices of the distortion function $g(\cdot)$:

$$g(u) = \Phi[\Phi^{-1}(u) + \lambda] \quad (6)$$

and

$$g(u) = Q[\Phi^{-1}(u) + \lambda] \quad (7)$$

where Φ is the standard normal distribution function, Q is the student- t distribution function, and λ is the market price of risk parameter. Formula (6) is also known as the *Wang transform*. A coherent risk measure can be then developed by taking expectation with respect to the distorted distribution function.^c Risk measures can also be formed to account for other parts of the distribution functions rather than the tail, e.g., they can be formed on the basis of expected utilities (see **Axiomatic Measures of Risk and Risk-Value Models**).

In practice, simplified approaches are sometimes adopted to obtain the aggregated risk measure. For example, one can derive the portfolio VaR as a weighted sum of VaR for each component risk, which implies perfect correlation between risks, or sometimes, multivariate normality is assumed for the individual risk components and a VaR measure is obtained accordingly. However, these simplified measures should be used with caution since they may lead to biased total risk estimation in some situations [13].

Risk Prioritization. To realize effective risk integration, ERM also promotes risk prioritization. Risk prioritization stems from the fact that risks are not equally important to corporations. Prioritization should reflect different aspects of the company's strategies and risk-management philosophy, e.g., cost to tolerate that risk, reduce it, elicit and apply management's risk preferences, etc.

A two-dimensional risk map (see Figure 1) can be used to rank risks. The vertical axis represents impact of the underlying risks (the severity of losses) and the horizontal axis represents likelihood of the underlying risks (the frequency of losses). Different alert levels and risk-management strategies are placed on each quarter panel. The low-likelihood, low-impact area usually needs minimum alarm; the high-likelihood, low-impact area should be dealt with accordingly

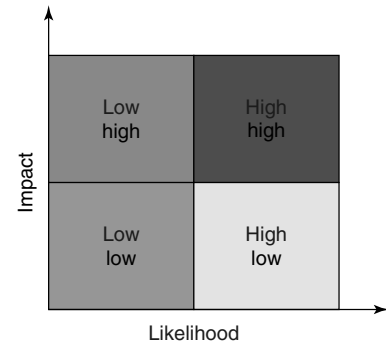


Figure 1 A two-dimensional risk map. The horizontal axis represents loss likelihood and the vertical axis represents loss impact. The four quarter panels stand for different combinations of likelihood and impact. Different colors are used to illustrate the overall impact of risks in each quarter panel to the corporation. Dark gray and black zones usually raise much higher concerns than the light gray and white zones. This map is used in prioritizing risks and designing risk-management techniques accordingly

by the risk-management routine (e.g., risk retention); the low-likelihood, high-impact area requires appropriate monitoring and special risk-management techniques (e.g., insurance); the high-likelihood, high-impact area can be disastrous to the corporation and thus demands full alert and tight control (e.g., risk avoidance) [18]. According to the ranking suggested by the risk map, corporations may want to prioritize those risks with high impact, as they may bring down the entire corporation if incurred. Risk-management activities should then be executed on the basis of assigned priority and characteristics of risks.

Alternatively, risks can also be ranked and prioritized on the basis of their respective impact on KPIs [4]. As explained above, KPIs describe corporations' strategic targets. The ultimate aim of ERM is to assist corporations in achieving these strategic targets by managing risks in the most effective way. Thus, risks that have higher potential influence on KPIs (or other chosen measures of objectives) should be prioritized and treated with special focus.

Risk Reporting and Risk Communications. Despite the extensive attention given to its technical aspects, ERM is not just about numbers. A key factor for success is effective risk communication (*see Role of Risk Communication in a Comprehensive Risk Management Approach*) from the board and executive management to operational units, and across different business departments of corporations. One way to implement and improve risk communication is through a well-designed risk reporting system [19]. The risk reporting system should both provide succinct summaries of critical risk information covering the broad range of corporate risks for board members and executives and allow access to more detailed information for those responsible for specific risks at the operational level. Moreover, both qualitative and quantitative analysis should be incorporated into a single system.

ERM software has been developed for this purpose. For example, an ERM dashboard, an interface providing "role-based information to key decision makers" is recommended for risk reporting [19]. Risk registers are also widely used for risk reporting and management. Risk registers record relevant information including risks, risk assessments, impact on KPIs, risk-management tools, and responsible personnel, to keep track of the risk-management activities and allow interactions among different parties

[18]. There are also other commercial ERM software products that are under development for general use or for use by particular corporations.

ERM and Compliance

ERM at first arises from corporations' efforts to comply with laws and regulations. To this end, ERM is seen more as an efficient internal control process. Within a corporation, it is often conducted with internal control functions and supervised by internal auditors. The most significant regulatory forces responsible for the rise of ERM are the Sarbanes Oxley Act of 2002, the Basel Capital Accord II, and rating criteria set forth by rating agencies such as Standard & Poor's (S&P).

Sarbanes Oxley Act of 2002

In the United States, the Sarbanes Oxley Act of 2002 [20] greatly raised compliance difficulty for corporations. Section 404 of the act governs the corporations' internal control activities over financial reporting and disclosure to the public. External auditors are also involved in assessing and certifying the adequacy of corporations' internal controls. Corporations have invested great amounts of time and money to comply with the act. In this process, they turn to ERM as a solution to the need for adequate, well reasoned and documented, and efficient internal controls, rather than for general risk-management purposes. Meanwhile, Sarbanes Oxley Act itself poses compliance risks for many corporations. ERM can potentially provide an effective toolkit for managing this type of risk.

Basel Capital Accord II

The Basel Capital Accord II [10] has also probably contributed to the development of ERM. This new Basel Capital Accord clearly describes the determination of capital requirements for the banking industry from a regulatory point of view. Besides minimum capital requirements, it also highlights the importance of supervisory review processes for managing major risks. Also for the first time, Basel II explicitly reflects regulatory interest in operational risk. Regulatory capital requirements and review process should stipulate ERM adoption by corporations,

to attain unification of risk and capital management, and to fulfill compliance needs.

Rating Agency

As a major constituency for corporations, rating agencies may have a more direct influence on promoting ERM practice as compared to the previous two forces. S&P started to evaluate ERM practice and incorporate it in the rating process for insurers in 2005 [21] and refined the criteria in 2006 [22]. The rating criteria span important components of the ERM process. Risk-management culture, risk control techniques, methodologies and principles employed by risk models, and the ability to deal with emerging risks all contribute to insurers' overall ERM assessment. S&P also gives positive weight to the articulation of risk appetite (and resulting risk tolerance, risk limits, etc.) because of its fundamental role in the ERM process.

In 2006, S&P extended its ERM evaluation to the financial industry by developing rating criteria specifically for financial institutions [23]. The ERM assessment framework is built up in three dimensions: infrastructure, policies, and methodology. The evaluation process focuses on five aspects: risk governance, operational risk, market risk, credit risk, and funding and liquidity. Among these, risk governance includes risk culture, risk appetite, risk aggregation/quantification, and risk disclosure. Highly rated financial institutions are those that use effective methodologies and procedures to control each important category of risks and that have a holistic view of the overall risk profile. S&P's rating criteria will undoubtedly encourage continuous adoption and elaboration of ERM in these industries. Moreover, it is likely that rating agencies will start to establish rating criteria for general industries in the foreseeable future, which will provide an even stronger incentive for corporations to aggressively advance in the ERM process.

Conclusion: ERM Future – Value Creation

ERM practices may have been initially driven by compliance needs, but ERM development should continue to serve as an internal control function for better corporate governance. Furthermore, the forces upon which ERM thrives are related to the potential economic value generated by better managing risks under identified objectives. One

common objective for corporations is to maximize firm value. ERM provides a framework for corporations to consciously optimize the risk/return relationships for their businesses. This optimization is achieved through the alignment of corporate strategic goals and risk appetite. At the operational level, the alignment guides virtually all activities conducted by the corporation. Specific risks are identified and measured. They are prioritized and integrated by recognizing the interrelations and relative influences affecting different risky outcomes. Risk-management strategies are developed for the entire portfolio of risks and their effects are assessed and communicated. In this way, ERM can reduce waste of resources caused by inadequate communication and cooperation under silo-based risk-management frameworks. ERM can also increase the capacity and free up space for new opportunities to be explored. In addition to these two primary sources of value, more effective risk management also creates benefits from higher credit ratings, lower distress costs, more favorable contract provisions, etc.

Testing the added value of ERM itself is another present challenge. Wang [17] proposes that value creation can be calculated as the increase in economic value of the portfolio after implementing ERM, where the economic value is obtained by discounting the expected net revenue using the distorted distribution function. Zenios [24] demonstrates using operations research methods that effective integration of risks under ERM can create value by pushing out the risk/reward frontier of the entire portfolio. More theoretical and empirical analyses are needed to demonstrate/test the added value from ERM.

ERM is still at an early stage of development. Conceptual and practical frameworks are still being constructed through the combined efforts of regulators, industries, and academia. More advanced methodologies, techniques, and tools are rapidly emerging. Therefore, some of the aspects (e.g., what ERM is, its real effects, in what way it can be best implemented, etc.) described will become refined and clarified as advances accumulate, allowing more concrete and analytical discussions.

End Notes

^a. A coherent risk measure should satisfy a set of properties: monotonicity, subadditivity, positive

homogeneity, and translation invariance. For details, see Artzner *et al.* [25].

^b. For any risks X and Y , a risk measure ρ is said to be subadditive if $\rho(X + Y) \leq \rho(X) + \rho(Y)$, which implies that portfolio risk should be no greater than the sum of individual component risks.

^c. Readers interested in quantile-based measures are directed to Dowd and Blake [12].

References

- [1] Buchanan, L. (2004). Breakthrough ideas for 2004, *Harvard Business Review* **2**, 13–16.
- [2] Committee of Sponsoring Organizations (COSO). (2004). *Enterprise Risk Management – Integrated Framework: Executive Summary*, New York, http://www.coso.org/Publications/ERM/COSO_ERM_Executive_Summary.pdf.
- [3] Standard & Poor's. (2006). *Evaluating Risk Appetite: A Fundamental Process of Enterprise Risk Management*, New York, U.S.A. <http://www2.standardandpoors.com/portal/site/sp/en/us/page.article/2,1,6,4,1148-332051802.html>.
- [4] Casualty Actuarial Society. (2003). *Overview of Enterprise Risk Management*, Arlington, Virginia, U.S.A. <http://www.casact.org/research/erm/overview.pdf>.
- [5] Towers Perrin. (2006). *A Changing Risk Landscape: A Study Of Corporate ERM in the U.S.*, http://www.towersperrin.com/tp/getwebcachedoc?webc=HRS/USA/2006/200611/ERM_Corporate_Survey_110106.pdf.
- [6] Beasley, M., Clune, R. & Hermanson, D. (2005). Enterprise risk management: an empirical analysis of factors associated with the extent of implementation, *Journal of Accounting and Public Policy* **24**, 521–531.
- [7] Kleffner, A., Lee, R. & McGannon, B. (2003). The effect of corporate governance on the use of enterprise risk management: evidence from Canada, *Risk Management and Insurance Review* **6**, 53–73.
- [8] Harrington, S. & Niehaus, G. (2003). United grain growers: enterprise risk management and weather risk, *Risk Management and Insurance Review* **6**, 193–217.
- [9] Liebenberg, A. & Hoyt, R. (2003). The determinants of enterprise risk management: evidence from the appointment of chief risk officers, *Risk Management and Insurance Review* **6**, 37–52.
- [10] Basel Committee on Banking Supervision. (2004). *Basel II: International convergence of capital measurement and capital standards: a revised framework*, Basel, <http://www.bis.org/publ/bcbs107.htm>.
- [11] Holmer, M. & Zenios, S. (1995). The productivity of financial intermediation and the technology of financial product management, *Operations Research* **43**, 970–982.
- [12] Dowd, K. & Blake, D. (2006). After VaR: the theory, estimation, and insurance applications of quantile-based risk measures, *Journal of Risk and Insurance* **73**, 193–229.
- [13] Rosenberg, J. & Schuermann, T. (2006). A general approach to integrated risk management with skewed, fat-tailed risks, *Journal of Financial Economics* **79**, 569–614.
- [14] Embrechts, P., McNeil, A. & Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls, in *Risk Management: Value at Risk and Beyond*, M. Dempster, ed, Cambridge University Press, Cambridge, pp. 176–223.
- [15] Acerbi, C. (2002). Spectral measures of risk: a coherent representation of subjective risk aversion, *Journal of Banking and Finance* **26**, 1505–1518.
- [16] Wang, S. (2000). A class of distortion operators for pricing financial and insurance Risks, *Journal of Risk and Insurance* **67**, 15–36.
- [17] Wang, S. (2002). A set of new methods and tools for enterprise risk capital management and portfolio optimization, *Paper Presented at the Casualty Actuarial Society Forum*, Arlington VA, <http://www.casact.org/pubs/forum/02sforum/02sf043.pdf>.
- [18] Pickett, K.H.S. (2006). *Enterprise Risk Management: A Manager's Journey*, John Wiley & Sons, Englewood Cliffs.
- [19] James Lam & Associates. (2006). *Emerging Best Practices in Developing Key Risk Indicators and ERM Reporting*, [http://www.gsm.pku.edu.cn/stat/public_html/ifirm/reports/ERM%20Dashboard_KRI%20White%20Paper_October%202006\(james's%20paper2\).pdf](http://www.gsm.pku.edu.cn/stat/public_html/ifirm/reports/ERM%20Dashboard_KRI%20White%20Paper_October%202006(james's%20paper2).pdf).
- [20] Sarbanes-Oxley Act of 2002. (2002). *United States Public Law No. 107–204*, Government Printing Office, Washington, DC.
- [21] Standard & Poor's. (2005). *Insurance Criteria: Evaluating the Enterprise Risk Management Practices of Insurance Companies*.
- [22] Standard & Poor's. (2006). *Insurance Criteria: Refining the Focus of Insurer Enterprise Risk Management Criteria*.
- [23] Standard & Poor's. (2006). *Criteria: Assessing Enterprise Risk Management Practices of Financial Institutions*.
- [24] Zenios, S. (2001). *Managing Risk, Reaping Rewards: Changing Financial World Turns to Operations Research*, OR/MS Today, October 2001.
- [25] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**, 203–228.

Related Articles

Considerations in Planning for Successful Risk Communication Early Warning Systems (EWSs) for Predicting Financial Crisis

JING AI AND PATRICK L. BROCKETT

Environmental Carcinogenesis Risk

Cancer is a major public health concern, with approximately one-half of all men and one-third of all women likely to develop the disease during their lifetime [1]. Cancer is the second leading cause of death in the United States [2], with approximately 200 deaths per year per 100 000 population (all ages, genders, and races combined). The causes of cancer are complex and multifactorial, involving genetic factors and natural aging. However, scientists agree that a significant proportion of cases are likely attributable to the environment, especially when broadly interpreted to include lifestyle choices such as smoking and diet, as well as chemical exposures encountered in the workplace, air, water, or in our food (see **Air Pollution Risk**; **Water Pollution Risk**).

There are several excellent sources of information about known cancer causing agents, or carcinogens. For example, in response to a U.S. Congressional mandate, the National Toxicology Program (NTP; see <http://ntp-server.niehs.nih.gov/>) reports biennially on all substances that are known or suspected to be human carcinogens, and also issues a periodic *Report on Carcinogens* [2]. The International Agency for Research on Cancer (IARC; see <http://www.iarc.fr/>) is another excellent source.

Both the NTP and IARC distinguish substances for which there is clear evidence of carcinogenicity in humans from those where carcinogenicity is suspected, but not confirmed. Suspicion of carcinogenicity in humans may be based on experimental data in animals (see **Cancer Risk Evaluation from Animal Studies**) or knowledge of the chemical structure and function of the compound itself. Classification as a known carcinogen usually requires the existence of reliable epidemiological data. Not surprisingly, the list of known and suspected carcinogens is relatively small. Classifications can also change over time in response to evolving science. For example, while the 2000 report listed 65 known carcinogens, this number was reduced to 58 for the 2005 report [3]. Saccharin is a high profile substance that was listed as a suspected carcinogen in the eighth report [4], but subsequently delisted. Among some of the more familiar known carcinogens

are cigarette smoke (see **Environmental Tobacco Smoke**), radon (see **Radon Risk**), solar radiation, aflatoxins, several metals (see **Arsenic**, cadmium), a variety of air pollutants products (1,3-butadiene, soot, coke oven emissions), several chemotherapeutic agents (e.g. tamoxifen, thiotepa) and some dyes. The list of suspected carcinogens is much longer. Almost always, the basis of the latter listing are positive findings in controlled animals experiments, although data from studies of human populations is also important in this regard. Statistical methodology to support analysis of such data is constantly evolving; see, e.g. [3, 5–11].

Progress in the fields of genetics and microbiology has opened up many fruitful avenues of research into the mechanistic causes of cancer [12]. The study of biomarkers such as DNA adducts [13] and other forms of DNA damage is of particular interest to many environmental scientists and also raises many challenging and interesting statistical questions.

References

- [1] Murphy, S. (2000). *Deaths: Final Data 1998*, National Center for Health Statistics, Hyattsville.
- [2] U.S. National Toxicology Program (2001). *11th Report on Carcinogens (Released January 2005)*, U.S. Department of Health and Human Services, Public Health Service, Research Triangle Park, <http://ehis.niehs.nih.gov/roc/>.
- [3] Portier, C.J. & Dinse, G.E. (1987). Semiparametric analysis of tumor incidence rates in survival/sacrifice experiments, *Biometrics* **43**, 107–114.
- [4] U.S. National Toxicology Program (1998). *8th Report on Carcinogens*, U.S. Department of Health and Human Services, Public Health Service, Research Triangle Park, <http://ehis.niehs.nih.gov/roc/>.
- [5] Archer, L.E. & Ryan, L.M. (1989). Accounting for misclassification in the cause-of-death test for carcinogenicity, *Journal of the American Statistical Association* **84**, 787–791.
- [6] Burnett, R.T., Ross, W.H. & Krewski, D. (1995). Non-linear mixed regression models, *Environmetrics* **6**, 85–99.
- [7] Dunson, D.B. & Dinse, G.E. (2002). Bayesian models for multivariate current status data with informative censoring, *Biometrics* **58**, 79–88.
- [8] Goddard, M., Krewski, D. & Zhu, Y. (1994). Measuring carcinogenic potency, in *Environmental Statistics, Assessment, and Forecasting*, C.R. Cothorn & N.P. Ross, eds, Lewis Publishers, Boca Raton, pp. 193–208.

- [9] Haseman, J.K. (1995). Data analysis: statistical analysis and use of historical control data, *Regulatory Toxicology and Pharmacology* **21**, 52–59.
- [10] Ryan, L.M. (1985). Efficiency of age-adjusted tests in animal carcinogenicity experiments, *Biometrics* **41**, 525–531.
- [11] Peddada, S.D., Dinse, G.E. & Haseman, J.K. (2005). A survival-adjusted quantal response test for comparing tumor incidence rates, *Journal of the Royal Statistical Society, Series C: Applied Statistics* **54**, 51–61.
- [12] Perera, F.P., Mooney, L.A., Dickey, C.P., Santella, R.M., Bell, D., Blaner, W., Tang, D. & Whyatt, R. (1996). Molecular epidemiology in environmental carcinogenesis, *Environmental Health Perspectives* **104**(Suppl. 3), 441–443.
- [13] Poirier, M.C. (1997). DNA adducts as exposure biomarkers and indicators of cancer risk, *Environmental Health Perspectives* **105**(Suppl. 4), 907–912.

Further Reading

Johnson, B.A., Kupper, L., Taylor, D.J. & Rappaport, S.M. (2005). Modeling exposure-biomarker relationships: applications of linear and nonlinear toxicokinetics, *Journal of Agricultural, Biological, and Environmental Statistics* **10**, 440–459.

Related Articles

Environmental Hazard

Environmental Risks

What are Hazardous Materials?

LOUISE RYAN

Environmental Exposure Monitoring

The opinions expressed here are those of the author and do not reflect the opinions or policies of the United States Environmental Protection Agency. Mention of any trade names or commercial products does not constitute an endorsement or recommendation for their use.

Exposure monitoring has become increasingly important since the mid-1970s due to greater awareness about chemical pollution in the environment and concern about the role pollutants may play directly or indirectly in the early onset of disease (*see Environmental Hazard*). Drawing ingredients from environmental chemistry, epidemiology, industrial hygiene, and pharmacology, exposure monitoring is the means to estimate contact with environmental pollutants contained in air, water, food, products, and transport media. Table 1 shows the array of parameters and constituents that can be associated with exposure monitoring. Exposure monitoring is used in environmental health surveillance and is a key ingredient of the risk assessment process. In the former, exposure monitoring, sometimes including collection and analysis of biological media (blood, breath, feces, saliva, or urine), can provide an indication of a person's contact with one or more chemicals of interest (Figure 1). In this context, exposure monitoring is used to determine if chemical concentrations exceed allowable limits in specific media or to evaluate the trend of exposure to a chemical in the environment over time. This kind of trend analysis is frequently done to assess the efficacy of some effort or activity to reduce or eliminate emission of a chemical into the environment. In the latter, exposure monitoring information is used to construct an exposure assessment that links information about the inherent toxicity of a chemical (hazard assessment) and information about the quantity of the chemical that is capable of eliciting an adverse health response (dose-response assessment) to characterize public health risk to that chemical (Figure 2). In this context, exposure monitoring provides information (exposure assessment) about who is exposed, the route of chemical exposure (inhalation, ingestion, or contact), and the duration and frequency of contact with the chemical. This kind of information is used frequently to set limits

for chemical intake (allowable limits, tolerances) or to determine if additional intervention is warranted to mitigate risk to protect public health [1].

Classically, exposure is defined as the condition of a biological, chemical, or physical agent contacting the outer boundary of a human over time [3, 4]. In this context, contact can mean the visible exterior of the person at the skin or include openings into the body such as the mouth and the lungs. An agent entering the body can be described in two steps: contact (exposure) followed by crossing the boundary of the body (potential dose (PD)) with subsequent absorption, transport, and distribution of an amount of the agent to one or more sites in the body (internal dose) (see Figure 1). Thus, exposure monitoring measures the amount of an agent that an individual contacts over time in air, water, soil, food, a product, or a transport or carrier medium *via* inhalation, ingestion, or dermal absorption (*see What are Hazardous Materials?; Environmental Risks*).

Exposure over a period of time can be represented by a time-dependent profile of the exposure concentration. The area under the curve of this profile is the magnitude of the exposure, in concentration-time units [5, 6].

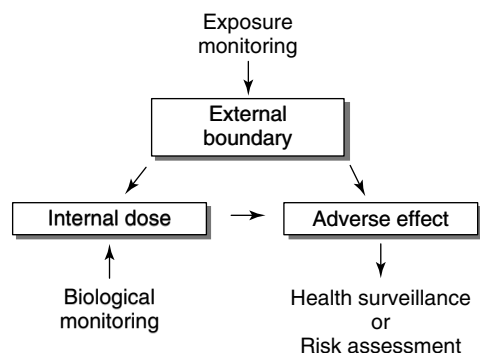
$$E = \int_{t_1}^{t_2} C(t) dt \quad (1)$$

where E is the magnitude of exposure, $C(t)$ is the exposure concentration as a function of time, and t is time, $t_2 - t_1$ being the exposure duration (ED). Integrated exposures are done typically for a single individual, a specific chemical, and a particular pathway or exposure route over a given time period. The integrated exposures for a number of different individuals (a population or population segment, for example) may then be displayed in a histogram or curve (usually, with integrated exposure increasing along the abscissa or x axis, and the number of individuals at that integrated exposure increasing along the ordinate or y axis). This histogram or curve is a presentation of an exposure distribution for that population or population segment (Figure 3).

Dose refers to the amount of an agent that crosses the external boundary of an individual. Dose is dependent upon the contaminant concentration and the rate of intake (ingestion or inhalation) or uptake (dermal absorption). PD is the amount of an agent that could be ingested, inhaled, or deposited on the skin. The absorbed dose is the amount of an agent

Table 1 Exposure monitoring parameters and constituents^(a)

Exposure parameter	Range of constituents
Agent	Biological, chemical, physical
Source categories	Anthropogenic/natural, area/point, stationary/mobile, indoor/outdoor
Media	Air, water, soil, sediment, dust, food, product, surface
Pathways	Inhalation, ingestion, skin contact
Duration	Seconds, minutes, hours, days, weeks, months, years
Frequency	Continuous, periodic, intermittent
Subjects	Individuals, groups, populations
Scope	Site specific, local, regional, national

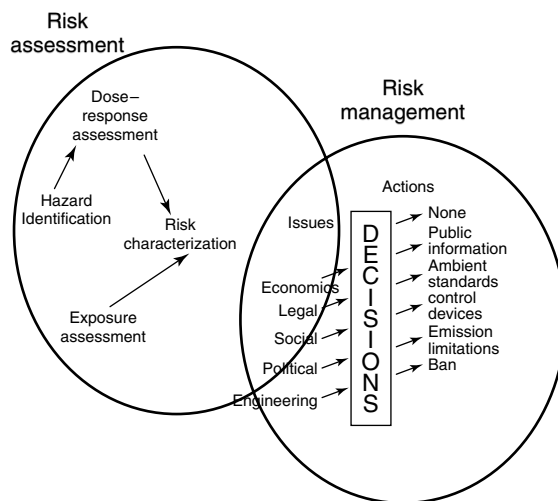
^(a) Modified from reference [2]**Figure 1** Exposure monitoring, biological monitoring and their relationship to health surveillance, and risk assessment

absorbed into the body through the gastrointestinal tract, lungs, or skin. PD may be calculated as follows:

$$PD = C \times IR \quad (2)$$

where

PD: potential dose (mg day^{-1}); *C*: contaminant concentration in the media of interest (milligrams per square centimeter, mg cm^{-2} ; milligrams per cubic meter, mg m^{-3} ; milligrams per gram, mg g^{-1} ; or milligrams per liter, mg l^{-1}); and *IR*: intake or contact rate with that media per unit time (minutes, hours, days, weeks etc.). Frequently, PD rates are normalized to body weight as a function of time (e.g., milligrams per kilogram per day, $\text{mg}^{-1} \text{kg}^{-1} \text{day}^{-1}$) by multiplying by ED and frequency, and dividing by

**Figure 2** The risk assessment paradigm showing the role of exposure assessment in risk characterization and risk assessment

body weight and averaging time to yield an average daily dose. For example

$$I = \frac{(C \times CR \times EF \times ED)}{(BW \times AT)} \quad (3)$$

where

I: intake

C: concentration of the agent

CR: contact rate (ingestion, inhalation, and dermal contact)

EF: exposure frequency

ED: exposure duration

BW: body weight

AT: averaging time

More details and specific examples can be found elsewhere [4, 7].

Exposure monitoring can be performed by direct, indirect, or reconstructed methods. Direct exposure monitoring methods involve actual collection of microenvironmental or personal samples on the individual or in their proximity at the time contact with the agent occurs. Personal monitoring devices are used to collect air samples at the breathing zone or breath samples to measure inhalation of an agent. A duplicate portion of all food and beverages that is prepared and consumed is collected and analyzed to measure the agent in the diet. Areas of the skin are covered with an absorbent patch or are wiped

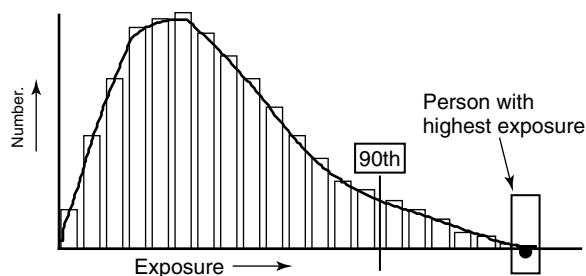


Figure 3 Hypothetical exposure distribution showing 90th percentile and individual with greatest exposure

with a solvent moistened gauze pad periodically to measure skin contact with the agent from a treated or contaminated surface. Additionally, there may be collection and analysis of biological samples (blood, breath, feces, saliva, or urine) from the individual being monitored to provide another measurement of exposure to the agent. These measurements are frequently combined with administration of an interview or questionnaire to capture information about the individual (age, gender, occupation, lifestyle, household size, housing characteristics) and the amount of time and frequency an individual spends performing a variety of tasks in different locations (time location activity patterns) during the monitoring period. This information is used to relate the monitoring measurements to contact with one or more agents and to determine if age, gender, socioeconomic level, and cultural, or lifestyle behavior may influence exposure to the agent being measured. This kind of information may provide an opportunity to associate an exposure to a specific age group (children or adults) or to specific activities (cleaning, food preparation, driving a car). These methods may be conducted for a single specific route of exposure (inhalation, ingestion, or contact) or for several routes to provide an aggregate estimate of exposure depending on the nature of the pollutant and the information desired. The strength of these approaches is that they measure exposure directly as it occurs by measuring the agent concentration at the interface of the person and the environment as a function of time, resulting in an exposure profile. Consequently these methods provide the most accurate estimate of exposure for the period of time of the measurements for the individuals being studied. They are capable of identifying major sources of exposure because they monitor individuals as they perform specific tasks or go about their normal daily routine throughout the day. They

can distinguish among various sources and pathways of pollution, which may be important for regulatory purposes and public health protection. They have the advantage of permitting selection of specific individuals for study such that monitoring can be conducted on individuals with specific personal characteristics (children in day care, females), socioeconomic characteristics (urban apartment dwellers, rural residents), or lifestyle (bicycle riders, gardeners, or sport fishermen) that may be pertinent to exposure to the agent. For example a study to estimate exposure of urban suburban and rural residents in an area to specific air pollutants would likely monitor selected air pollutants in the ambient environment, in indoor environments in the home and workplace, and in the modes of travel for the individuals being studied to obtain a complete picture of exposure to these air pollutants and to identify major pollutant sources (*see Air Pollution Risk*). A study of the contribution of local fish consumption to exposure to specific heavy metals such as mercury or organic pollutants such as *polychlorinated biphenyls* (*see Polychlorinated Biphenyls*) would likely focus on recreational fisherman who consume their catch regularly compared to residents who do not eat fish from local sources or who do not eat fish at all. In this case the study would recruit individuals with these dietary habits, investigate where fish were caught, how they were prepared and consumed, the quantity of any pollutants present in the edible portions, and the frequency of consumption to estimate the contribution of these fish to pollutant exposure. This kind of study might incorporate collection and analysis of biological samples in an effort to relate the amount of exposure measured to fish consumption. Direct exposure methods are limited by the fact that measurement devices and techniques do not exist for all agents and that assumptions must be made about relating short-term

monitoring measurements to estimates of long-term exposure. Generally, direct exposure monitoring is expensive, time consuming, and labor intensive to conduct. This is due to the requirements to select individuals with the characteristics of interest for study, collect, and analyze microenvironmental, personal, and biological samples, and record the time location activity pattern information to identify the sources of exposure for the individuals being studied. These methods have been used successfully to monitor individuals' exposure to volatile organic compounds, pesticides, particles, and heavy metals, and provide the best estimates of exposure to environmental agents [8–10].

Indirect exposure monitoring methods do not entail personal exposure monitoring. They combine environmental monitoring data or determinations about the concentration of an agent in various media in the environment with fate and transport models and assumptions about how contact with an agent takes place over time (scenario) to estimate the exposure that would occur to an individual engaged in that activity at that location for a specific period of time. As such exposure is estimated by combining estimates of the concentration of the agent in various media with assumptions about the time location activity patterns (scenario) of the individual where contact with the agent is expected to occur [4]. Using the air pollution example, ambient air monitoring network data for the pollutants of interest could be combined with a fate and transport model to estimate air pollution in the localities of interest. These estimates could be combined with assumptions about where and how individuals spend their time indoors with assumptions about the concentration of air pollutants in those microenvironments to construct an estimate of the exposure to individuals to these air pollutants. In the fish consumption example, residue levels in fish obtained from a local or regional survey could be combined with assumptions about preparation and consumption patterns to estimate dietary exposure to these compounds. The strength of the indirect methods is their economy, efficiency, and flexibility. Environmental monitoring data or concentration determinations are readily available or easily obtained for many agents in a variety of media. These data can be combined with fate and transport models for the appropriate pathways to construct reasonable estimates of exposure for specific scenarios or events.

The major weakness of these methods is the reasonableness of the estimates that are made about the concentrations of the agent in the environment at the point of contact of the individual generated by the fate and transport models and the assumptions about how the contact takes place over time that is described in the scenario to construct the exposure event. This is important because the further away the agent is measured from the point of contact with the individual the greater the uncertainty about the estimate of exposure in most cases. Likewise, care must be exercised in constructing the scenario in terms of the nature, duration, and frequency of the activity to correctly approximate time location activity patterns that would be monitored by direct exposure monitoring methods. Notwithstanding these limitations, indirect exposure monitoring methods are widely used to evaluate exposure to a variety of agents for regulatory purposes, estimate public health risk, or determine if additional exposure monitoring is needed to provide more detailed information about exposure to the agent.

Biological monitoring itself can be considered a form of direct exposure monitoring. But in this case, samples of biological media (blood, breath, feces, saliva, or urine) are collected and analyzed to measure the agent or its metabolites (biological markers). These measurements provide an indication of the amount of the agent that enters the body (dose). When this information is combined with information about time location activity patterns obtained from an interview or questionnaire, these measurements can be used to reconstruct the past exposure events that led to the measured agent or its metabolite in the body [4]. Thus, exposure reconstruction relies on measuring the agent or its metabolites in the body to back-calculate dose and exposure. Moreover, these methods can be used to provide information about the ability of the agent to exert a toxic effect at one or more target organs (target dose or target organ toxicity) when combined with physiological modeling, provided that sufficient information is known about the biological distribution and fate of the agent in the body. In the air pollution example, breath samples could be collected from individuals who would be administered a questionnaire about their recent time activity patterns and likely exposure to the air pollutants of interest. The breath sample analyses and questionnaire results would be evaluated to determine if the breath results could be related with

particular activities indicated in the questionnaire responses for specific air pollutants to indicate the source of exposure. Similarly, a series of blood samples could be collected from individuals who were reported to have recently consumed locally caught fish in an effort to estimate exposure to specific chemical residues from this source. The biological monitoring approach is useful because it provides an integrated estimate of exposure and does not require the measurement of individual routes of exposure. Moreover, it can provide information about the likelihood of the exposure to cause adverse health effects. Thus, biological monitoring may be more economical and efficient than direct personal exposure monitoring methods, provided that there is adequate information about time location activity patterns to relate the biological measurements to an exposure event. However, this approach is often limited to those agents that are not metabolized in the body or that have unique metabolites. Often metabolites lack specificity, which makes it difficult to assign them to a specific agent or to a specific exposure especially when the time–location activity pattern information is incomplete.

References

- [1] Ott, W.R. & Roberts, J.W. (1998). Everyday exposure to toxic pollutants, *Scientific American* **2**, 1–7.
- [2] Sexton, K., Callahan, M.A., Ryan, E.F., Saint, C.G. & Wood, W.P. (1995). Informed decisions about protecting and promoting public health: rationale for a national human exposure assessment survey, *Journal of Exposure Analysis and Environmental Epidemiology* **5**, 233–256.
- [3] National Research Council (NRC), Committee on the Institutional Means for Assessment of Risks to Public Health, Commission on Life Sciences (1983). *Risk Assessment in the Federal Government: Managing the Process*, National Academy Press, Washington, DC.
- [4] United States Environmental Protection Agency (US EPA) (1992). *Guidelines for Exposure Assessment*, EPA/600/Z-92/001, Washington, DC.
- [5] Liroy, P.J. (1990). Assessing total human exposure to contaminants, *Environmental Science and Technology* **24**, 938–945.
- [6] National Research Council (NRC), Committee on Advances in Assessing Human Exposure to Airborne Pollutants, Committee on Geosciences, Environment, and Resources, National Research Council (1990). *Human Exposure Assessment for Airborne Pollutants: Advances and Applications*, National Academy Press, Washington, DC.
- [7] United States Environmental Protection Agency (US EPA) (2004). *Example Exposure Scenarios*, Washington, DC. <http://cfpub.epa.gov/ncea/cfm/recorddisplay.cfm?deid=85843>.
- [8] Liroy, P.J., Wallace, L. & Pellizzari, E. (1991). Indoor/outdoor, and personal monitor and breath analysis relationships for selected volatile organic compounds measured at three homes during New Jersey team-1987, *Journal of Exposure Analysis and Environmental Epidemiology* **1**, 45–61.
- [9] Whitmore, R.W., Immerman, F.W., Camann, D.E., Bond, A.E., Lewis, R.G. & Schaum, J.L. (1994). Non-occupational exposure to pesticides for residents of two U.S. cities, *Archives of Environmental Contamination and Toxicology* **26**, 45–79.
- [10] Ozkaynak, H., Xue, J., Spengler, J., Wallace, L., Pellizzari, E. & Jenkins, P. (1996). Personal exposure to airborne particles and metals: results from the particle team study in Riverside, California, *Journal of Exposure Analysis and Environmental Epidemiology* **6**, 57–78.

Further Reading

- Ott, W.R., Steinman, A.C. & Wallace, L.A. (eds) (2006). *Exposure Analysis*, CRC Press, Boca Raton.
- World Health Organization (WHO), International Programme on Chemical Safety (IPCS) (2000). *Environmental Health Criteria 214, Human Exposure Assessment*, Geneva.

Related Articles

Environmental Performance Index

Soil Contamination Risk

Water Pollution Risk

MICHAEL DELLARCO

Environmental Hazard

Preliminaries to Environmental Hazard Analysis

Risk assessment has evolved in the past few decades as a rigorous quantitative approach with applications in numerous disciplines. In the fields of public health and environmental protection, for instance, the US National Academy of Sciences published its seminal “Red Book” [1] on risk assessment, in 1983. This influential work suggested a systematic, sequential process for researching risk, which included hazard evaluation as the first step. Since then environmental scientists have dealt with the concept of environmental hazard. After defining, evaluating, and ranking environmental hazards, more specific assessments of their risks have been developed. What is an environmental hazard? An environmental hazard is any condition, process or state that adversely affects the environment. Environmental hazards often manifest as physical or chemical pollution from unwanted waste or contamination usually due to economic activities of humans, such as mining, power generation, manufacturing, transportation, farming, forestry, and pesticide use in air, water, and soils (*see Air Pollution Risk; Water Pollution Risk; What are Hazardous Materials?*). Monitoring, remote sensing, and other techniques have shown that environmental hazards, singly or in combination, have affected and continue to impact the most remote places on earth. This ranges from ocean bottoms, threatening the oceans with extinction as living ecologic zones, up to the stratosphere, which has lost part of its screening capability of protecting humans, plants, and animals from the sun’s harsh ultraviolet rays.

Environmental hazards have potential for widespread harm to humans and the physical environment and appear to be increasing in number and extent. Today, global warming appears to be emerging as possibly the most dire and pervasive future environmental hazard. A few of the more common environmental hazards are reviewed briefly, below.

Lead

Lead is a ubiquitous environmental hazard. Its use historically was thought possibly to be injurious, but,

without scientific, epidemiologic methods, its actual risk could not be quantified (*see Lead*). Industries and consumers therefore disregarded the hazard and used lead in such products and processes involved in the manufacture of wine preservatives, coins, pewter, candle wicks, ceramic dishware, water pipes, paints, and, most significantly, as an additive in gasoline. Tetraethyl lead was found effective for increasing gasoline octane, and was prominently used in gasoline from the 1920s until the late 1980s. With gasoline lead residues in the streets, young children typically obtained doses in contaminated soil, or through ingestion of lead paint. Regulations now prohibit lead in gasoline or house paint.

PCBs

Industrial chemists synthesized a family of 209 simple molecules with marvelous properties as lubricants and heat absorbers. *Polychlorinated biphenyls* (PCBs) (*see Polychlorinated Biphenyls*) are virtually indestructible. After their introduction into commercial use [2], information began to emerge from environmental and occupational epidemiologic research, that PCBs caused chloracne, cancers, and other adverse health effects and were bioaccumulative and persistent. PCBs have been banned but since they are virtually indestructible they continue to circulate through the atmosphere, in river beds, soils, and the oceans, migrating up food chains to top predators such as the polar bear and man [3].

Asbestos

Asbestos was widely used as a flame retardant material and a component in automobile brake linings. Well after its established commercial use, scientists documented increases in lung disease among workers occupationally exposed to asbestos. Researchers eventually showed that asbestos exposure was causally associated with cancers of the lip, nasal passages, oral cavity, trachea, bronchus and lungs, mesothelial tissue, stomach, colon, and rectum (*see Asbestos*). Asbestos fibers – very thin, sinuous fibers akin to needles – are capable of migrating long distances within the body and lodging as foreign bodies. After this, the US Environmental Protection Agency banned the material and implemented a national program of remediation in schools and public buildings.

References

- [1] U.S. National Research Council (NRC), Committee on the Institutional Means for Assessment of Risks to Public Health, Commission on Life Sciences (1983). *Risk Assessment in the Federal Government: Managing the Process*, The National Academies Press, Washington, DC.
- [2] Muir, D. & Howard, P. (2006). Are there other persistent organic pollutants? A challenge for environmental chemists, *Environmental Science and Technology* **40**, 7157–7166.
- [3] Chiu, A., Chiu, N. & Beaubier, N. (2000). Effects and mechanisms of PCB ecotoxicology in food chains: algae, fish, seal, polar bear, *Environmental Carcinogenesis and Ecotoxicology Reviews* **C18**(2), 127–152.

Further Reading

Covello, V.T. & Merkhoher, M.W. (2006). *Risk Assessment Methods: Approaches for Assessing Health and Environmental Risk*, Springer, The Netherlands.

Field, R.A., Eisenberg, N.A. & Compton, K.L. (2007). *Quantitative Environmental Risk Analysis for Human Health*, John Wiley & Sons, Hoboken.

Smith, K. (2004). *Environmental Hazards: Assessing Risk and Reducing Disaster*, Rutledge, London.

Yassi, A., Kjellstrom, T. & Guidatti, T.L. (2001). *Basic Environmental Health*, Oxford University Press, Oxford.

Related Articles

Mercury/Methylmercury Risk

What are Hazardous Materials?

.

JEFF BEAUBIER AND BARRY D. NUSSBAUM

Environmental Health Risk

Risk is defined as the probability that a substance or situation will produce harm under specified conditions. Safety, the probability that harm or loss will not occur under specified conditions, is the reciprocal of risk. Both the risk and the safety are intimately associated with issues of perception [1]. Risk assessment is the process of estimating the potential impact of a chemical, physical, microbiological, or psychosocial hazard on a specified human population or ecological system under a specific set of conditions and for a certain timeframe [2].

Traditionally, risk has been assessed and managed by trial and error. Societies have survived natural disasters by building structures resistant to collapse under severe environmental conditions and exercising preventive measures to minimize diseases resulting from exposure to the environmental factors and contaminants. Risk assessment refers to the technical assessment of the nature and magnitude of the risks.

The complexity of the nature, interdependencies of environmental systems, and correlation between the hazard, exposure, and risk have resulted in development of the systematic approaches to risk assessment [3]. During the last decades, scientists, engineers, epidemiologists, among others, began conducting analyses of environmental systems, material, and technologies to understand risk and its relation to hazards, exposure conditions, and associated harms due to the exposure (*see, for example, Air Pollution Risk; Soil Contamination Risk*).

In the United States, efforts toward defining and managing risk properly have resulted in legislations that began in the 1970s with the formation of the regulatory frameworks, such as Environmental Protection Agency (EPA) and Occupational Safety and Health Administration (OSHA) in conjunction with other newly emerging industry of firms and academic centers [4].

Risk can be assessed with regards to safety (low probability, acute, high consequence, and accidental), human health (high probability, chronic, low consequence, and ongoing), ecology, public welfare, and goodwill. *Ecological risk assessment* (*see Ecological Risk Assessment*) considers indirect effects

of hazards on human health through disruption of the environment. The assessment of ecological risk follows the same steps as human risk assessment does. Public welfare and goodwill risks are value focused and are based on perceptions and property value concerns. Financial risk analysis focuses on the business viability, liability, insurance, and investment return [2, 3, 5].

Environmental Risk

Environmental factors can significantly impact human development, health, and welfare. Exposure to hazardous material in the air, water, soil, food, and physical hazards in the environment could result in unwanted morbidity and mortality. *Hazardous materials* (*see What are Hazardous Materials?*) impact humans and other organisms through multiple routes, including inhalation, ingestion, and dermal contact [3].

Environmental risk assessment is a scientific endeavor, which uses the facts and assumptions to estimate the potentials for adverse effects on human health, environment, and ecosystems from exposure to specific pollutants, toxic substances, and environmental hazards [6]. The process of risk assessment as used in public health and environmental settings is inherently a complex and highly technical process [7].

Risk must be determined both qualitatively and quantitatively from clarified facts and well-defined and accepted scientific approaches. Risk assessment provides a systematic scientific description of potential adverse effects from exposure to contaminants. The process should clearly define and describe hazards, exposure pathways, the nature of the adverse effects and their severity, uncertainties associated with data and their sources, as well as potentials for both the reversibility and preventability of risks [2, 7].

The anticipated risks posed by a specified problem or hazard needs to be expressed further in the context of multisource, multimedia, multichemical, and multirisk.

The cumulative risk assessment framework developed by the United States Environmental Protection Agency (USEPA) identifies procedure for the analysis, characterization, and possible quantification of the combined risks to health or the environment from multiple agents or stressors [2, 8].

Framework for Environmental Health Risk Assessment

Risk is determined from the concentration of the chemical or substances the person or the environment is exposed to, the rate of intake or dose, and toxicity of the chemical or substance [6]. Risk assessment uses either an engineering approach, focusing on the process control and safety, or toxicological approaches. A toxicological approach involves many assumptions and has limitations and uncertainties. Limitations of a toxicological approach are mainly related to the process uncertainties, difficulties associated with the correct estimate of the long-term, cumulative effects of low doses of toxins, and lack of sufficient and reliable data on the pathways and exposure levels [4].

The USEPA recommends a four-step procedure for assessing human health risk. The procedure consists of (1) hazard identification, (2) dose-response assessment, (3) exposure assessment, and (4) risk characterization [2].

As illustrated in Figure 1 the USEPA risk assessment procedure puts strong emphasis on the development of knowledge about the extent to which people are exposed to potentially harmful contaminants in their daily lives, level of the chemicals in the air, water, food, or other environmental media, the types of chemicals to which they are most often exposed, the source and levels of such exposures, the variability in exposures with time, and assumptions made about the mechanism and the rate of absorption of

the chemicals by the body. In order to eliminate the possibility of underestimating human risks, the procedure suggests a series of assumptions, preferably in a conservative range, about the chemical hazards, exposure scenarios, and the toxicological factors and constants [7].

An overview of the steps involved in the USEPA risk assessment procedure is presented.

Hazard Identification. Hazard is the inherent adverse effect that a chemical, process, or object poses with potentials for generating adverse consequences. Hazard both in its natural (geophysical and biological) and technological form can cause harms, such as human injury, illness, death, property damage, and ecological and environmental damage. Hazards identification is the quantitative measure of likelihood and severity of the harm [4, 9].

In terms of health risk assessment, hazard identification focuses on determining whether exposure to an agent can cause or increase a particular adverse health effect [9].

Hazard assessment calls for collection of scientific data on the characterization of the chemicals and hazards and the extent of contamination. Assessment evaluates both potentials for the release and types of contaminants. Since it is neither feasible nor cost effective to conduct risk assessment on all chemicals, hazard identification focuses on identifying chemicals of concern (COC) (*see What are Hazardous Materials?*). The common

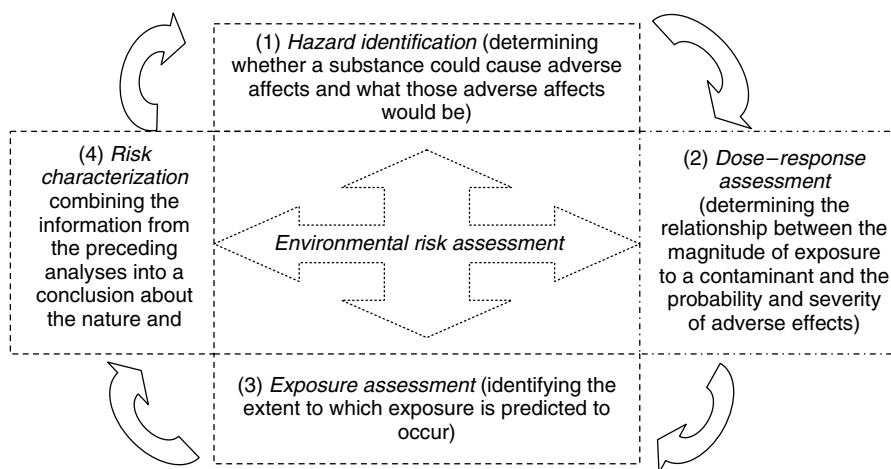


Figure 1 EPA's four-step process for quantitative risk assessment

methods used to select COC involves comparing information obtained from environmental samples with (a) detection limits, (b) laboratory artifact based on blank data, (c) natural background, and the frequency, and (d) concentration-toxicity and chemical-specific screening criteria set by the regulations [10].

In quantifying human health risks, with respect to the toxicity, hazards can be characterized as carcinogens and noncarcinogens. Since risk assessment process is different for carcinogens and noncarcinogens, hazard identification also involves determining whether or not COC have the potential to increase the incidence of cancer in exposed humans. Principal evidence for such an analysis is obtained from human or epidemiologic, and long-term, controlled animal studies (*see Cancer Risk Evaluation from Animal Studies*).

Dose–Response and Toxicity Assessment. Dose–response is the quantitative relationship between the dose and effect caused due to the exposure. Toxicity assessment evaluates toxicity of a chemical from the laboratory data and dose–response curve (*see Dose–Response Analysis*), a graphical representation of the relationship between the degrees of exposure to a chemical substance and observed or predicted biological effects or responses [9].

Toxicological assessment based on animal studies and/or human epidemiology studies provides quantitative measures of chemical concentrations or doses above which carcinogenic potency or risk is expected. Examples include dose–response constants such as slope factors (SFs) expressed in units of $\text{mg}\cdot\text{kg}^{-1}\cdot\text{day}^{-1}$ for carcinogens, and reference dose (RfD) expressed in units of $\text{mg}\cdot\text{kg}^{-1}$ of body weight/day for evaluating risks of noncarcinogenic effects. SF is a plausible upper-bound estimate of the probability of an individual developing cancer as a result of lifetime exposure to a particular level of a potential carcinogen. These mathematical constants are determined by the upper 95% confidence limit of the slope of the linearized dose–response curves [10, 11].

The dose–response assessments are also used to determine carcinogenicity and the threshold levels for the COC. Carcinogens are identified according to the weight of evidence from human epidemiologic studies and data from laboratory experiments. Probable or possible carcinogens are further classified as A, B, C, and D [11]. Threshold is the lowest dose or exposure

of a chemical at which a specified measurable effect is observed and below which such effect is not observed [9].

The data obtained during the dose–response assessments are almost always taken from laboratory tests involving animals exposed to high levels of a substance for a relatively short period of time. Extrapolation using linear or curvilinear projections is then made to estimate the level of risk associated with the exposure to such chemicals for human populations. Owing to the limitations with the data and assumptions made to estimate toxicity constants, this process involves a large level of uncertainties [12]. Toxicity assessment process, therefore, also defines the types of uncertainty inherent in the estimates of mathematical constants and the ways these uncertainties could affect the estimated values for risk [11].

The USEPA suggests the use of direct methods to measure the exposure in order to minimize and/or eliminate uncertainties associated with the estimates from mathematical models commonly used in exposure studies, extrapolations from experiments involving animals [1], and measurements of the chemicals in the environment [7].

Exposure Assessment. Exposure assessment is the qualitative or quantitative estimation or the measurement of the dose or amount of a chemical to which potential receptors have been exposed or could potentially be exposed to. Exposure assessment defines the magnitude, frequency, duration, route, and extent of exposure to the chemicals or hazards of concern [9].

Exposure pathways are the course(s) a chemical or physical agent takes from a source to an exposed population or organism. The avenue by which an organism contacts a chemical, such as inhalation, ingestion, and dermal contact, are defined as exposure routes. Evaluation and quantification of exposure in a given situation requires data and a set of assumptions about the exposure scenarios, hazard resources, exposure pathways, type and levels of chemicals, and the potential receptors.

Exposure assessments generally focus on individual chemicals, media, and potential sources. It can be simple or complex depending on the needs and specifications of risk assessment. Exposure can be estimated from the direct measurements or application of models that involve a range of assumptions about the source, chemical properties, and receptor characteristics [13].

During the course of exposure assessment, general and sensitive populations in risk are identified along with the sources of contamination and potential scenarios for the exposure. Scenarios are the conditions under which the population may be potentially exposed to the contaminants. The spatial distribution of contaminants, potentials for environmental fate and transport are also assessed and short- and long-term exposures in terms of doses by exposure routes are estimated [11].

Exposure can be defined as acute, chronic, or sub-chronic. Chronic exposure is the long-term low-level exposure to chemicals, i.e., the repeated exposure or doses to a chemical over a long period of time (exposure that can last weeks to years). Acute exposure is a large exposure to a chemical or dose over a short period of time (exposure for minutes to hours with some exceptions that allow acute exposures to encompass up to 3 months).

Subchronic exposure is defined as a short-term, high-level exposure to chemicals i.e., the maximum exposure or doses to a chemical over a portion of a lifetime (exposure that lasts days to weeks) [9, 10].

Mathematically (equations (1) and (2)), exposure can be estimated from the average concentration of the chemicals and the duration of exposure [14]:

$$E = \int_{t_1}^{t_2} C(t) dt \quad (1)$$

$$E = C \cdot \Delta t \quad (2)$$

where E is the exposure, C is the average chemical concentration (chemical concentration at the point of contact), and Δt is the duration of exposure. Exposure concentrations are typically expressed in units such as $\mu\text{g}\cdot\text{M}^{-3}$, $\text{mg}\cdot\text{m}^{-3}$, $\text{mg}\cdot\text{kg}^{-1}$, $\mu\text{g}\cdot\text{l}^{-1}$, or $\text{mg}\cdot\text{l}^{-1}$.

In different microenvironments (j), the exposure to any chemical i can be simply estimated from the filed information using equations (3a and b):

$$E_{j,k} = C_{j,k} \cdot \Delta t_{j,k} \quad (3a)$$

$$E_{i,j,k} = C_{i,j,k} \cdot \Delta t_{i,j,k} \quad (3b)$$

Where $E_{j,k}$ is the exposure of target person k to a given pollutant during the time interval Δt , and $C_{j,k}$ is the average concentration to which person k is exposed during the time interval Δt and in microenvironment j . The term $\Delta t_{j,k}$ represents the time spent by person k in microenvironment j [14].

For acute health effect Δt is a short interval that reflects the peak concentration for an effect. This equation can be expanded to multiple pollutants where the receptor (k th person) is exposed to a concentration of the i th chemical contaminant in the j th microenvironment ($C_{i,j}$). This concentration is constant during the interval $\Delta t_{j,k}$. The time integrated exposure or total exposure, E^T , can be also calculated for the exposure of the k th person (equation (4)) or total population ($k \rightarrow K$ from equation (5)) to the i th contaminant, over the possible microenvironments ($j \rightarrow J$):

$$E_{i,k}^T = \sum_j C_{i,j} \cdot \Delta t_{j,k} \quad (4)$$

$$E_{i,k}^T = \sum_k C_{i,j} \cdot \Delta t_{j,k} \quad (5)$$

Equation (6) assists estimation of a time-environmental integrated population exposure to a single contaminant for a population of persons and a series of microenvironments:

$$E_i^T = \sum_j \sum_k C_{i,j} \cdot \Delta t_{j,k} \quad (6)$$

The population averaged exposure (intake) of a chemical *via* an exposure pathway in units of $\text{mg}\cdot\text{kg}^{-1}\cdot\text{day}^{-1}$ can be estimated as follows:

$$I = C \cdot \left(\frac{CR}{BW} \right) \cdot FI \cdot \frac{EFD}{AT} \quad (7)$$

In this equation, C is the chemical concentration in the exposure medium, CR is the rate of chemical in medium contacted, BW is the body weight (average over exposure period), FI is the fractional intake from the exposure, EFD is exposure duration, and AT is the averaging time [14].

Risk Characterization. Risk characterization is a quantitative estimate of risks for the assumed exposure scenarios (source, chemicals, pathways, and receptor). During the risk characterization step, risk is estimated for the most probable exposed population and for the maximally exposed individuals. For example, carcinogenic risk is estimated by multiplying the estimated chronic daily intake dose (I_C) during the

exposure assessment by the value of SF from toxicity assessment:

$$\text{Risk} = I_C \times SF \quad (8)$$

The noncarcinogenic risk is normally characterized in terms of hazard index, which is calculated as the ratio of the estimated chronic daily intake of concentration (I_N) from exposure studies to the reference concentration (RfC):

$$HI = \frac{I_N}{RfC} \quad (9a)$$

$$HI = \frac{\text{Concentration}}{RfC}$$

$$HI = \frac{\text{Dose}}{RfD} \quad (9b)$$

Hazard index is a standard unit for quantifying noncarcinogenic risk. Noncarcinogens exhibit a threshold effect such as RfC and RfD below which they do not cause adverse health effects [11, 14]. An index greater than 1.0 indicates the possibility of an adverse health effect due to exposure.

Risk characterization can also identify all potential sources and their impact on the estimated risks, along with their relative positioning and distribution in relation to other risks to the community. Risk characterization addresses the hazard and the exposure, the nature and likelihood of the health risk; individuals or groups at risk; the likelihood and magnitude of risks; and the severity anticipated for the adverse impacts or effects. This step also involves evaluation and characterization of the scientific evidence supporting the conclusions about the risk with respect to their strength, accuracy, and uncertainty [14]. Risk characterization is also used to evaluate weather besides adverse effects on the health or the environment, if risk can impact social or cultural consequences [2].

Sources and Methods of Capturing Data

The quantitative risk assessment is a data-intensive process that requires collection of reliable data and selection of methodologies and models that matches the need of the risk assessment. Risk models commonly assume different scenarios for the source terms, and use equations and thermophysical

properties for the chemicals in question. The performance of the methods and risk models, however, needs to be characterized and verified in order to ensure their integrity and reliability. Evaluation and validation might include sensitivity testing for evaluating parameters with the greatest influence on the model output, model accuracy, precision, and predictive power [3, 15].

The principal methods of data collection for risk assessment programs include, but are not limited to, environmental sampling; laboratory studies; review of the historical data; analysis of the process, sources and the receptors; review and analysis of the chemical data; and assessment of the environmental parameters and potentials for chemicals fate and transport [3]. The success of environmental sampling programs depends on the proper sampling and analysis plans and consideration of a variety of factors in selecting sample locations and analytical detection limits. These programs must demonstrate clear methodologies for data collection, data verification, and evaluation of the accuracy and conservativeness of the risk analysis [10].

Data from both field observations and laboratory studies in controlled settings can be used in the risk modeling process. Laboratory and environmental (field) tests can provide strong evidence linking a chemical to observed adverse effects. Since laboratory data tends to be less variable than those from the field studies, responses may differ from those observed in the environment. Field studies, however, have potentials to provide environmental realism that laboratory studies lack [16].

Environmental data can also be obtained from data files or organizations that perform sampling, analysis, and data validation. Owing to the variable nature of the sampling and analysis, data is commonly screened and evaluated prior to being utilized in the risk assessment models. Toxicity values such as RfDs, RfCs, and SF can be estimated from independent research or received from the USEPA approved sources such as integrated risk information system (IRIS) [17]. The IRIS is an electronic database containing information on human health effects that may result from exposure to various chemicals in the environment. The system is a collection of computer files, which contain descriptive and quantitative information on oral RfDs and inhalation RfCs for chronic noncarcinogenic health effects and hazard identification, oral SF,

and oral and inhalation unit risks for carcinogenic effects [17, 18].

Since data could be incomplete or vary depending on the source of information, particularly in the areas of toxicity and human errors, an expert opinion is necessary to assess the existing data and their suitability for a quantitative risk analysis.

Risk Management

Environmental risk analysis involves both risk assessment and management of natural and technological hazards. Risk assessment refers to the technical assessment of the nature and magnitude of risks. Risk management, however, is the process of evaluating and selecting appropriate responses for mitigating risk and the consequences [4].

Risk assessment is used to determine health guidance values for environmental hazards. It is also used to establish regulatory exposure standards to adequately protect public health and the environment. Risk management, however, is the process of weighing policy alternatives and selecting the most appropriate action through integrating results of risk assessment with social, economic, and political concerns [1, 4]. Risk management often involves the public; it requires examination of the term *acceptable risks* and the perception of the assessed risks [11] (see **Risk and the Media; Stakeholder Participation in Risk Management Decision Making**).

Having characterized the risk and the degree of uncertainty associated with the risk, the best decisions must be made in order to eliminate and/or mitigate the risks. Effective risk management must include the stakeholder perceptions of risk and any other social or cultural impacts associated with risks posed by a specified problem and/or hazards [2].

Integration of the stakeholder perceptions of the risks with scientific and contextual aspects of the risks to human health or the environment assists policymakers with development of guidelines for effective risk management and mitigation.

Risk Communication

Risk communication is the continental exchange of information about health and environmental risks among risk assessors, risk managers, the public, the media, interested groups, and others. It is

possible to arrive at a meaningful idea of success for risk communication by considering a broad public purpose that successful risk communication serves [19].

References

- [1] Leonard, E.J. & Robinson, G.D. (2002). *Managing Hazardous Material*, Institute of Hazardous Materials Management, Rockville, MD, pp. 160–170.
- [2] The Presidential/Congressional Commission on Risk Assessment and Risk Management (1997). *Risk Assessment and Risk Management in Regulatory Decision Making*, Final Report; Volume 2, pp. 2, 3, 5, 21, 200.
- [3] Kolluru, R., Bartell, S., Pitblado, R. & Stricoff, S. (1996). *Risk Assessment and Management: Hand Book for Environmental, Health, and Safety Professionals*, McGraw-Hill, pp. 8–34.
- [4] Chechile, R.A. & Carlisle, S. (1991). *Environmental Decision Making: A Multidisciplinary Perspective*; Tuft University Center for the Study of Decision Making, Van Nostrand Reinhold, pp. 217–219.
- [5] Crosby, D.G. (1998). *Environmental Toxicology and Chemistry*, Oxford University Press, p. 143.
- [6] Burke, G.H., Singh, B.R. & Theodore, L. (2000). *Handbook of Environmental Management and Technology*, 2nd Edition, John Wiley & Sons, pp. 36, 661–665.
- [7] Human Health Risk Assessment, Committee on Environment and Public Works. (2006). *EPA; United States Government Accountability Office Report to the Chairman*, GAO-06-595, U.S. Senate, pp. 8, 10.
- [8] Framework for Cumulative Risk Assessment (2003). *Risk Assessment Forum*, EPA/630/P-02/001F, U.S. Environmental Protection Agency, Washington, DC, p. xvii.
- [9] Kofi, A.D. (1993). *Hazardous Waste Risk Assessment*, Lewis Publishers, pp. 283–293.
- [10] Air and Waste Management Association (1995). *Risk Assessment and Risk Communications Strategic Tools*, Environmental Toxicology International, Kansas City, pp. 19–22.
- [11] LaGrega, M.D., Buckingham, P.L. & Evans, J.C. (2001). *Hazardous Waste Management*, McGraw-Hill, pp. 872–875, 887–888, 892.
- [12] Manaham, S.E. (1990). *Hazardous Waste Chemistry: Toxicology and Treatment*, Lewis Publishers, pp. 102–105.
- [13] Commission on Risk Assessment and Risk Management (1996). *Risk Assessment and Risk Management in Regulatory Decision-making*, Draft Report for Public Review and Comment, pp. 16–18, 28.
- [14] Gratt, L.B. (1996). *Air Toxic Risk Assessment and Management: Public Health Risk from Normal Operation*, ITP® A Division of International Thomson Publishing, Inc., Van Nostrand Reinhold, pp. 109–114, 144.
- [15] Science and Judgment in Risk Assessment (1994). *Committee on Risk Assessment of Air Pollutants*, National

Research Council, National Academy of Press, Washington, DC, pp 106–112.

- [16] Framework for Ecological Risk Assessment (1992). *Risk Assessment Forum*, EPA/630/R-92/001, U.S. Environmental Protection Agency, Washington, DC, pp. 21–23.
- [17] U.S. Department of Energy Office of Environmental Management (1997). *Risk Assessment Program Data Management Implementation Plan ES/ER/TM-232, Prepared by Environmental Restoration Risk Assessment Program*, Part 2, pp. 4–6.
- [18] USEPA (2007). *Integrated Risk Information System*, <http://www.epa.gov/iris/intro.htm>.
- [19] Improving Risk Communication (1989). *National Research Council, Committee on Risk Perception and Communication*, National Academy of Sciences, p. 27.

Related Articles

Environmental Health Risk Assessment

Scientific Uncertainty in Social Debates Around Risk

Role of Risk Communication in a Comprehensive Risk Management Approach

NASRIN R. KHALILI

Environmental Health Risk Assessment

General Considerations

Quantitative health risk assessment is a tool for making predictions about the impact of substances in the environment on health. Despite the use of scientific principles and a reliance on the most appropriate data from the scientific literature, risk assessment, in the strict sense of the word, is not, by itself, a science. This is because the predictions of risk and/or safety that derive from a risk assessment are rarely verifiable. This is the case because the risks that are generally addressed are at, or just above, the *de minimis* level. That is, risk assessments are generally carried out in order to specify a level of acceptable environmental exposure that poses little or no significant elevation of risk. The residual risk remaining after risk-based controls are put in place should, therefore, rarely be measurable above background levels. Risks that are identified through a risk assessment process as having the potential to pose more than a *de minimis* level of risk are generally reduced without waiting for statistically verifiable evidence of impacts on public health. When environmental conditions are such that they result in observable and quantifiable levels of an adverse health effect in a population (as opposed to predicted or potential levels of effect), then the prediction of risk through the application of formal risk assessment is largely superfluous. Furthermore, nearly all health effects known to result from exposure to environmental contaminants are multifactorial in origin, and can occur from genetic and lifestyle factors in the absence of significant exposure to environmental contaminants. Thus, in many cases, the increase in the incidence of a given disease resulting from an environmental exposure that may be predicted from a risk assessment is within the confidence limits of the background incidence of that disease in the population. Health risk assessment is, therefore, a predictive tool rather than an experimental or even an observational science. In practice, risk assessment is used both to predict the inherent risk (or safety) associated with a substance (e.g., what is the risk associated with a unit exposure to benzene? what is the safe dose of methyl mercury?), and to describe the risk associated with a given scenario (e.g., what

is the risk from drinking water containing $x \mu\text{g l}^{-1}$ of arsenic?) (see **Water Pollution Risk**).

Despite more than 50 years of the practice of environmental health risk assessment, its practice and application continue to provoke controversy. On a fundamental level, the notion of quantitative risk is not well understood by the public (see **Role of Risk Communication in a Comprehensive Risk Management Approach; Risk and the Media**), and so the results of risk assessments are often misinterpreted and misused. This is compounded by the fact that risk assessment is a highly interdisciplinary practice. The synthesis of an overall assessment of risk across disciplines in the face of multiple uncertainties makes the communication of the findings and their implications difficult. Furthermore, as a tool to inform regulatory decision making, risk assessment is used by different agencies and groups with differing agendas. This can lead to controversy over its proper goals and application and thus there is often no obvious consensus as to what the appropriate goal of a risk assessment should be. In addition, the use of risk assessment for standard setting and environmental regulation is seen by some as reflecting a willingness to accept environmental risk (see **Environmental Hazard**). For those with such a viewpoint, the precautionary principle is often cited as a more appropriate and uncompromising alternative. The two approaches, however, are not necessarily incompatible, and there are situations such as those requiring risk–benefit balancing (e.g., consumption of moderately contaminated fish) where there is no obvious role for the precautionary approach. Just as there are those for whom environmental health risk assessment is, by its nature, insufficiently protective, there are those, often in the regulated community, for whom it is unnecessarily conservative (i.e., protective). This point of view arises in response to the various conservative default assumptions in much of the risk assessments carried out by governmental agencies. To a large extent, these conservative assumptions are intentional and represent a public health protective response to the uncertainty inherent in many steps in the risk assessment process. Nonetheless, the proliferation of such assumptions can result in an unintended level of conservatism in the overall assessment. This has led to several approaches that attempt to restrict the level of conservatism within intended bounds (see below).

Two different forms of quantitative risk assessment are practiced, often by the same governmental agencies, and interested groups that seek to influence the policies and decisions of those agencies. These two different forms have somewhat different orientations and address different regulatory and communication goals. One form of risk assessment can be referred to as *safety assessment*. This type of assessment seeks to identify the largest level of exposure or dose that can be reasonably expected to be without adverse effect. Safety assessment does not necessarily provide quantitative information on the risk associated with exposure above this “safe” level. In the United States, such assessments are most commonly applied to substances judged to be noncarcinogenic, or to the noncarcinogenic effects of carcinogens (*see Environmental Carcinogenesis Risk*). The rationale for this application is based on the long-standing notion that, for most types of noncancer toxicity, there is a dose (the threshold dose) below which no significant adverse effects will be experienced even over a lifetime of exposure (*see Threshold Models*). In the United States, the most common examples of safety assessment are the United States Environmental Protection Agency’s (US EPA’s) reference dose (RfD) for oral exposure or reference concentration (RfC) for inhalation exposure. The RfD is defined as “an estimate (with uncertainty spanning perhaps an order of magnitude) of a daily oral exposure to the human population (including sensitive subgroups) that is likely to be without an appreciable risk of deleterious effects during a lifetime” [1]. The RfC is similarly defined as “an estimate (with uncertainty spanning perhaps an order of magnitude) of a continuous inhalation exposure to the human population (including sensitive subgroups) that is likely to be without an appreciable risk of deleterious effects during a lifetime” [2, 3]. The RfD is expressed in units of dose (e.g., milligrams of substance per kilogram body weight per day), while the RfC is expressed in units of concentration in air (e.g., milligrams of substance per cubic meter of air). In some other countries, particularly in Europe, this type of assessment is also applied to carcinogens. More recently, the notion of a sharp threshold for adverse effects has been challenged by the recognition of variability in response within human populations as well as the apparent continuous effect of some toxicants through the lowest doses observable (e.g., the neurodevelopmental effects of gestational lead exposure). The population

variability in response is best seen in epidemiological studies when dose (or exposure) response data (*see Dose–Response Analysis*) are viewed without grouping into arbitrary exposure categories. In contrast to the much neater collection of observations generated in classical animal toxicology studies, such dose–response data typically appear as a point cloud. A good example of this is the data on the association of neurodevelopmental test scores with gestational methyl mercury exposure [4]. Both the variability in response and the continuous dose–response can make the notion of identifying a clear threshold for adverse effects a practical impossibility.

The other type of risk assessment seeks to quantify the risk associated with any exposure or dose of an environmental agent. In the United States, this form of risk assessment is generally applied to carcinogens. The most common example of this type of assessment is the cancer potency slope, which (as employed by the US EPA) gives the probability of developing a dose-related tumor for a unit dose (e.g., per milligram of carcinogen per kilogram body weight per day). Historically, in contrast to the threshold assumption applied to noncancer safety assessment, cancer risk has been assumed to be linear and to have no threshold (i.e., the risk is assumed to be zero only at zero dose). This assumption was based on the idea that a heritable genetic change in the DNA is a necessary causative step in carcinogenesis and that, in principle, one molecule of a carcinogen has a small but finite chance of causing such a change (*see Gene–Environment Interaction*). The most recent version of the US EPA’s Guidelines for Carcinogen Risk Assessment [5] still employs a linear no-threshold approach as the default assumption (although linearity is assumed only between the origin and a “point of departure” defined by the observable dose–response data) for chemicals, for which the mode of action of carcinogenesis is not known, and for carcinogens with an established mutagenic mode of action. However, for carcinogens with clearly defined modes of action that lead to nonlinear dose–response models and/or dose–response models with quantifiable thresholds, cancer risk may be based on threshold and/or nonlinear models rather than on the linear assumption. In practice, given the evidence required to establish a specific mode of action, departures from the linear, no-threshold assumption are still rare. A notable recent exception to this is chloroform. On the basis of a detailed mode of action

information indicating that the primary mechanism for the carcinogenicity of chloroform is based on cell lethality and regenerative hyperplasia, the US EPA has addressed the carcinogenicity of chloroform through the same noncancer (safety assessment) risk assessment approach used to derive the RfD that is, by itself, intended to be protective against chloroform-induced cell lethality [6].

It has been recognized for some time that the distinction between the determination of noncancer risk through safety assessment and cancer risk through a continuous dose–risk relationship is largely artificial and an artifact of the history of the development of risk assessment. Movement toward a unified cancer/noncancer risk assessment framework (sometimes referred to as *harmonization*), however, has been slow. In part, this is because various risk-based regulatory policies have been developed with specific reference to one or the other risk assessment approach. The US EPA's recent approach to risk assessment for chloroform, applying a safety assessment approach to protect against noncancer effects that occur “upstream” of tumor production, is an important step in the direction of reconciling these two approaches.

Another controversial aspect of health risk assessment is that health risk can include a very broad range of effects – from mild irritation to lethality. Not all risks are of equal severity or concern (even if they occur at the same dose); therefore, risk is a function both of the likelihood of occurrence and of severity. It has proven difficult to arrive at a common metric of severity. For example, while it may be readily agreed that progressive kidney dysfunction is a more severe outcome than mild, transient neurological effects, it is not clear in objective quantitative terms how much more severe an outcome it is. For these reasons, environmental health risk assessment is, in practice, defined either as the process of estimating the probability of a given adverse outcome (most usually cancer) at a given dose or as the determination of the largest dose that will not result in *any significant* adverse outcome. It is generally acknowledged that some effects may be adverse, but insufficiently significant from a public health standpoint to support guidelines or standards. The issue of what constitutes a “significant” adverse outcome is somewhat subjective. In practice, even effects that do not, strictly speaking, impair health, such as mild dental mottling or skin discoloration,

have formed the basis for risk-based standards and guidelines. Mild transient health effects (e.g., mild skin irritation or itching), on the other hand, have generally not been considered “significant” adverse effects.

One of the major limitations of environmental health risk assessment that has only recently begun to be fully appreciated is the fact that historically, this tool was developed to address the risk associated with exposure to individual substances. Humans, however, exist in an ocean of stressors with a myriad of exposures occurring at any point in time. Given ubiquitous coexposures, the risk from coexposures to a specific substance may be significantly greater or less than that calculated on a single substance basis. While some combinations of exposures are known to result in additive or synergistic risks (e.g., smoking and asbestos) there are few data that are useful for estimating the risk from even common combinations of different environmental substances. For some closely related chemicals with common mechanisms of toxicity, however, metrics of relative potency (toxic equivalency factors or TEFs and relative potency factors or RPFs) have been developed that express the overall potency of a mixture in terms of an equivalent concentration or dose of the most potent constituent. TEFs have been developed for PCDDs (polychlorinated dibenzo-dioxins), and related PCDFs (polychlorinated dibenzo-furans) and coplanar PCBs (polychlorinated biphenyls) relative to 2,3,7,8-TCDD (tetrachlorodibenzo-p-dioxin) [7], and also for carcinogenic PAHs (polycyclic aromatic hydrocarbons) relative to benzo-*a*-pyrene [8]. RPFs have been developed for organophosphate pesticides [9]. Various approaches have been suggested for estimating cumulative risk from chemicals that are less closely related structurally or through their mechanisms [10], but not surprisingly, given the range of possible interactions, progress in this area has been slow (*see What are Hazardous Materials?; Environmental Risks*).

Despite uncertainties, limitations, and various areas of controversy, quantitative health risk assessment is an essential tool for the setting and carrying out of environmental policy. It provides a scientifically based and practical, if imperfect, metric for setting regulatory policy for the protection of public health, for prioritizing environmental action, and for comparing risks.

The Structure of Health Risk Assessments

Since its publication in 1983, the US National Research Council's report, *Risk Assessment in the Federal Government: Managing the Process* (often referred to as the *Red Book*) has defined the structure of a formal health risk assessment [11]. This structure includes four key steps: hazard identification, dose–response assessment, exposure assessment, and risk characterization. Each of these steps is discussed in more detail.

Hazard Identification

In this step (sometimes referred to as *hazard assessment*), both the substances (or more generally, the stressor) in question and the health endpoints resulting from exposure to the substance are determined and documented. The identification of the substance in question is not necessarily a trivial matter. Chemicals may be complex and variable mixtures of related molecules. PCBs (see **Polychlorinated Biphenyls**) and PCDDs, for instance, include many different congeners, each with different numbers and arrangements of chlorine atoms. Each congener and each mixture of congeners may have qualitatively and quantitatively different potencies. Metals, such as mercury (see **Methylmercury**), arsenic (see **Arsenic**), chromium (see **Hexavalent Chromium**), and nickel can have different valence states each of which can form different compounds that differ markedly in their toxicological properties. In addition, even well-characterized compounds may commonly contain chemical impurities that, themselves, have toxicological properties. Care must be taken in specifying, as precisely as possible, the substance or mixture that is under study. Although sometimes overlooked, equal care must be taken in not applying risk assessments to exposures that do not clearly correspond to the specific substance addressed in the assessment.

Hazard identification should reflect a thorough review of the relevant scientific literature for the substance. In addition to toxicologic studies *per se*, this can also include pharmaceutical/pharmacological studies, medical studies, and nutritional studies. The review should focus on the absence of adverse effects as well as on their occurrence. Studies should be evaluated for their overall quality, their physiological and toxicological relevance to humans (e.g., does an effect observed in an animal study reflect a

toxicological mechanism that is likely to be present in humans?), their relevance to potential human routes of exposure (e.g., are toxicological findings from an inhalation study relevant to risk from ingestion?), and their ability to support quantitative dose–response assessment. Studies that cannot support quantitative dose–response assessment can still be qualitatively useful in supporting the relevance of other studies. This review of the scientific literature results in the identification of a key study or of several closely related studies. In addition, the review should identify and describe, if possible, a mode of action for the critical effect. The mode of action describes the key steps leading to the critical effect, but is a less precise description than a formal toxicological mechanism at the molecular level. In addition to assisting in the evaluation of the critical effect in terms of its relevance to humans and its potential to progress to more severe effects, identification of a mode of action can influence the nature of the dose–response analysis for carcinogens under the current US EPA's Guidelines for Carcinogen Risk Assessment [5] (see above).

Generally, in hazard identification, a single health endpoint (the critical effect) is chosen from the key study as the basis for the dose–response and exposure assessments that follow. As a general rule, the critical effect is the most sensitive adverse health outcome, i.e., the significant health effect that occurs with the lowest dose. Judgment must, however, be exercised in this selection, since the most sensitive endpoint may be of doubtful health significance (e.g., enzyme induction, minimal organ weight gain, or minimum whole body weight loss), may not be well documented, and/or may not have sufficient quantitative data to support dose–response assessment. For example, in its 2000 report, the US National Research Council on the toxicology of *methyl mercury* identified cardiovascular effects as possibly the most sensitive health endpoint, but chose neurodevelopmental effects as the critical effect, because they were more extensively studied and documented [4]. Approaches for combining several health endpoints of varying severity have been proposed [12], but have not been used for the formal derivation of guidelines and standards.

Dose–Response Assessment

The overall goal of the dose–response assessment is to derive a generalizable quantitative relationship between the dose (or exposure) and the occurrence

or risk of the critical effect identified in the hazard identification step. A good general discussion of the dose–response process is given in the online background information to the US EPA’s Integrated Risk Information System (IRIS) (<http://www.epa.gov/ngispgm3/iris/rfd.htm>). Given the historical dichotomy between the cancer and noncancer risk assessment approaches, the specific goals of dose–response assessment for each type of assessment have been fundamentally different. For cancer risk assessment, the goal of *dose–response analysis* has been to estimate the risk at any given dose through the fitting of a dose–response model to the empirical data. In practice, the cancer dose–response data have largely been animal chronic bioassay data, in which the doses have been far in excess of human environmental doses, and the response rates have been far in excess of those that are appropriate for setting standards and guidelines for protection of public health (see **Cancer Risk Evaluation from Animal Studies**). The observed response rates of tumors in animal tests is generally in the whole number percentage range, while standards and guidelines are generally set on the basis of risks in the 1×10^{-4} to 1×10^{-6} range. This means that mathematical dose–response models are fit to the observed data in the very high-dose range, and extrapolated downward across 3–5 orders of magnitude to the very low dose range where little information on the nature of the dose–response relationship exists. Estimates of cancer risk are thus often driven more by generic assumptions about the type of dose–response model that is applied than by the empirical toxicological data. On account of large uncertainties about the shape of the dose–response curve in the (low) dose range of interest, and because of a general lack of information about the mechanisms underlying carcinogenicity in the low-dose range, standards and guidelines for carcinogens have traditionally been based on an admittedly conservative, linear, no-threshold dose–response model. Only recently has the US EPA begun to entertain the use of nonlinear and threshold dose–response models [5] (see above).

When human epidemiological data on cancer incidence and exposure data in a well-defined cohort are available, such uncertainties can be reduced, although the observed risks must be sufficiently large to be detectable and statistically meaningful, given the available cohort size, exposure range, and background incidence of the cancer in the population.

The exposures (doses) encountered in epidemiologic studies, although often originating from relatively contaminated occupational environments, are, by definition, within the range of human exposures. Thus, when epidemiological data are used as the basis for cancer dose–response assessment, the uncertainties generally originate less from issues of the selection of the dose–response model or high-dose to low-dose extrapolation than from considerations of study power and exposure misclassification.

As with cancer risk assessment, much of the data for noncancer (safety) risk assessment derives from animal studies. In noncancer risk assessment, the goal has been the assessment of a safe dose rather than a determination of the risk at any given dose. This has led to an approach to dose–response assessment that has, until recently, concentrated on the lowest portion of the dose–response curve to the virtual exclusion of the information contained in the remainder of the curve. In this approach, the largest dose at which there is no significant increase in the incidence of an adverse effect above the background incidence is defined as the no observed adverse effect level (NOAEL) dose [1]. When no such dose can be identified, the smallest dose at which a statistically significant increase in the incidence of an adverse effect is observed (the lowest observed adverse effect level or LOAEL) is identified. The NOAEL or LOAEL then becomes the point of departure for setting the virtually “safe” dose. The possible values that the NOAEL or LOAEL can assume are limited to the experimental doses (or in an epidemiological study with grouped data, the arbitrary exposure groupings). The NOAEL is thus dependent on the size of the study (i.e., the number of dose groups) and on the range and spacing of the doses in the original study. NOAELs (and LOAELs) for the same or different substances derived from different studies are, therefore, not easily comparable. The shape and slope of the overall dose–response relationship is largely irrelevant in this approach and much potentially useful information remains unused. Furthermore, since NOAELs and LOAELs can only be derived from epidemiologic data that have been binned into discrete dose or exposure categories, this approach also requires loss of valuable information when applied to epidemiologic data derived from studies with continuous measures of exposure or dose.

During the last two decades, the *benchmark dose estimation* (see **Benchmark Dose Estimation**)

approach has gradually gained acceptance as an alternative in dose–response analysis [13–15]. Initially, the benchmark dose approach was applied to noncancer risk assessment, and more recently, to cancer risk assessment for identifying a point of departure in linear extrapolation [5]. In this approach, a dose–response model is fitted to the empirical data, and the probability of a small but observable response (e.g., 1%, 10%) is chosen. The lower confidence limit on the dose corresponding to that probability is then identified as the *point of departure* for setting the virtually “safe” dose (e.g., an RfD). The original intent of such a point of departure was to provide a generalizable (i.e., study-independent) estimate of the NOAEL. However, it is clear that such a value is not conceptually equivalent to the NOAEL. In part, this is because the NOAEL is, by definition, study-dependent, and therefore, the notion of a generalized NOAEL is not clearly defined. This is also the case because the benchmark dose is, by definition, not a no-effect level; rather, it is a lower bound estimate of the dose corresponding to a 1, 5, 10%, etc. effect level. Various uncertainties in the application and interpretation of this approach still need to be resolved. Among these is the fact that, in most cases, the dose–response functions applied to the observed data are chosen from among a standard set of functions (e.g., Weibull, gamma, logistic) identified on the basis of their mathematical utility. The particular mathematical function ultimately used to derive the benchmark dose is generally chosen on the basis of its suitability to the data rather than on the basis of its underlying biological plausibility. Additionally, as discussed above, the public health interpretation of the benchmark dose remains uncertain.

Efforts have been underway for at least two decades to harmonize cancer and noncancer dose–response assessment [16]. This effort is complicated, however, by the fact that cancer and noncancer dose–response assessments tend to have data on opposite extremes of the dose–response curve as their starting points. While, for reasons of practicality in study design, cancer dose–response often starts from unrealistically high doses, noncancer dose–response seeks to identify doses spanning the threshold for response.

One approach that has received considerable attention involves the identification of noncancer endpoints for carcinogens. This approach is exemplified by US EPA’s recent treatment of chloroform

as discussed above [6]. Such effects (enzyme induction, hyperplasia, DNA adduct formation, mutagenesis, etc.), which may be predictive of eventual tumor formation, can then be modeled using noncancer dose–response techniques such as the benchmark dose [15]. Such an approach holds promise but raises questions about the extent to which such responses are necessarily adverse and truly predictive of carcinogenicity. Such approaches cannot be generalized to all carcinogens and will likely find application on a case-by-case basis. A more broad-based approach to harmonization of cancer and noncancer dose–response does not appear likely in the foreseeable future.

Exposure Assessment

The role of the exposure assessment portion of environmental health risk assessments depends on the purpose and application of the risk assessment. In predictive risk assessments, the purpose of the assessment is to predict the risk that will occur under a future or hypothetical set of exposure conditions. Examples of predictive risk assessments are those intended to set generic acceptable levels of a contaminant in drinking water and those intended to set generic acceptable levels of contaminants in soil. Since these exposures are not occurring in the present, they cannot be measured, and so must be predicted under a specific exposure scenario. In such cases, the exposure assessment is required to quantitatively characterize those aspects of the scenario that determine the dose. This is generally accomplished with a model that integrates the dose over time to predict the cumulative dose. Such models require identification of the relevant routes of exposure (ingestion, inhalation, dermal absorption), the receptors (children, adults, workers, pregnant women, etc.), the concentration of the contaminant in the relevant media (soil, air, water, and food), the uptake or intake rate for each medium (e.g., grams of soil ingested per day, cubic meters of air inhaled per day, liters of water ingested per day), the frequency of exposure (continuous, daily, monthly, etc.), the duration of exposure (lifetime, childhood, occupational lifetime, single day, gestation, etc.), and factors such as the bioavailability of the substance in the particular medium that can alter the nominal dose [17]. Such scenarios, even when realistic, are often based on assumed or hypothetical populations with typical activity patterns. Generic parameters for common

activities associated with environmental exposures can be found in the US EPA's Exposure Factors Handbook [18]. In contrast, descriptive risk assessments seek to describe the ongoing risk from measurable exposures in an actual setting. Examples of descriptive risk assessments are those intended to characterize workplace risks and those describing the risk to consumers of specific species of contaminated fish from a specific water body. Although the exposures in such cases are theoretically measurable on a real-time and individual basis, it is more common for exposures to be characterized by temporally and geographically discrete sampling of the environment or by using generic exposure assumptions such as the ones discussed above. Characterizing the exposure of specific populations obviously requires more effort, is more data intensive, and is more expensive than applying generic exposure assumptions. Thus, even when exposures are ongoing, descriptive risk assessments tend to rely on such generic assumptions.

When generic exposure assumptions are used, point estimates (single numerical values) have traditionally been assigned to each exposure variable (e.g., the volume of daily water consumption). Recognizing, however, that each variable can assume a range of possible values, attempts have been made to select upper percentiles for exposure variables so as to be inclusive of most of the potentially exposed population. Such values can often be located in the US EPA's Exposure Factors Handbook [18]. When dealing with the suite of exposure variables in an exposure model, however, the mathematical combination of upper percentile values in the exposure model can quickly compound into an extreme upper percentile estimate of the overall exposure. This has been referred to as *redundant* or *compounding conservatism* [19, 20]. Various combinations of central tendency estimates and upper percentiles have been proposed to remedy this situation [21], but such approaches are necessarily arbitrary, and result in exposure estimates with an unknown degree of population inclusiveness. These problems arise because no single value can adequately represent variables that are, in fact, distributed quantities. Over the last decade, probabilistic (largely Monte Carlo) approaches have been proposed to address the underlying variability and uncertainty in exposure assessment [22, 23]. Such approaches allow for realistic descriptions of the full range of exposure variables and yield a distributional description

of the integrated exposure estimate. This distributional description of the overall exposure estimate allows risk managers to select the exposure and related risk at a specific percentile of the estimated population distribution (e.g., 90th, 95th, 99th upper percentile of exposure) for a given exposure scenario. Probabilistic approaches, therefore, hold the potential for more realistic (if still generic) exposure assessment. The acceptance of such approaches has, however, been slow among regulatory agencies. In part, this is because the probabilistic description of many key exposure parameters is uncertain. This uncertainty stems from the lack of information about the specifics of distributions, particularly in their extreme values. Another factor in the slow progress in using probabilistic approaches is that differences among populations, including regional and ethnic differences, make the derivation of generic distributions difficult since different populations may not lie on the same continuous distribution. In principle, probabilistic approaches are also applicable in dose-response assessment. However, uncertainty regarding the probabilistic description of several key aspects of dose-response extrapolation (e.g., animal-to-human, subchronic-to-chronic, LOAEL-to-NOAEL) as well as the general lack of data on interindividual human variability in dose-response have traditionally limited probabilistic approaches to the exposure assessment portion of risk assessments (see **Low-Dose Extrapolation**).

When risk assessments are conducted to characterize the inherent risk associated with a substance (rather than to characterize a specific exposure scenario), the role of exposure assessment is to relate the toxicokinetics underlying the key dose-response study or studies to the conditions of human exposure that may ultimately be encountered. Thus, for example, in an animal study where the route of exposure may be ingestion or injection, it may be necessary to apply appropriate toxicokinetic adjustments to address human inhalation exposure. Or, in a study in which the dose to the test animals was contained in food, it is necessary to consider how the bioavailability of the test substance in the food would affect the dose-response when considering human exposure from drinking water. Such toxicokinetic considerations are often based on models of varying complexity and realism (e.g., one compartment, multicompartment, physiologically-based pharmacokinetic). Traditionally, such considerations have been

addressed using point estimates for the various variables in toxicokinetic models. However, biological parameters (e.g., body weight, half-lives, blood volume) are truly distributed quantities reflecting population variability. Recently, probabilistic approaches have begun to be applied to such toxicokinetic models (see, for example [24] and [25]).

Risk Characterization

The risk characterization portion of the risk assessment process is intended to synthesize the outcomes from the earlier portions of the assessment into an overall quantitative estimate of the risk, to quantitatively address uncertainties in the estimate, and to provide a qualitative judgment of the overall confidence in the estimate. Having completed the dose–response and exposure assessment steps, the generation of the quantitative risk estimate is generally a straightforward process that involves applying the dose estimate from the exposure assessment to the dose–response relationship derived in the dose–response assessment. Quantitative treatment of uncertainties in the assessment (essentially the consideration and application of a series of adjustments) has traditionally been confined to the dose–response portion of noncancer risk assessments. The lack of such adjustments in cancer risk assessments reflects the general consensus that common cancer dose–response models already employ conservative assumptions and require no further adjustments to be sufficiently protective. In noncancer risk assessments, uncertainty adjustments have generally been applied according to the categories defined by the US EPA [26]. These are interindividual variability in sensitivity among healthy humans; interspecies differences in sensitivity (animal-to-human extrapolation); extrapolation from the results of subchronic exposure studies to lifetime exposure; extrapolation from a LOAEL dose to a (hypothetical) NOAEL; and a general modifying factor to address considerations of quality and reliability of the critical study (e.g., database insufficiency, small study size). Traditionally, uncertainty has been quantitatively adjusted separately for each category *via* the use of generic factors of 10, and, less commonly, 3. For example, a LOAEL derived from an animal study might be divided by three factors of 10 (i.e., a single factor of 1000) to account for (a) LOAEL-to-NOAEL, (b) animal-to-human, and (c) average human-to-sensitive human extrapolations. The original rationale for the use of

values of 10 (or 3) for these adjustments appears to have been rather arbitrary. Attempts to provide an empirical and probabilistic basis for such uncertainty adjustments are ongoing [12, 27–29], and appear to provide some (after-the-fact) justification for the traditional approach. Even with uncertainty adjustments generally limited to a maximum of 3000 or 10 000, in practice, the final estimate of a noncancer risk may be driven more by the uncertainty adjustments than by the dose–response data from the critical study.

The description of the overall confidence in the assessment is generally qualitative, and based on professional judgment about the completeness of the overall database (including the availability of data on reproductive, and developmental effects as well as chronic toxicity), the appropriateness of the critical effect, the adequacy of the design and execution of the key study, the data quality, and the appropriateness of the models employed. For noncancer assessments, confidence is generally ranked as high, medium, or low. For cancer assessments, the weight of evidence has historically been ranked according to one of several scales [5, 30, 31] that express the likelihood that the substance is a human carcinogen independent of the quantitative estimate of cancer risk. This determination involves several factors, including whether there is adequate evidence of human carcinogenicity from human studies, the extent to which the substance is carcinogenic in various species strains and/or sexes of test animals, the extent to which tumors observed in the animal models are physiologically and biochemically applicable to humans, and data from *in vivo* and *in vitro* studies, which give insight into the compound's mode of action.

References

- [1] Barnes, D.G. & Dourson, M. (1988). Reference dose (RfD): description and use in health risk assessment, *Regulatory Toxicology and Pharmacology* **8**, 471–486.
- [2] Jarabek, A.M. (1995). The application of dosimetry models to identify key processes and parameters for default dose-response assessment approaches, *Toxicology Letters* **79**, 171–84.
- [3] USEPA (1994). *Methods for Derivation of Inhalation Reference Concentrations and Application of Inhalation Dosimetry*, Research Triangle Park, North Carolina, EPA/600/8-90-066F.
- [4] National Research Council (2000). *Toxicological Effects of Methylmercury*, National Academy Press, Washington, DC.

- [5] USEPA (2005). *Guidelines for Carcinogen Risk Assessment*, Accessed at <http://www.cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=116283&CFID=61473&CFTOKEN=35305184&jsessionid=f430fb68fd9649642154TR> (accessed Nov 2006).
- [6] USEPA (2006). *Integrated Risk Information System (IRIS), Chloroform*. Accessed at <http://www.epa.gov/iris/subst/0025.htm> (accessed Nov 2006).
- [7] USEPA (2003). *Exposure and Human Health Reassessment of 2,3,7,8-Tetrachlorodibenzo-p-Dioxin (TCDD) and Related Compounds. NAS Review*, Washington, DC. Draft, EPA/600/P-00/001Cb, December.
- [8] USEPA (1993). *Provisional Guidance for Quantitative Risk Assessment of Polycyclic Aromatic Hydrocarbons* EPA/600/R-93/089. Environmental Criteria and Assessment Office, Office of Health and Environmental Assessment, Cincinnati, OH 45268.
- [9] USEPA (2006). *Organophosphate Pesticides: Preliminary OP Cumulative Risk Assessment*, Accessed at http://www.epa.gov/pesticides/cumulative/prap_op_methods.htm (accessed Dec 2006).
- [10] USEPA (2003). *Framework for Cumulative Risk Assessment* EPA/630/P-02/001F, May Risk Assessment Forum, Washington, DC 20460.
- [11] National Research Council (1983). *Risk Assessment in the Federal Government: Managing the Process*, National Academy Press, Washington, DC.
- [12] Dourson, M.L., Hertzberg, C., Hartung, R. & Blackburn, K. (1985). Novel methods for the estimation of acceptable daily intake, *Toxicology and Industrial Health* **1**, 23–33.
- [13] Crump, K.S. (1984). A new method for determining allowable daily intakes, *Fundamentals of Applied Toxicology* **4**, 854–871.
- [14] Crump, K.S. (1995). Calculation of benchmark doses from continuous data, *Risk Analysis* **15**, 79–89.
- [15] USEPA (1995). *The Use of the Benchmark Dose Approach in Health Risk Assessment*, Risk Assessment Forum, EPA/630/R-94/007.
- [16] Bogdanffy, M.S., Daston, G., Faustman, E.M., Kimmel, C.A., Kimmel, G.L., Feed, J. & Vu, V. (2001). Harmonization of cancer and non-cancer risk assessment, *Toxicological Sciences* **61**, 18–31.
- [17] National Research Council (1991). *Human Exposure Assessment for Airborne Pollutants*, National Academy Press, Washington, DC.
- [18] USEPA (1999). *Exposure Factors Handbook*. EPA/600/C-99/001. National Center for Environmental Assessment, Washington, DC.
- [19] Burmaster, D.E. & Harris, R.E. (1993). The magnitude of compounding conservatism in superfund risk assessments, *Risk Analysis* **13**, 131–134.
- [20] Cullen, A.C. (1994). Measures of compounding conservatism in probabilistic risk assessment, *Risk Analysis* **14**, 389–393.
- [21] USEPA (1989). *Risk Assessment Guidance for Superfund, Vol. 1, Human Health Evaluation Manual (Part A) Interim Final*, United States Environmental Protection Agency, Washington, DC. EPA/540/01-89/002.
- [22] Burmaster, D.E. & Von Stackelberg, K. (1994). Using Monte Carlo simulations in public health risk assessments: estimating and presenting full distributions of risk, *Journal of Exposure Analysis and Environmental Epidemiology* **1**, 491–512.
- [23] Smith, R.L. (1994). Use of Monte Carlo simulation for human exposure assessment at a superfund site, *Risk Analysis* **14**, 433–439.
- [24] Clewell, H.J., Gearhart, J.M., Gentry, P.R., Covington, T.R., Van Landingham, C.B., Crump, K.S. & Shipp, A.M. (1999). Evaluation of the uncertainty in an oral reference dose for methyl mercury due to interindividual variability in pharmacokinetics, *Risk Analysis* **19**, 547–558.
- [25] Stern, A.H. (2005). A revised probabilistic estimate of the maternal methyl mercury intake dose corresponding to a measured cord blood mercury concentration, *Environmental Health Perspectives* **113**, 155–163.
- [26] Dourson, M.L. (1994). Methods for establishing oral reference doses (RfDs), in *Risk Assessment of Essential Elements*, W. Mertz, C.O. Abernathy & S.S. Olin, eds, ILSI Press, Washington, DC, pp. 51–61.
- [27] Gaylor, D.W. & Kodell, R.L. (2000). Percentiles of the product of uncertainty factors for establishing probabilistic reference doses, *Risk Analysis* **20**, 245–250.
- [28] Renwick, A.G. (1993). Dose-derived safety factors for the evaluation of food additives and environmental contaminants, *Food Additives and Contamination* **3**, 275–305.
- [29] Hattis, D., Ginsberg, G., Sonawane, B., Smolenski, S., Russ, A., Kozlak, M. & Goble, R. Differences in pharmacokinetics between children and adults—II. Children's variability in drug elimination half-lives and in some parameters needed for physiologically-based pharmacokinetic modeling, *Risk Analysis* **23**, 117–142.
- [30] IARC (1987). Overall evaluations of carcinogenicity: an updating of IARC monographs 1–42 International Agency for Research on Cancer, *Inhalation Toxicology* **6**(4), 301–325.
- [31] USEPA (1986). *Guidelines for Carcinogen Risk Assessment*, Risk Assessment Forum, Washington, DC. 51 FR 33992.

Further Reading

Hattis, D., Banati, P., Goble, R. & Burmaster, D.E. (1999). Human interindividual variability in parameters related to health risks, *Risk Analysis* **19**, 711–726.

Related Articles

Environmental Monitoring

Environmental Performance Index

Statistics for Environmental Justice

ALAN H. STERN

Environmental Health Triage, Chemical Risk Assessment for

Preventing and protecting against environmental, occupational, and food-related chemical hazards is typically the primary goal of quantitative methods developed to assess environmental chemical risks (*see Environmental Health Risk; Environmental Hazard; Occupational Risks*) [1–5]. Protective regulatory and industrial hygiene measures are not designed to predict chemical risk or casualties. Extensive human pharmacokinetic, epidemiological, and/or clinical-trial data that allow reasonably unbiased estimates of risk for ionizing radiation and medical pharmaceuticals simply do not exist for most toxic industrial chemicals (TICs) or chemical threat agents (*see What are Hazardous Materials?; Environmental Risks*). Consequently, assessments of environmental chemical risks generally entail substantial uncertainty that is typically addressed by routine, health-protective application of “uncertainty” factors designed to guarantee acceptable levels of minimal risk [6]. However, less biased approaches may be preferable in some circumstances. Civil litigations (“toxic tort”) or criminal prosecutions require “more likely than not” or “beyond a reasonable doubt” standards of evidence, respectively, and in such cases unbiased assessments of environmental chemical harm may be required. Application of appropriate quantitative methods would be necessary to help meet such standards. Timely, unbiased assessment of chemical risk can also be important for effective insurance and disaster planning. In general, such unbiased assessment is needed whenever circumstances require choosing among trading off or prioritizing options associated with different chemical risks or different combinations of chemical and other risks [7].

Trade-off contexts concerning health are most dramatically illustrated in the case of urgent military coarse-of-action decisions, or emergency-response situations involving, e.g., airborne chemical threats (*see Counterterrorism*). Airborne chemical threat assessment systems now link real-time plume modeling and mapping capabilities with a variety of

toxic exposure guidelines, but none provide a systematically consistent way to compare risks that involve chemical exposures. If planners and responders cannot reliably compare chemical risks, or compare chemical *versus* nonchemical risks, then they cannot prevent or minimize associated casualties in a risk-based fashion. Reliable comparison of consequences associated with response alternatives will be critical whenever resource constraints or exposure-related complexities force a choice among possible options. Under such conditions, current approaches do not permit risk-based choices, and so will inevitably produce *ad hoc* decisions that tend to allow preventable casualties.

A risk-based approach to trade-offs is possible only if responders use unbiased information on chemical toxicity to predict residual casualties associated with alternative response strategies [7]. Current protective guidelines for acute and chronic chemical exposures [8–12] cannot serve this purpose because they incorporate a number of “uncertainty” factors and conservative assumptions that may vary quite substantially in combined magnitude from chemical to chemical, depending on scope and quality of available toxicity data. Planning and response strategies that effectively prevent exceedance of such conservative guidelines will clearly achieve risk-based consequence management. However, accidental or terrorism-related dispersion of toxic chemicals may involve either of two factors that complicate response and recovery decisions in ways that can make conservative chemical exposure guidelines irrelevant to effective consequence management. First, widespread dispersal of hazardous materials may create exposure scenarios that are potentially *complex*, by involving multiple exposed populations, agents (chemical, perhaps also biological and/or radiological), contaminated media, exposure routes, exposure durations, and/or types of casualty (acute, chronic, quasi-threshold, stochastic, short-term, and/or long-term). Secondly, logistic constraints, clustering of multiple events in time, and the sheer magnitude of damage incurred may limit response and recovery resources, either in the short-term or for extended or indefinite periods. Resource constraints and exposure-scenario complexities will, at times, make it impossible for planners and responders to operate within a traditional *environmental health protection* paradigm, which focuses on preventing the realization of conservatively estimated risks. Instead, decision makers will

be forced to operate within an *environmental health triage* paradigm [13], which focuses on making trade-off (triage) decisions that mitigate the greatest amount of expected casualty.

Protective chemical exposure guidelines appropriately incorporate substantial and widely varying degrees of conservatism, which depend on the nature of available dose–response data. Trade-off and priority-setting decisions require unbiased estimates of expected casualties. Such best estimates are not readily available, in easily accessible form, for most of the wide array of chemicals of concern. Consequently, no readily accessible, reliable basis for risk-based planning or consequence management is available for incidents that require comparing casualties from different chemical agents, or from mixtures of chemicals and other sources of risk (e.g., radiation, biological agents, or evacuation measures). What is missing is a comprehensive database of unbiased estimates of chemical toxicity and associated uncertainty characterizations. To address chemical emergency situations that require trade-off and prioritization decisions, a fundamentally new, general approach to assessing and managing chemical risks is needed [13, 14]. A scientifically peer-reviewed process, actively coordinated by agencies tasked with supporting emergency response, would also clearly contribute to progress toward new data structures and methods to support effective environmental health triage.

The following example of unbiased chemical risk assessment illustrates how and why a triage context should alter the protection-oriented, safety-factor approach traditionally applied to evaluate the acceptability of chemical exposure scenarios. The hypothetical triage situation considered posits that, due to practical short-term limits on available resources, emergency responders must choose between two immediate scenarios: (1) allow a group of trapped office workers to inhale 120 ppm Methyl Isocyanate (MIC) in air for 10 min, or (2) evacuate the workers by means that entail inhalation of 500 ppm Hydrogen Fluoride (HF) in air for 10 min. To protect against severe or lethal toxicity in a general population, acute exposure guideline levels (so-called AEGL-3 values) imply that no 10-min respiratory exposure should exceed an air concentration of 1.2 ppm MIC, or of 170 ppm HF [9]. Scenario 1 involves an exposure that is 10-fold, and scenario 2 an exposure that is threefold, greater than the corresponding AEGL-3

value. However, the AEGL-3 levels for MIC and HF have built-in safety factors of 6 and 1000 respectively, relative to corresponding (LC_{50}) concentrations likely to be lethal to 50% of exposed adults [13]. So scenario 2 implies exposure to about 50%, and scenario 1 to only 1%, of an LC_{50} . If forced to choose, responders should therefore choose scenario 1 rather than 2. More explicit, probabilistic methods are required to address more complicated triage scenarios that involve near- LC_{50} exposures, multiple chemicals, susceptible subpopulations, and/or time-varying concentrations [9, 13].

References

- [1] National Academy of Sciences (NAS) (1983). *Risk Assessment in the Federal Government: Managing the Process*, National Academy Press, Washington, DC.
- [2] Zimmerman, R. (1990). *Governmental Management of Chemical Risk: Regulatory Processes for Environmental Health*, Lewis Publishers, Chelsea.
- [3] American Conference of Governmental Industrial Hygienists (ACGIH) (2001). *Documentation of the Threshold Limit Values and Biological Exposure Indices*, 7th Edition, ACGIH, Cincinnati.
- [4] Dunnette, D.A. (1989). Assessing risks and preventing disease from environmental chemicals, *Journal of Community Health* **14**, 169–186.
- [5] International Programme on Chemical Safety (IPCS) (1999). *Principles for the Assessment of Risks to Human Health from Exposure to Chemicals*, IPCS Report, Environmental Health Criteria 210 (and related Concise International Chemical Assessment Documents and Environmental Health Criteria documents), World Health Organization (WHO), Geneva.
- [6] Dourson, M.L. & Stara, J.F. (1983). Regulatory history and experimental support of uncertainty (safety) factors, *Regulatory Toxicology and Pharmacology* **3**, 224–238.
- [7] National Research Council (NRC) (1994). *Science and Judgment in Risk Assessment*, National Academy Press, Washington, DC, pp. 186–187.
- [8] National Research Council (NRC) (1993). *Guidelines for Developing Community Emergency Exposure Levels for Hazardous Substances*, National Academy Press, Washington, DC.
- [9] National Research Council (NRC) (2000). *Acute Exposure Guideline Levels for Selected Airborne Chemicals*, National Academy Press, Washington, DC, Vol. 1. [Corresponding volumes 2, 3 and 4 published in 2002, 2003 and 2004, respectively].
- [10] National Research Council (NRC) (2001). *Standing Operating Procedures for Developing Acute Exposure Levels for Hazardous Chemicals*, National Academy Press, Washington, DC.

- [11] U.S. Department of the Army, Center for Health Promotion and Preventative Medicine (USACHPPM) (2002). *Technical Guide (TG) 230: Chemical Exposure Guidelines for Deployed Military Personnel*, and corresponding Reference Document (RD 230, January 2002), Aberdeen Proving Ground.
- [12] U.S. Environmental Protection Agency (EPA) (2008). Integrated Risk Information System (IRIS), National Center for Environmental Assessment, Cincinnati at <http://cfpub.epa.gov/ncea/iris/index>.
- [13] Bogen, K.T. (2005). Risk analysis for environmental health triage, *Risk Analysis* **25**, 1085–1095.
- [14] National Research Council (NRC) (2004). *Review of the Army's Technical Guides on Assessing and Managing*

Chemical Hazards to Deployed Personnel, The National Academies Press, Washington, DC.

Related Articles

Air Pollution Risk

Environmental Risk Assessment of Water Pollution

KENNETH T. BOGEN

Environmental Monitoring

Conceptual Framework for Environmental Risk Assessment

Risk assessment has become an increasingly important tool for environmental decision making (*see Environmental Health Risk*). Risk assessments are being performed in application to the spectrum of environmental issues including such concerns as hazardous waste cleanups, permitting activities for water and air discharges, for land and water management decisions related to open land, forests, watersheds and estuaries, and for establishing environmental quality standards and guidelines (*see Air Pollution Risk; Hazardous Waste Site(s); Water Pollution Risk*). Inherently interdisciplinary, the field of risk assessment applied to the environment is evolving rapidly, despite associated uncertainties in assessment methodologies and data limitations. The result is that risk assessment and management procedures are now widely employed, including being dictated by the mandates of regulatory agencies throughout developed countries. However, it is also clear there are substantial needs for the development, use, and sharing of information including environmental data, assessment methods, and insights gained from interacting with decision makers and the public, to improve risk assessment protocols and risk management.

For the public to participate in making risk-based decisions, it is essential that the concepts of risk assessment be understood by the general public (*see Risk and the Media*). Hence, the scientific community must translate risk assessment approaches to simple terms, to ensure understanding is gained by the general public.

To put the concept of environmental risk assessment into more perspective, quantitative risk assessment (QRA) in the environmental arena may include three broad situations: (a) a source of contamination release/site exists, (b) a future activity is to be planned which will release some level of contaminants, or (c) an accident or terrorist activity may involve release of a toxic chemical (*see What are Hazardous Materials?*). To respond to any and/or all of these, routine and specific environmental

monitoring is essential. While specific environmental monitoring for a particular site may be designed to provide direct and accurate information for an existing source, for a future activity or accidental release, background data pertinent to the location are essential. Although models relating emission sources and their impacts to human receptors are employed in all situations, confidence in the model predictions is improved with data obtained by environmental monitoring.

The basic framework of any environmental health risk assessment is (a) exposure/dose assessment, (b) dose-response toxicity relationship (*see Dose-Response Analysis*), and (c) risk characterization. Apparent from the diversity of components, this basic framework is inherently interdisciplinary. The first part of the framework, exposure/dose assessment, is generally addressed by environmental engineers and scientists who specialize in environmental monitoring and modeling. The overall framework of the exposure/dose assessment is illustrated in Figure 1.

Source Characterization

Source characterization begins with hazard identification in terms of a chemical and/or an array of chemicals. This includes contaminated sites, industrial pollution sources (both point and nonpoint) for water, air, and solid wastes, and uses and transportation of toxic chemicals. Source monitoring represents an essential part of comprehensive source characterization with source characterization including information such as the location, quantity of contaminant release, frequency of discharge (e.g., continuous or intermittent), and so on.

Methods for Estimating Chemical Releases

The methods most commonly used in estimating emissions/discharge of pollutants include the following:

- *Continuous emission/discharge monitors (CEMs)*

CEMs continuously measure (with very short averaging time) and record actual emissions/discharge during operations. CEM data may also be used to estimate emissions for different operating and longer averaging times.

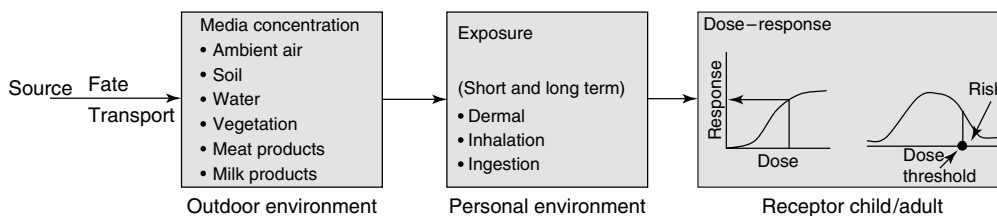


Figure 1 Conceptual framework for an exposure/dose assessment

- *Source testing*

Emission/discharge rates are derived from short-term measurements taken at an emission stack or waste discharge point. Emission data may then be extrapolated to estimate long-term emissions/discharge from the same or similar sources.

- *Material balance*

Emissions/discharges may be determined based on the amount of material that enters a process, the amount that leaves the process, and the amount shipped as part of the product itself.

- *Emission factors*

An emission factor is a ratio that relates the emission of a pollutant to an activity level at a plant that can be easily measured, e.g., an amount of material processed or an amount of fuel used. Given an emission factor and a known activity level, a simple multiplication yields an estimate of the emissions.

- *Fuel analysis*

Emissions may be determined based on the application of conservation laws. The presence of certain elements in fuels may be used to predict their presence in emission streams. For example, sulfur dioxide (SO₂) emissions from oil combustion may be calculated based on the concentration of sulfur in the oil. This approach assumes complete conversion of sulfur to SO₂. Therefore, for every kilogram of sulfur (molecular weight = 32 g) burned, 2 kg of SO₂ (molecular weight = 64 g) are emitted.

- *Emission estimation models*

Emission estimation models are empirically developed process equations used to estimate emissions from specific sources.

- *Surveys and questionnaires*

Surveys and questionnaires are commonly used to obtain facility-specific data on emissions and their sources.

- *Engineering judgment*

Engineering judgment is employed when the specific emission estimation techniques, such as stack testing, material balance, or emission factor are not possible; instead, the estimation is made by an engineer familiar with the specific process.

Examples of Emission Factor Information

- *AP-42*

One of the most frequently cited resources for emission factors to the air is the USEPA document, *Compilation of Air Pollutant Emission Factors (AP-42)*. This document contains pollutant emission factors for point and area sources. AP-42 is available at <http://www.epa.gov/ttn/chief/ap42/index.html>.

- *Emission factor databases*

Several emission factor databases are currently available in easy-to-access formats. These tools include Factor Information Retrieval (FIRE) Data System (<http://www.epa.gov/ttn/chief/fire.html>) and Air Clearinghouse for Inventories and Emission Factors – CD-ROM, <http://www.epa.gov/ttn/chief/airchief.html#order>.

- In addition, one should conduct a search of technical papers for source test and background information for the emission source category or pollutants in question. One can conduct this search using USEPA library services or through government document depositories at local universities. Examples of references and documents that one should review include miscellaneous private sector resources. For example, the National Council of the Paper Industry for Air and Stream Improvements, Inc. (NCASI) compiles, through a highly focused research program, reliable environmental data and information on the forest products industry.

Source characterization for existing sources is often accomplished by emission/discharge monitoring that includes source sampling and laboratory chemical analyses to ascertain quantity of emissions and discharges. For facilities still in the planning stage, best estimates of emission estimates must be obtained through other sources, which are expected to be similar in operation.

The Toxics Release Inventory (TRI) (<http://www.epa.gov/tri/>) database [1] contains information on specific toxic chemical releases for more than 650 toxic chemicals that are being used, manufactured, treated, transported, or released into the environment.

Broad classifications of sources are listed below.

- *Soils*
 - application of agricultural chemicals;
 - application of sewage sludge and treated/untreated wastewater for irrigation;
 - residues from transportation sources;
 - waste disposal.
- *Surface water*
 - runoff from agricultural and urban land;
 - treated, untreated, and accidental discharges from cities and factories;
 - atmospheric deposition of pollutants to lakes.
- *Groundwater*
 - leaking underground tanks and pipelines;
 - wastewater injection wells;
 - leachate from landfills.
- *Air*
 - stacks and vents;
 - leakages from storage tanks, flanges, valves;
 - handling of raw materials, products and their evaporation/volatilization.

Fate and Transport of Chemicals in the Environment

Once the source characteristics have been established, fate and transport of contaminants through monitoring and/or by mathematical modeling is essential to estimate the exposure and/or dose experienced by the receptor, as the next phase in the risk assessment (see **Fate and Transport Models**). This requires gathering data to quantify the constituents in each

of the compartments (air, water, soil, etc.) plus understanding the transformations which may occur from one compartment to another (e.g., volatilization of a chemical from water to air). Because constituents can change media, for many situations it is necessary to employ multimedia modeling to complete the hazard analysis. Before discussing these models which account for physical, chemical, and biological transformations, the properties of chemicals that influence the fate and transport must be addressed. The considerations must include the various attributes of the chemicals, e.g., volatility, viscosity, and density, since these features will influence the pathways by which the chemical may migrate; e.g., a highly volatile chemical may volatilize and hence travel as a gas and be inhaled by a receptor. Alternatively, a dense solvent such as tetrachloroethylene (PCE) may sink as a separate phase to the saturated groundwater and create an exposure pathway to a receptor *via* drinking water.

Important Fate and Transport Properties, Processes, and Parameters of Environmental Contaminants

Several factors contribute to the migration and transport of chemicals in environmental media. For example, in groundwater, does a chemical released into the groundwater cause an exposure hazard *via* consumption of groundwater as a water supply? Alternatively, the volatility of the chemical may be sufficiently high that the chemical volatilizes and creates an exposure scenario by the vapor entering into the basement of a house and hence an inhalation exposure to the resident.

Examination of a contaminant's physical and chemical properties may determine the extent of environmental partitioning, migration, and/or attenuation. A number of important physical and chemical properties, processes, and the parameters may affect the environmental fate and/or cross-media transfers of environmental contaminants including the physical state, water solubility, diffusion, dispersion and volatilization, biodegradation and photolysis (or photodegradation), hydrolysis, and oxidation/reduction (redox reactions). More details on these processes are provided, for example, in reference [2].

One of the important influences on risk exposure scenarios, and used herein to demonstrate the

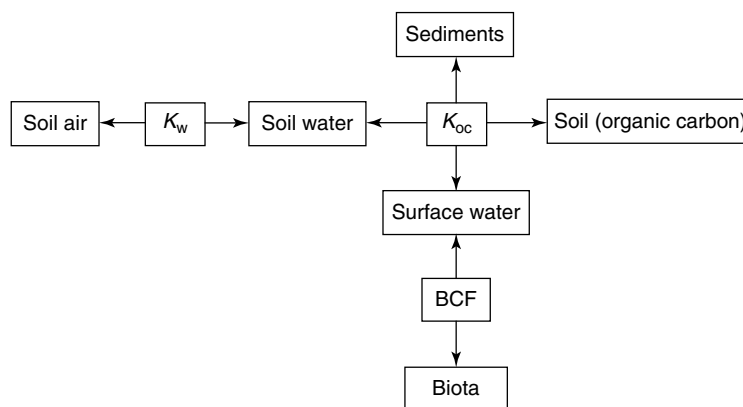


Figure 2 Major components of the partitioning of contaminants between environmental compartments [Reproduced from [2]. © John Wiley & Sons, Ltd, 1993.]

complexity of risk assessment procedures, is the partitioning of a chemical between several phases within a variety of environmental matrices. The partition coefficient is a measure of the distribution of a given compound between two phases and is expressed as a concentration ratio. Several measures of the partitioning phenomena are enumerated below and their application in environmental phase transformation is shown in Figure 2.

- *Water/air partition coefficient*

The water/air partition coefficient (K_w) relates the distribution of a chemical between water and air. It consists of an expression that is equivalent to the reciprocal of Henry's law (H), i.e.,

$$K_w = \frac{C_{\text{water}}}{C_{\text{air}}} = \frac{1}{H} \quad (1)$$

Where C_{air} is the concentration of the chemical in air (expressed in units of $\mu\text{g l}^{-1}$). This coefficient is referred to as the *washout ratio* for the removal of gaseous air pollutants from air to the water phase during rain or snowfall.

- *Octanol/water partition coefficient*

The octanol/water partition (K_{ow}) is defined as the ratio of a chemical's concentration in the octanol phase (organic) to its concentration in the aqueous phase of a two-phase octanol/water system, represented by:

$$K_{ow} = \frac{\text{Concentration in octanol phase}}{\text{Concentration in aqueous phase}} \quad (2)$$

This dimensionless parameter provides a measure of the extent of chemical partitioning between water and octanol at equilibrium. It has become a particularly important parameter in the study of environmental fate of organic chemicals.

K_{ow} can be used to predict the magnitude of an organic constituent's tendency to partition between the aqueous and organic phases of a two-phase system, such as surface water and aquatic organisms. The higher the value of K_{ow} , the greater the tendency of an organic constituent to adsorb to soil or waste matrices containing appreciable organic carbon or to accumulate in biota.

- *Organic carbon adsorption coefficient*

The sorption characteristics of a chemical may be normalized to obtain a sorption constant based on organic carbon, which is essentially independent of any soil material. The organic carbon adsorption coefficient (K_{oc}) provides a measure of the extent of partitioning of a chemical constituent between soil or sediment organic carbon and water at equilibrium. Also called the *organic carbon partition coefficients*, K_{oc} , is a measure of the tendency for organics to be adsorbed by soil and sediment, and is expressed by

$$K_{oc} (\text{ml g}^{-1}) = \frac{\text{Microgram of chemical adsorbed per gram weight of soil or sediment organic carbon}}{\text{Microgram of chemical dissolved per milliliter of water}} \quad (3)$$

- *Bioconcentration factor (BCF)*

The BCF is the ratio of the concentration of a chemical constituent in an organism or whole body (e.g., a fish) or specific tissue (e.g., fat) to the concentration in its surrounding medium (e.g., water) at equilibrium, given by

$$BCF = \frac{(\text{Concentration in biota})}{(\text{Concentration in surrounding medium})} = \frac{\left[\left(\frac{\text{Microgram of chemical per gram biota such as fish}}{\text{Microgram of chemical per milliliter medium such as water}} \right) \right]}{\left[\left(\frac{\text{Microgram of chemical per gram biota such as fish}}{\text{Microgram of chemical per milliliter medium such as water}} \right) \right]} \quad (4)$$

The BCF indicates the degree to which a chemical residue may accumulate in aquatic organisms, coincident with ambient concentration of the chemical in water; it is a measure of the tendency of a chemical in water to accumulate in the tissue of an organism. In this regard, this is related to the concentration of the chemical where the fish are feeding, or biota BCF for that chemical. Thus, the average concentration in fish or biota is given by

$$C_{\text{fish-biota}} (\mu\text{g kg}^{-1}) = C_{\text{water}} (\mu\text{g l}^{-1}) \times BCF \quad (5)$$

The BCF is important in determining the risk, for example, *via* food-chain modeling where necessary assumptions involve tracing of successive levels of uptake through the food chain. For understanding this type of modeling, data for one of the environmental media are necessary and hence samples must be collected, as was done in reference [3] and then transformed to other parts of the food chain by procedures outlined in the following paragraphs.

Subsequently, food-chain modeling as input to the environmental risk assessment requires quantification of the BCF, which defines the equilibrium ratio of the concentration of the substance in the exposed organism (including vegetation and plants) to the concentration of the substance in the surrounding environment. In simple mathematical terms, soil-to-plant BCFs are defined as (after [4])

$$BCF_{\text{veg}} = \frac{C_{\text{veg}}}{C_{\text{soil}}} \quad (6)$$

where, BCF_{veg} is the BCF for vegetation, C_{veg} is the concentration of the chemical in vegetation ($\mu\text{g kg}^{-1}$ (dry weight)), and C_{soil} is the concentration of the chemical in soil ($\mu\text{g kg}^{-1}$ (dry weight)). The value of BCF for a toxicant depends on its chemical

properties, environmental conditions, and biological factors such as the organism's ability to metabolize the substance.

BCF magnitudes are related to *n*-octanol/water partition coefficient, K_{ow} (e.g., see [5]) and are constituent-specific. Uncertainties in estimation of BCF will be directly transformed into uncertain estimates of risk magnitudes. While wishing to remain conservative in the quantification of risk to err on the side of protection of humans and the environment, an underestimated BCF may eclipse safety and may result in a false sense of protection.

Standardized protocols for estimating BCFs exist for a number of aquatic species including, for example, the USEPA's recently updated fish and oyster BCF test guidelines [6, 7]. However, experimental determination of BCF is expensive and demanding and, as a consequence, measuring the BCFs of the many thousands of chemical substances that are of potential regulatory interest is simply not possible. This is obviously true for new chemicals such as those reviewed under the US Toxic Substances Control Act [8]. The dimensions of uncertainty also include the BCF from soil to vegetation to examine intake of toxic pollutants directly from the plants. Several options are commonly relied upon for estimation of the BCF. As per reference [9], one method for dioxins and furans is to rely upon bioaccumulation equivalency factors (BEFs) as a measure of congeners bioaccumulation potential relative to 2,3,7,8-(tetra chloro dibenzo dioxin (2,3,7,8-TCDD); see **Health Hazards Posed by Dioxin**). BEFs are estimated as a ratio between each dioxin and furan congener-specific biota-sediment accumulation factor to that of 2,3,7,8-TCDD [10, 11] as

$$BEF_i = \frac{BSAF_i}{BSAF_{\text{TCDD}}} \quad (7)$$

where BEF_i is the bioaccumulation equivalency factor for the i th congener, $BSAF_i$ is the biota-sediment accumulation factor for i th congener, and $BSAF_{\text{TCDD}}$ is the biota-sediment accumulation factor for 2,3,7,8-TCDD.

The estimates of BCF_i for each of the i congeners is then determined from

$$BCF_i = BCF_{\text{TCDD}} \cdot BEF_i \quad (8)$$

where BCF_i is the soil-to-plant bioconcentration factor for the i th congener (kg^{-1}), BCF_{TCDD} is

the soil-to-plant bioconcentration factor for 2,3,7,8-TCDD (1 kg^{-1}).

An alternative procedure, based on $\log K_{ow}$, provides estimates of BCF_i , from regression equations of the general form

$$\log BCF = a \log K_{ow} + b \quad (9)$$

where a and b are empirically determined constants, and K_{ow} is the n -octanol/water partition coefficient. Numerous versions of equation (9) have been published, but the broadly applicable (i.e., not class-specific) equations generally have slopes close to unity and negative intercepts [8]. The USEPA [9] currently uses the equation of reference [12] for estimating soil-to-plant BCF in which $a = -0.578$ and $b = 1.588$.

Such equations (e.g., see [13]) provide approximate magnitudes of the BCF for nonionic, nonmetabolized substances with $\log K_{ow}$ in the range of 1–6 for aquatic animals like fish and has been used extensively (as cited in Meylan *et al.* [8]). Alternatively, for the BCF in plants from soil, Travis and Arms [12] is extensively used. The main limitation of the Travis and Arms equation is that they have considered only 29 chemicals having $\log K_{ow}$ up to 6.89 (and polybrominated biphenyl, $\log K_{ow} = 9.35$). However, Travis and Arms have considered only one dioxin, TCDD (out of 29 chemicals) and, therefore, estimating BCF for other dioxins and furans using their equation may add to uncertainty in risk assessments. However, recently, Sharma *et al.* have demonstrated that observed BCFs for dioxins and furans from the field study are generally much higher (3–575 times – for 15 compounds and for octa chlorinated dibenzofurans (octa-CDF) 2770 times) than those computed by the USEPA method and are higher by a factor of 3–650 (for eight compounds) computed by the K_{ow} model [3]. These findings indicate that use of model-based BCFs may result in underestimation

of the risk or hazard and, pertinent to the completion of environmental risk assessments, demonstrates the need for careful development of data and, to the extent feasible, the necessity for collection of environmental data.

Contaminant Transport, Fate, and Modeling

The modeling of contaminants describes the pathways, phase transformation, and chemical transformation. Most of the models are based on simple concepts of mass conservation. Figure 3, in a simplistic way, presents the environmental models.

A specific example model for transport of soil to groundwater is depicted in Figure 4.

Another example of transport and transformation of air contaminant is illustrated in Figure 5.

As apparent from Figure 5, the exposure concentrations of a contaminant depend on several factors including:

- dispersion by atmospheric advection and diffusion;
- reduced by deposition, transfer to other media, and transformation reactions; and,
- meteorological parameters
 - wind parameters (direction, speed, and turbulence)
 - thermal properties (stability).

A typical multimedia equilibrium distribution of a fixed quantity of a chemical, in a closed environment at equilibrium, with no degrading reactions, no advective processes, and no intermedia transport processes (e.g., no wet deposition or sedimentation) is shown in Figure 6.

A multimedia environmental model like the one shown in Figure 6 uses physical–chemical properties of chemical. Three types of chemicals are treated

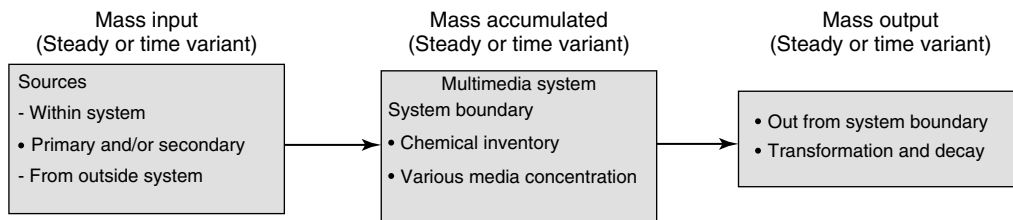


Figure 3 Conceptual environmental model: equations are solved to equate mass gained and mass lost

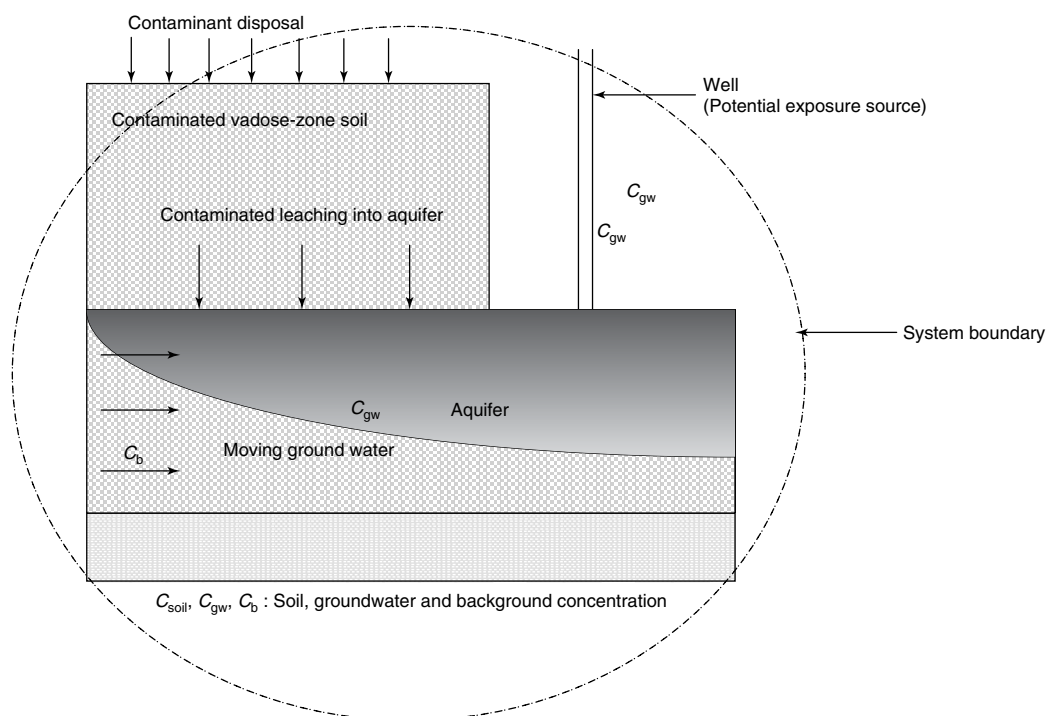
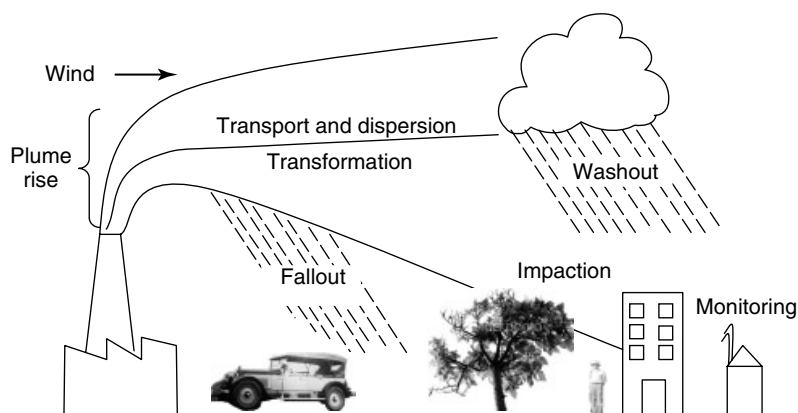


Figure 4 Transport from soil to groundwater



Source	Transport	Receptor
Emissions time and duration	Wind speed and direction turbulence	Humans plants buildings
Characteristics height volume flow temperature	Physical and chemical transformations	Monitoring

Figure 5 Transport and transformation of air contaminants (after Boubel *et al.* 1994)

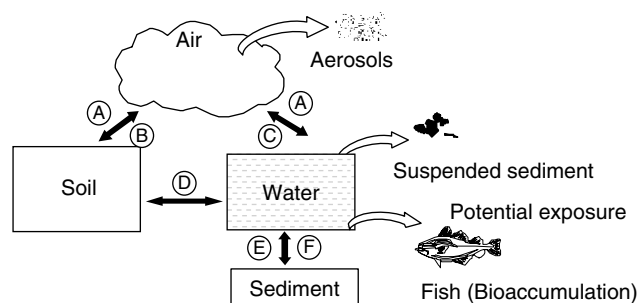


Figure 6 Equilibrium multimedia model description. A: deposition/rain washout; B: resuspension/evaporation; C: evaporation/super saturation; D: ground water movement; E: sedimentation; F: resuspension

The success of a modeling assessment may depend on how well the model fits with the measurements. Thus, environmental monitoring that incorporates sampling, analysis, interpretation of data, and quality assurance and quality control is essential for the success of a modeling exercise. We may need, for example, to utilize monitoring data to provide quantification of coefficients in the mathematical models, to allow accurate characterization of risks. The availability of monitoring data for purposes of calibrating mathematical models greatly improves the veracity of the monitoring results and *vice versa*.

While the sampling of water, soil, and food products present challenges, environmental sampling may be much more challenging. For example, to ensure capture of measurable quantities of contaminants (e.g., for air monitoring, approximately 200–1000 m³ must pass through the sampling device and chemicals of interest) need to be retained on the sampling device. Further, there may be complexities due to, for example, diurnal and seasonal variation. For example, significant levels of the ozone may be just 3–4 h in a day while for the remainder of the day, the ozone may be at background concentrations. This also implies that the averaging time of the measurements becomes crucial. For example, a short-term peak concentration may be more harmful/toxic than long-term values (chronic) but the relevance depends on the source chemical, the migration pathways, and the scenarios of exposure of the receptor. For some compounds, especially those having long-term chronic carcinogenic effects, annual averages may be necessary (e.g., for benzene, a known carcinogen).

Receptor Characteristics

The attributes of the receptor influence the potential for the receptor to become ill. For example, what is the exposure timeframe of the receptor? The size of the receptor (child *versus* adult) is important since the potential to cause cancer is a function of the size of the individual (milligram per kilogram of weight).

Statistical Analyses of Monitoring Data in Support of Risk Assessment

As apparent from the preceding, environmental monitoring data are important inputs to environmental risk assessments (*see Environmental Health Risk*), to allow assignment of coefficients within the models. Unfortunately, environmental monitoring data tend to have characteristics that make the application of standard statistical methods difficult and/or inappropriate [14].

Examples of issues with interpreting the data include the data record length may be brief and may have skewness, thereby violating the assumptions implicit in standard statistical tests. The requirements of many of the more frequently used statistical tests must be validated to ensure the application of the tests is appropriate (e.g., see [15] for guidance).

Another problematic characteristic of environmental monitoring data is that the environmental data are often “censored” (*see Censoring*). That is, the data have a lower limit or, less frequently, an upper limit, beyond which a monitoring result is reported as “not detected” or “off-scale”. Statistically, this causes a change in the shape of the data distribution that may violate assumptions of specific statistical models and

require the use of alternate methods. Finally, natural systems tend to be highly complex and variable, and confounding effects, such as seasonality, temperature, or matrix stratification may need to be extracted, to remove noise from the data.

These features stress the difficulty with statistical interpretation of environmental data. There is a need for the analysis procedure to be sensitive to small changes (e.g., early detection of contamination is desirable) and yet a point of diminishing marginal returns for data collection may also occur. The result is that the analyst must be inventive in terms of how statistical analyses proceed. For the variety of reasons previously indicated, there is not a single approach to statistical analyses as input to environmental risk assessment. Instead, what is frequently needed is a series of approaches, each of which possesses useful attributes that may be appropriate in addressing a particular question.

Samples of environmental quality are just that, namely subsets of the populations that are of interest. To clarify, a limited number of monitoring samples will never fully describe every condition found in the population, i.e., every possible subset of the environment. The challenge then is to acquire a representative sample for use as part of the environmental risk assessment, to estimate population parameters with a given probability. The result is that in many cases, estimates of environmental quality must be developed from only very brief data records, brief being with respect to both temporal and spatial considerations. Even in cases with substantial quantities of data, there may only be a limited amount of information that is usable to address a specific question.

In addition to natural variability, the acquisition of monitoring data is also subject to collection and analysis errors. Errors in sampling procedures, inadequate sample storage and preservation techniques, and laboratory analytical errors are examples of errors in environmental data sets.

There is always a degree of uncertainty associated with each discrete measurement of environmental quality. In interpreting data, each discrete measurement represents a range of statistically probable values instead of a single value. Uncertainty can be estimated by considering replication and repeatability. Replication is a measure of the variation in results obtained by the same operator in a given laboratory using the same apparatus to make successive determinations on identical test material within a short period

of time. Often this is done during quality assurance and quality control testing of laboratories, to ensure the results are trustworthy. Replication represents the analytical uncertainty in acquiring monitoring data. Conversely, repeatability is a quantitative expression of the random error associated in the long run with a single operator in a given laboratory obtaining successive results with the same apparatus under constant operating conditions on identical material. Repeatability is a measure of the natural or sample variability. Note that a systematic error or bias in analytical results may be introduced in the field or laboratory due to sampling, handling, and/or analysis techniques, and such error may only be detected by making interlaboratory or intertechnician comparisons using the same samples.

The fact that sampling errors are inherent in random data does not mean, however, that statistical manipulations and sophistication can in any way overcome faulty environmental data. The quality of any statistical analysis as input to a risk assessment, is no better than the quality of the data utilized. Furthermore, statistical considerations cannot be used to replace judgment and careful thought in analyzing data. Statistics must be regarded as a tool or an aid to understanding, but never as a replacement of careful thought [15].

Statistical methods also suggest the ways to ascertain and handle the uncertainties in the monitoring data. In practice, with the help of a few measurements (e.g., number of samples <20) probability distribution functions can be generated for each parameter and from the derived distributions, random numbers can be generated as the surrogate of actual observations. This implies that infinite number of realizations can be generated and utilized in a model to develop probabilistic exposure scenarios, a procedure referred to as *Monte Carlo simulation*. For example, such an example for lead distribution in various food items is shown in Figure 7.

Conclusions

Use of environmental risk assessment and management is becoming increasingly important in decision making for environmental problems. There is substantial information available in the technical literature upon which to rely, and there is important merit in considering the collection of additional, site-specific data to ensure that the risk assessments are

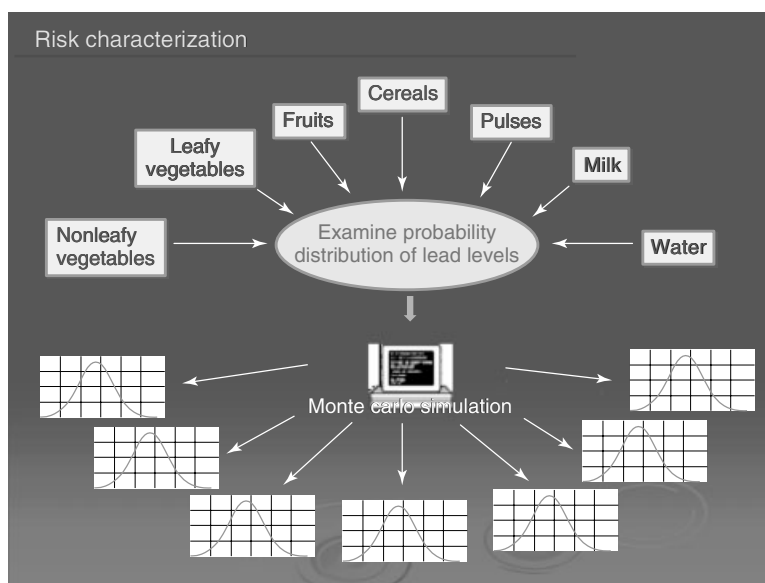


Figure 7 Monte Carlo simulation to address uncertainties of measurements

pertinent, and appropriate, to a particular application. Frequently, there are so many environmental variables (chemical-specific, as well as media-specific) that the value of relying on site-specific data is very real, as long as it is feasible to do so. Clearly, there is an essential linkage between environmental risk assessment, and the data upon which the environmental risk assessment relies. While there is merit in obtaining additional data, there are points at which diminishing marginal returns occur for additional data collection efforts.

References

- [1] USEPA (1995b). *Ecological Effects Test Guidelines*, Aquatic Toxicity Information Retrieval AQUIRE Database, U.S.EPA, Environmental Research Laboratory, Duluth.
- [2] Asante-Duah, K. (1993). *Risk Assessment in Environmental Management*, John Wiley & Sons, New York.
- [3] Sharma, M., McBean, E.A. & Glowing, A. (2007). Bioconcentration of dioxins and furans in vegetation, *Journal of Air, Soil and Water Pollution* **179**(1–4), 117–124.
- [4] USEPA (1994). *Estimating Exposure to Dioxin-Like Compounds, Volume II: Properties, Sources, Occurrence and Background Exposures*, EPA/600/6-88/005Cb, U.S.EPA, Office of Health and Environmental Assessment, Exposure Assessment Group.
- [5] Lyman, W.J., Reehl, W.F. & Rosenblatt, D.H. (1990). *Handbook of Chemical Property Estimation Methods*, American Chemical Society, Washington, DC.
- [6] USEPA (1996a). *Ecological Effects Test Guidelines*, EPA/712/C-96/129, OPPTS 850.1730 Fish BCF, Public Draft, Office of Prevention, Pesticides and Toxic Substances, Washington, DC.
- [7] USEPA (1996b). *Ecological Effects Test Guidelines*, EPA/712/C-96-127, OPPTS 850.1710, Oyster BCF, Draft Report, Office of Prevention, Pesticides and Toxic Substances, Washington, DC.
- [8] Meylan, W.M., Howard, P.H., Boethling, R.S., Aronson, D., Printup, H. & Gouchie, C. (1999). Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient, *Environmental Toxicology and Chemistry* **18**(4), 664–672.
- [9] USEPA (1999). *Screening Level Ecological Risk Assessment Protocol for Hazardous Waste Combustion Facilities*, Report No. EPA 530-D-99-001A, Office of Solid Waste, Vol. 1.
- [10] Boubel, R.W., Fox, D.L., Turner, D.B. & Stern A.C. (1994). *Fundamentals of Air Pollution*, 3rd edition, Academic Press, New York.
- [11] USEPA (1995a). *Great Lakes Water Quality Initiative Technical Support Document for Wildlife Criteria*, EPA-820-B-95-009, Office of Water, Washington, DC.
- [12] Travis, C.C. & Arms, A.D. (1988). Bioconcentration of organics in beef, milk, and vegetation, *Environmental Science and Technology* **22**(3), 271–274.
- [13] Veith, G.D. & Kosian, P. (1983). Estimating bioconcentration potential from octanol/water partition coefficients, in *Physical Behavior of PCBs in the Great*

Lakes, D. Mackay, S. Paterson, & S.J. Eisenreich, eds, Ann Arbor Science, Ann Arbor, pp. 269–282.

- [14] McBean, E., Schmidtke, K., Dyck, W. & Rovers, F. (2001). Monitoring of contaminants and consideration of risk, in *Geotechnical and Geoenvironmental Engineering Handbook*, R.K. Rowe, ed, McGraw-Hill, Chapter 28.
- [15] McBean, E.A. & Rovers, F.A. (1998). *Statistical Procedures for Analysis of Environmental Monitoring Data and Risk Assessment*, Prentice-Hall, Englewood Cliffs.

Related Articles

Environmental Hazard

Evaluation of Risk Communication Efforts

Statistics for Environmental Justice

EDWARD A. McBEAN AND MUKESH SHARMA

Environmental Performance Index

Contrary to other policy fields such as the economy, health care, and education, the environment remains mired in deeply emotional debates and rhetoric that makes limited use of data and performance assessments (*see* **History and Examples of Environmental Justice**). While partially the result of a persistent lack of relevant information, environmental policy requires the same data-driven, rigorous assessments used to make decisions on economic and other issues. Moreover, effective environmental management is not possible without monitoring ecological conditions and measuring the effects of policy responses to environmental pressures (*see* **Ecological Risk Assessment**).

Recognizing this need and the prevailing gap of quantitative measures of environmental performance, a team of researchers and policy experts at the Yale Center for Environmental Law and Policy (Yale University) and Columbia University's Center for International Earth Science Information Network (CIESIN) with support from the World Economic Forum developed the Pilot 2006 Environmental Performance Index (EPI) [1].

The EPI measures the performance of country-level environmental activities on key policy concerns. It is designed to support and complement the environmental indicators of the United Nations Millennium Development Goals (MDG), which are the international community's most concerted effort, to date, to eradicate poverty and promote human development using goal-specific indicators with explicit time frames.

The EPI Framework

The EPI is a target-bound, composite index measuring environmental policy outcomes on two broad, inextricably linked objectives: (a) protecting human environmental health and (b) maintaining ecological vitality. National performance on each objective is measured on six widely adopted policy categories through a total of 16 indicators shown in Figure 1.

Although input and process indicators are required for choosing policy solutions within the context of

benefit–cost analysis (*see* **Risk–Benefit Analysis for Environmental Applications**), the ultimate goal is to achieve stated objectives for on-the-ground conditions. Therefore, the EPI indicators are outcome oriented to the extent data availability and understanding of causal relationships permit.

As shown in Table 1 the performance targets of the 16 indicators reflect relevant long-term public health or ecosystem sustainability objectives drawn from international agreements, international or national standards, and prevailing consensus among scientists.

Although orthogonal indicators are often desirable in index development, environmental pressures frequently have multiple impacts such that the separation of issues into mutually exclusive policy categories is not feasible in all cases. For this reason the indicators *particulate concentration*, *water consumption*, and *timber harvest rate* are each associated with two policy categories (*cf.* Figure 1).

Data Transformation

Aggregation of the 16 EPI indicators into performance indices for the six policy categories, and ultimately the EPI, requires transformation to the same scale. This is achieved by converting each indicator to a standardized proximity-to-target measure such that a value of 100 corresponds to meeting (or exceeding) the target and a value of zero to being farthest away from the target among the 133 countries included in the EPI. Prior to this conversion, extremely low values, which may represent data anomalies (*see* **Low-Dose Extrapolation**) rather than true measurements, are winsorized at both the fifth percentile and target values. This means that the lowest performing 5% of the observations of each indicator are set to the fifth percentile value and observations exceeding the corresponding target are reduced to the target value. The calculation of proximity-to-target values $\tilde{y}_{ij}$ is shown in equation (1), where y_{ij}^w denotes the winsorized value for the i th country ($i = 1, 2, \dots, 133$) and j th indicator ($j = 1, 2, \dots, 16$), and t_j is the target of the j th indicator.

$$\tilde{y}_{ij} = \begin{cases} 100(t_j - y_{ij}^w)/(t_j - \min(y_{ij}^w)), & t_j \geq \max(y_{ij}^w) \\ 100(y_{ij}^w - t_j)/(\max(y_{ij}^w) - t_j), & t_j \leq \min(y_{ij}^w) \end{cases} \quad (1)$$

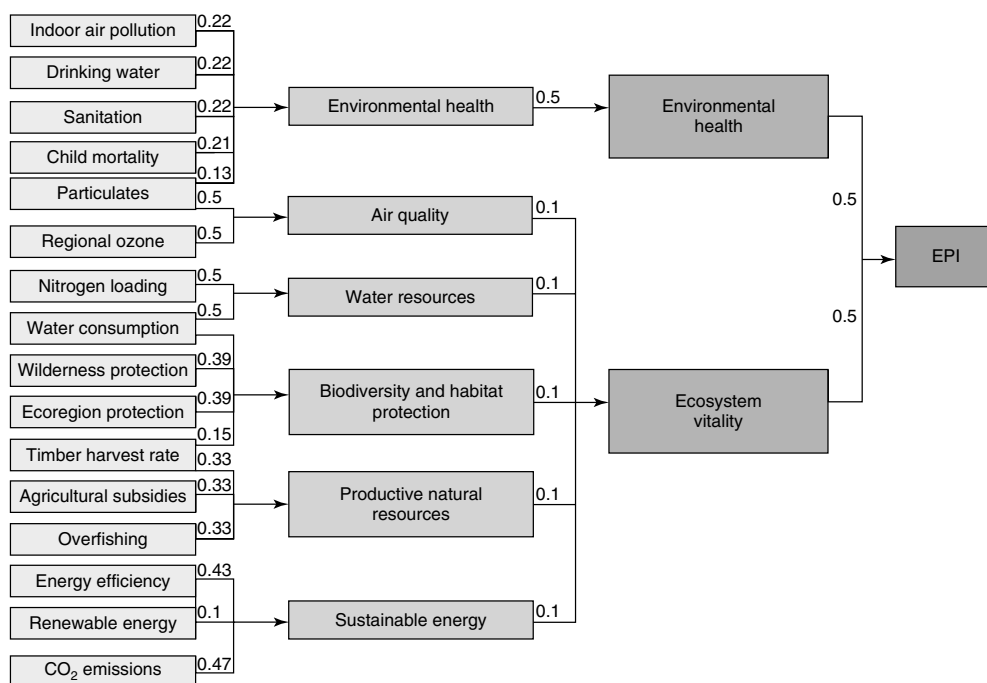


Figure 1 The EPI is constructed in a two-step procedure by averaging 16 proximity-to-target indicators using weights derived from principal component analysis and expert judgment. Numbers indicate the weight of the component within the EPI

Weighting

The six policy subindices are calculated as the weighted average of their underlying proximity-to-target indicators. Since weights can be interpreted as the importance given to the respective indicator, which is likely to vary among countries, no uniformly “correct” weights exist. Thus, a combination of principal component analysis (PCA) and expert judgment is used to specify the indicator weights. The PCA identifies three distinct policy dimensions – environmental health, biodiversity and habitat protection, and sustainable energy – and the estimated indicator loads are used to calculate the weights (*cf.* Table 1). In the absence of sufficient justification for differential weights for the remaining three categories, equal weighting is applied. The EPI then gives 50% weight to the environmental health policy category and 10% weight each to the remaining five policy categories, thereby reflecting the political dominance of public environmental health concerns (*cf.* Table 1).

Results

Of the 133 countries included in the EPI none has achieved all 16 performance targets and likewise no country scored zero on all measures, although four countries scored less than 100 on every single indicator: Jordan, South Africa, Thailand, and Tunisia. The range of observed scores for the EPI and policy categories is substantial, as Table 2 illustrates for a sample of 25 countries.

Policy makers, interested in identifying what makes some countries successful environmental stewards while others struggle to meet their responsibilities, benefit from EPI “peer group” analyses on the basis of common geographical, political, environmental, and other factors (see Table 2). For example, per capita gross domestic product (GDP) measured in purchasing power parities (PPP), to account for differences in cost of living, is a useful discriminator to identify countries that are likely to perform differently on the indicators *child mortality* and *nitrogen loading*.

Table 1 Indicators and performance targets in the Pilot 2006 Environmental Performance Index^{(a)(b)}

Policy category	Indicator	Performance target
Environmental health	Concentration of particulates (urban)	10 $\mu\text{g m}^{-3(c)}$
	Indoor air pollution	0% households using solid fuels ^(d)
	Access to improved drinking water	100% access
	Access to sanitation	100% access
	Child mortality (1–4 years)	0 deaths per 1000 pop 1–4 years
Air quality	Concentration of particulates (urban)	10 $\mu\text{g m}^{-3(c)}$
	Regional ozone concentration	15 ppb ^(e)
Water resources	Nitrogen loading	1 mg l^{-1}
	Water consumption	0% oversubscription
	Wilderness protection	90% of wild areas protected
Biodiversity and habitat protection	Eco-region protection	10% for all biomes
	Timber harvest rate	3% ^(f)
Productive natural resources	Water consumption	0% oversubscription
	Timber harvest rate	3% ^(f)
	Overfishing	No overfishing (score of 0)
	Agricultural subsidies	0%
Sustainable energy	Energy efficiency	1650 TJ/million US\$ GDP
	Renewable energy	100%
	CO ₂ emissions	0 net emissions ^(g)

^(a) Reproduced with permission from [1]. © Yale Center for Environmental Law & Policy, 2006

^(b) The indicators concentration of particulates (urban), water consumption, and timber harvest rate are each constituent to two policy objectives. They and their corresponding performance target are therefore listed twice. The total number of indicators in the EPI is 16

^(c) Determined in consultation with Kiran Pandey from the World Bank and other air pollution experts

^(d) Determined in consultation with Kirk Smith and Daniel Kammen at UC Berkeley and the indoor air pollution literature

^(e) Determined in consultation with Denise Mauzerall and her air pollution team at Princeton University

^(f) Determined in consultation with Lloyd Irland and Chad Oliver from the Yale School of Forestry and Environmental Studies

^(g) Strict interpretation of the goal of the 1992 UN Framework Convention on Climate Change

Regardless of the choice of peer group, environmental policy is a heterogeneous mixture of successes and shortfalls for most countries. For example, although the average country is 82% along to meeting its water resources targets, it has covered only 51 and 54% of the distance to the Biodiversity and Habitat Protection and Air Quality goals, respectively. Similarly, timber harvest rate, which is measured

only for countries with existing natural or plantation forests, is near the sustainability target in most countries but almost every country fails to reach the wilderness and eco-region protection targets.

In addition to the EPI's environmental outcome measures, analyzing its relationship to other environmental, economic, and development indices also offers insights for policy. Without implying causal

Table 2 Sample of 25 countries ranked by EPI score. Shown are also the scores for the six policy objectives: environmental health, air quality, water resources, biodiversity and habitat protection, productive natural resources, and sustainable energy^{(a)(b)}

Rank	Country	EPI score	Peer group	Environ- mental health	Air quality	Water resources	Biodiversity and habitat protection	Productive natural resources	Sustainable energy
1	New Zealand	88.0	OECD	97.9	83.7	98.8	73.4	61.4	73.4
3	Finland	87.0	OECD, EU	98.8	65.3	99.5	54.3	81.5	75.7
5	United Kingdom	85.6	OECD, EU	98.9	61.6	91.9	58.8	71.6	77.8
8	Canada	84.0	OECD, FTAA	98.6	56.2	98.4	55.1	73.9	62.8
14	Japan	81.9	OECD, POP, ASEAN	97.6	52.6	94.8	70.4	33.3	79.7
15	Costa Rica	81.6	FTAA	81.1	60.6	100	80.2	83.1	86.0
26	Chile	78.9	FTAA	87.2	63.7	83.7	68.3	63.0	74.6
28	United States	78.5	OECD, FTAA	98.3	44.7	73.9	66.8	38.9	69.7
33	Hungary	77.0	OECD, EU	94.2	55.6	77.0	47.6	50.0	69.2
34	Brazil	77.0	FTAA	79.3	64.0	97.8	50.4	80.9	80.6
36	Lebanon	76.7	POP	93.4	52.1	89.3	20.2	76.6	61.2
46	Gabon	73.2	AU	61.0	96.1	99.9	62.5	88.9	79.8
51	Ukraine	71.2	NIS	93.8	56.6	65.2	40.0	77.8	3.7
53	Iran	70.0	D	85.7	31.1	72.4	47.9	83.3	36.6
64	Jordan	66.0	D	85.5	40.6	45.8	55.9	38.0	51.7
72	Ghana	63.1	AU	48.8	87.3	99.4	50.1	67.5	83.3
78	Uganda	60.8	LDC, AU	31.7	98.0	92.7	73.6	93.0	92.4
80	Kyrgyzstan	60.5	NIS	53.7	50.6	92.7	68.0	100	38.3
81	Nepal	60.2	LDC, POP	44.1	35.9	99.0	60.5	99.0	86.4
85	Egypt	57.9	D, AU	74.6	14.8	71.5	23.9	38.9	57.2
89	Rwanda	57.0	LDC, POP, AU	31.1	91.1	95.0	63.2	77.3	87.3
94	China	56.2	ASEAN	61.0	22.3	49.6	68.1	66.2	50.8
118	India	47.7	POP	43.8	28.4	67.6	39.7	62.1	59.7
127	Pakistan	41.1	D	46.1	8.2	37.9	23.0	44.4	66.6
133	Niger	25.7	LDC, D, AU	1.0	22.9	56.6	38.9	50.0	83.6

^(a) Reproduced with permission from [1]. © Yale Center for Environmental Law & Policy, 2006

^(b) To place the EPI and policy category scores into appropriate context selected peer group memberships are provided in the fourth column. They are in order of appearance OECD: Organization for Economic Cooperation and Development; EU: European Union; FTAA: Free Trade Agreement of the Americas; POP: high population density; ASEAN: Association of Southeast Asian Nations plus China, Japan, and South Korea; AU: African Union; NIS: Russia and Newly Independent States; D: desert countries; LDC: least developed countries

links, strong performance on the EPI correlates statistically significantly with per capita GDP ($R^2 = 0.70$), the 2003 Human Development Index (HDI) (the 2003 HDI refers to approximately the same period as the EPI) ($R^2 = 0.77$) [2], and the environmental governance indicator of the 2005 Environmental Sustainability Index ($R^2 = 0.56$) [3]. Explanations for these linkages proposed in the literature [3–6] include that integrated thinking about environmental, economic, educational, and health-related issues combined with a political governance system that permits and encourages the finding of solutions leads – on average – to greater environmental net benefits than sectoral policies alone.

Outlook

The Pilot 2006 EPI work is in progress and is only the beginning of a deeper quantitative analysis of environmental performance and its drivers. At this stage of development, the emphasis is on identifying the appropriate components of a performance matrix, the building of the best possible indicators, and solving methodological questions in the aggregation procedure. The EPI is thus meant to stimulate discourse on these issues as well as to promote the use of quantitative assessments in environmental policy.

References

- [1] Esty, D.C., Levy, M.A., Srebotnjak, T., Sherbinin, A., Kim, C.H. & Anderson, B. (2006). *Pilot 2006 Environmental Performance Index*, Yale Center for Environmental Law & Policy, New Haven. Available at <http://www.yale.edu/epi/>.
- [2] United Nations Development Programme (2003). *Human Development Report 2003*, Oxford University Press, placeNew York, pp. 237–240.
- [3] Esty, D.C., Levy, M. A., Srebotnjak, T. & Sherbinin, A. (2005). *2005 Environmental Sustainability Index: Benchmarking National Environmental Stewardship*, Yale Center for Environmental Law & Policy, New Haven. Available at <http://www.yale.edu/esi/>.
- [4] UN Millennium Project (2005). *Environment and Human Well-being: A Practical Strategy*, Summary version of the report of the Task Force on Environmental Sustainability, The Earth Institute at Columbia University, New York.
- [5] European Environment Agency (2004). *EEA Signals 2004 – A European Environment Agency update on Selected Issues*, European Environment Agency.
- [6] Leiserowitz, A.A., Kates, R.W. & Parris, T.M. (2004). *Sustainability Values, Attitudes, and Behaviors: A Review of Multi-national and Global Trends*, CID Working Paper No. 113, Science, Environment and Development Group, Center for International Development, Harvard University, Cambridge.

Related Articles

Comparative Risk Assessment

Environmental Hazard

History and Examples of Environmental Justice

TANJA SREBOTNJAK

Environmental Remediation

Environmental remediation concerns the quality of the environment, and how to support decision making to decide on how to handle it. Research into exposure pathways has illustrated the dangers of pollutants, such as fine dust, *heavy metals* (see **Lead**; **Arsenic**), or volatiles exuded from former petrochemical sites. Additionally, there has been increasing insight provided into pollutant decay and reabsorption, as well as into complex phase transfer processes, e.g., between soil and groundwater, or air and surface water, allowing a more detailed and forward looking assessment of pollution evolution. Such understanding is necessary given (a) the growing demand for suitable land for habitation and agricultural/industrial use, (b) frequent changes in landuse types that tolerate different levels of contamination, and (c) growing awareness of the effects of pollution among the public. Strict laws detailing precise pollution and exposure limits exist, at least in industrialized nations. The cost of remediation is also rising, and identification of the most cost-effective approach requires detailed knowledge that allows correspondingly detailed treatment depending on specific local contamination.

Environmental remediation concerns assessment of the quality of the three main environmental compartments to define ways to reduce pollution levels to tolerable limits for a given landuse. This is done by assessing the hazard, the exposure, and the dose–response, together commonly called *risk characterization*.

We consider each of the three major environmental compartments: air, water, and soil (see **Air Pollution Risk**; **Environmental Risk Assessment of Water Pollution**; **Soil Contamination Risk**). All can be contaminated and all require a very careful treatment. Soil is the most stable compartment, in the sense that contamination may remain more or less stable for years. Water and air are more fluid and immediate changes in quality may occur. Our main concern will be the soil, as here the major statistical contributions can be made. Remediation ranges from leaving as is, to a total excavation and cleaning of contaminated soil in order to reach a contamination level appropriate for a given landuse. For residential use, a near zero pollution level may be required, whereas

only a reduced level may suffice for the soil below a parking lot. In the soil we generally work with two thresholds: a target value and an intervention value. The target value defines a clean quality of the environment, whereas the intervention value is a more operational value setting the threshold of intervention by either the government or local engineering companies.

We proceed as follows. First we consider the concept of risk in environmental remediation. Next, we consider spatial statistical methods (see **Spatial Risk Assessment**) to indicate their advantages in the remediation domain. We then proceed with *in situ* and *ex situ* remediation and decision support and make the step toward a model. We finish the entry with some concluding remarks.

What Is Risk?

While being easy to define at a general level, a detailed risk characterization makes data collection for a 100% accurate assessment for larger areas impossible or impractical. Various approaches have been developed to assess environmental pollution.

Risk is defined as the probability of a given adverse outcome, commonly calculated as the product of *hazard*, *social* and *physical vulnerability*, and *elements at risk*. A more complete assessment can also include *value*, i.e., the potential cost of a hazardous event. In the context of environmental remediation a similar definition is used [1]:

$$\text{Risk characterization} = \text{hazard} \times \text{exposure} \times \text{dose-response} \quad (1)$$

The overall risk is then a function of the hazard, in terms of its nature, concentration and distribution, the exposure of a given person or group of persons, in terms of intensity, frequency and duration, and the specific relationship between dose and adverse health effects [1, 2]. A similar definition also applies for adverse consequences on the environment or animals.

Hazard identification concerns identification of a substance or agent posing a risk, and assessing its potential to cause adverse effects to health or environment. In soils this translates into high spatial variability that depends on factors, such as soil types, impermeable layers, pollutant type and mobility, and groundwater seepage. In water and air, while

more homogenous, distribution and concentration are subject to currents and wind direction, respectively. In all cases this translates into strong spatial and potentially temporal variability.

Exposure assessment aims at characterizing and quantifying the intensity, duration, and frequency of exposure to the given hazard, ideally for every person. This considers the various pathways of such contact, e.g., *via* direct ingestion of polluted soil, dust, water or air, consumption of soil or dust attached to vegetables, or consumption of pollutants *via* vegetables or meat. For an accurate assessment, mobility of people and local landuse may also be included.

Dose-response considers the specific relationship between the dose of hazardous agent and potentially adverse health or environmental consequences. It is a function of factors such as age, gender, health status, as well as frequency and intensity of contaminant ingestion (see **Dose-Response Analysis**).

In environmental remediation, we are mainly considering spatial risks, i.e., risks to be related to spatial variables. We consider risk as the relation between a spatial variable and a threshold. This is denoted as a spatial variable $X(s)$ in relation to a threshold x_c . We shall not consider the obvious extension toward time variables or toward space-time variables in this article. We can then formulate the hazard as a probabilistic concept:

$$\text{Hazard} = \Pr(X(s) > x_c) \quad (2)$$

It specifies therefore the probability that at a location s , a threshold value x_c is exceeded. In order to make valid statements, stationarity assumptions have to be made, leading to applicability of geostatistical methods. Because of the exploratory nature of this article, we shall not go very deep into these matters [3].

The existence of threshold values depends on the effect of exceeding or not exceeding these values in relation to the quality of life. In all environmental compartments there are critical values, which in case of exceeding require additional measures by local authorities. In air quality measurements, typical thresholds of current scientific concern exist for ozone, NO_x and SO_x . These values typically have a direct effect on the well-being of the population. In groundwater the effects are usually indirect, as much of the drinking water and the irrigation water is cleaned before further application. In groundwater,

however, various chemical toxicants like heavy metals and polyaromatic hydrocarbons are of concern. Usually, the response time is relatively long between threshold exceeding and measurable effects on the (human) population. Even more, in soil studies the time between a measurement above a threshold level and measurable effects to the (human) population is usually rather long. Typical contaminants are heavy metals, mineral oils and polyaromatic hydrocarbons (see **What are Hazardous Materials?**).

The traditional approach to such a problem is a remedial action aimed at multifunctionality [4]. Thus, all the functions the soil can possess, given its natural characteristics, are to be reestablished. This single-perspective view is just one possibility to face the problem. Multiperspective views include elements other than soil protection and have caused a shift of attention from *ex situ* remediation to *in situ* remediation. Within the multifunctional framework, where concentrations are the norm, *in situ* techniques are aimed at reaching threshold values in the shortest possible time. In contrast, techniques in the triple-perspective REC-framework, where *risks*, *environmental merits*, and *costs* are taken into account simultaneously, aim at optimizing within the three-dimensional framework.

Risk Assessment Approaches

Deterministic Methods

A deterministic approach quantifies all parameters required in the risk equation. Soil pollutant concentrations are measured, as well as the amount a person is exposed to, and how that person will react to the pollutant. This leads to a simple source-pathway-receptor framework. First used in the early 1980s it is still being used, for example in the United Kingdom's site-specific assessment criteria (SSAC) procedure [2]. The SSAC approach is very detailed, taking account of exposure duration and time-averaged body weight or vegetable consumption rates of the receptor. The mathematical implementation of the approach, however, is straightforward. Point values used in the parameterization, e.g., pollutant concentrations collected from spot measurements, however, cannot reflect the high spatial-temporal variability of pollutant concentrations [1, 5]. They are, therefore, unsuitable for an assessment of how many people will be affected or exposed. Hence, such approaches

calculate the maximum exposure, thus aiding the *precautionary principle* [6, 7]. Increasing safety factors have traditionally been added, to be on the safe side, leading to risk overestimation. Irrespective of isolated peak concentrations or exposures the entire area is assigned the maximum risk value, leading to excessive and unnecessary remediation costs.

Probabilistic Methods

In the early 1990s, methods were developed that incorporated a characterization of variability, i.e., the natural variation, as well as uncertainty, or lack of knowledge [1]. This allows one to consider the variability in pollutant concentration within the soil, as well as that between individuals, but also to assess the uncertainty in exposure estimation. An important method is Monte Carlo simulation (MCS), which takes account of the variability of the natural phenomenon [8]. Instead of producing a single risk map, MCS produces a distribution of plausible maps that all reasonably match the sample statistics. It shows both the uncertainty of the model parameters and how this uncertainty influences the final risk estimate. In addition to MCS, regression analysis, and analysis of variance (ANOVA) have been used. However, the temporal dimension, critical given the dynamics of natural systems, is frequently neglected.

MCS is further useful to assess the combined effect of several input parameters [1]. Given the multitude of parameter combinations and resultant assumptions and generalizations, a further problem is that it is often not possible to provide risk values with a certainty desired by decision makers. A related problem is the frequent use of precise parameter values and probability distributions [7]. One way to overcome this is by using intervals or probability bound analysis, although these do not reflect well the distribution shape or central tendency. Statistical analysis, however, should be more than a simple tool, and hence should include expert knowledge, especially for site-specific parameter estimation.

Beyond simple MCS the utility of geospatial statistics combined with probabilistic human health risk assessment can be assessed [8]. This allows for a more realistic assessment of spatial variability and local uncertainty, e.g., in soil contamination levels. Such a combined approach better incorporates the spatial dimension in contaminant concentration and distribution, thus allowing a more realistic analysis.

Hybrid of Probabilistic and Fuzzy Methods

To address the uncertainty in the evaluation criteria, Chen *et al.* [9] state that complex parameters such as landuse patterns or receptor sensitivities cannot be modeled with probability functions. Thus, approaches based on fuzzy set theory have been developed that are better equipped to represent the parameter uncertainties. The authors describe a hybrid approach that uses probabilistic methods to address the variability in pollution source and transfer modeling, and fuzzy tools for the evaluation criteria. They also provide a detailed methodology and illustrate with an example. A similar hybrid approach was used by Li *et al.* [10] to evaluate risk from hydrocarbon contamination, who also used the fuzzy set methodology to model different remediation scenarios (from partial to nearly complete cleanup), as well as the spatio-temporal risk variation over different time periods.

In situ and ex situ Remediation

In environmental remediation a distinction is made between *ex situ* and *in situ* remediation. *In situ* remediation concerns techniques like bioremediation, soil washing or extraction, and soil venting [11], whereas containment techniques prevent contamination to migrate. Soil washing uses the solubility of a contaminant, which will dissolve in the percolate and by means of a special withdrawal system the percolate is pumped up and treated. Soil venting aims at volatilization and biodegradation of the contaminant in the unsaturated zone, followed by a vapor treatment system to remove the contaminants from the vapor. Air sparging involves injection of air into the saturated zone for the dual purpose of volatilizing organic components and enhancing biodegradation [11]. *Ex situ* soil remediation concerns excavation of contaminated soil, cleaning this soil at an independent location and putting it back after cleaning, possibly at the same location.

Spatial Statistical Methods

In many remediation studies, the need to quantify risks can be solved by means of geostatistical and other spatial statistical methods. Indicator kriging and probability kriging have been usefully used for risk mapping. Optimal sampling procedures have been

developed to sample an area such that maximum information is obtained given the limited amount of budget for taking observations. Optimization has also taken place for air quality networks and in groundwater (and surface water) observation networks. Spatial simulation methods have been used to show realizations of the random field underlying observations. Different types of observations (measured and organoleptic) have been combined in a geographical information system. Space–time statistics have in particular been useful to assess the changes in concentrations and quality of the environment (see **Spatiotemporal Risk Analysis**). Modern approaches by means of model-based Geostatistics are a recent step forward. Finally, the combination of spatial statistical methods with soil, air, and groundwater models has been performed in many systems and detailed case studies.

Example: a Cost Model for Environmental Remediation

Sanitation of contaminated soil is an expensive task, which is difficult to assess. Estimated costs are sometimes exceeded by 100% [12]. Here we show the use of a cost model to a former gasworks site in the harbor area of the city of Rotterdam [13]. In 1994, an integrated soil inventory was carried out (10 observations/ha). We concentrate on the upper layer (0–0.5 m), containing the largest set of the full vector of observations (Figure 1). The five contaminants indicative of the spatial extent of contamination are cyanide, mineral oil, lead, polyaromatic hydrocarbons PAH-total, and zinc.

Four environmental thresholds classify areas according to the degree of contamination: *S* value: (*Safe* or *Target level*): maximum concentration of a contaminant to maintain multifunctionality, *T* value: intermediate level between low and serious soil pollution, *I* value (*Intervention level*): if this value is exceeded the functional properties of the soil for humans, fauna, and flora are seriously degraded or are threatened to be seriously degraded and the BAGA level (“*Chemical waste*”): at this level the soil is chemical waste and should be carefully treated (Table 1).

A cost model combines the thresholds with remediation techniques. Ten classes with different processing costs are distinguished. The highest degree of

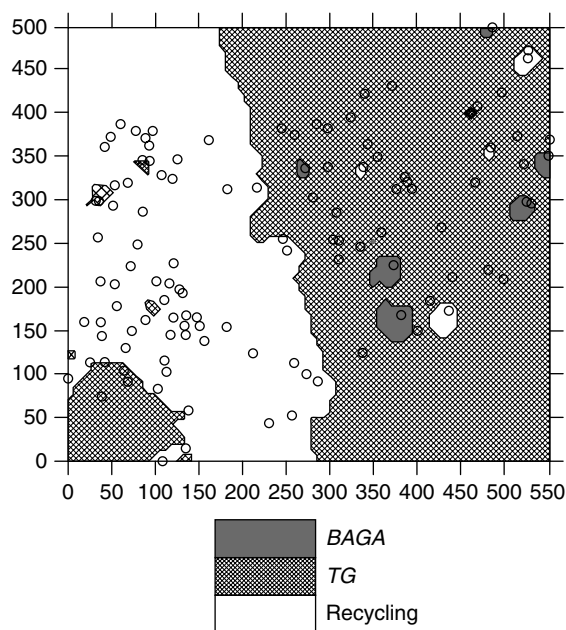


Figure 1 Location of sample points in the area

Table 1 Environmental thresholds for five contaminants. All thresholds are expressed in (mg kg^{-1})

Contaminant	<i>S</i> value	<i>T</i> value	<i>I</i> value	BAGA
Cn	5	27.5	50	50
Mineral oil	50	2525	5000	50 000
Pb	85	307.5	530	5000
PAH	1	20.5	40	50
Zn	140	430	720	20 000

pollution of any contaminant determines whether a soil cube should be excavated or not, depending upon different remediation scenarios. Dividing the area into 5760 squares of 7 by 7 m^2 , we find that treatment of the area costs €3.11 M. More details can be found in Stein and Van Oort [14].

Concluding Remarks

Environmental remediation has received quite some recent attention, with the advance a growing awareness of the effects of contamination for health and the environment, and with the advance of spatial statistical methods to quantify the associated risks. A current trend is a further focus on individual risks,

a better quantification of those, and a more realistic assessment. Advances can be expected at each of the components of risk assessment: realistic modeling, improved observation methods e.g., by using remote sensing derived methods, and realistic vulnerability assessments.

References

- [1] Öberg, T. & Bergbäck, B. (2005). A review of probabilistic risk assessment of contaminated land, *Journal of Soils and Sediments* **5**, 213–224.
- [2] Nathanail, P., McCaffrey, C., Earl, N., Foster, N.D., Gillett, A.G. & Ogden, R. (2005). A deterministic method for deriving site-specific human health assessment criteria for contaminants in soil, *Human and Ecological Risk Assessment* **11**, 389–410.
- [3] Webster, R. & Oliver, M.A. (2001). *Geostatistics for Environmental Scientists*, John Wiley & Sons, Chichester.
- [4] Robberse, J.G. & Denneman, C.A.J. (1993). Do target values help to protect the soil?, in *Proceedings of the Fourth International KfK/TNO Conference on Contaminated Soil*, F. Ahrendt, G.J. Annokkée, R. Bosman & W.J. van den Brink, eds, Kluwer Academic Publishers, Dordrecht.
- [5] Bates, S.C., Cullen, A. & Raftery, A.E. (2003). Bayesian uncertainty assessment in multicompartiment deterministic simulation models for environmental risk assessment, *Environmetrics* **14**, 355–371.
- [6] Morris, J. (2000). *Rethinking Risk and the Precautionary Principle*, Elsevier, pp. 294.
- [7] Sander, P., Bergback, B. & Oberg, T. (2006). Uncertain numbers and uncertainty in the selection of input distributions – consequences for a probabilistic risk assessment of contaminated land, *Risk Analysis* **26**, 1363–1375.
- [8] Gay, J.R. & Korre, A. (2006). A spatially-evaluated methodology for assessing risk to a population from contaminated land, *Environmental Pollution* **142**, 227–234.
- [9] Chen, Z., Huang, G.H. & Chakma, A. (2003). Hybrid fuzzy-stochastic modeling approach for assessing environmental risks at contaminated groundwater systems, *Journal of Environmental Engineering–ASCE* **129**, 79–88.
- [10] Li, J.B., Liu, L., Huang, G.H. & Zeng, G.M. (2006). A fuzzy-set approach for addressing uncertainties in risk assessment of hydrocarbon-contaminated site, *Water Air and Soil Pollution* **171**(1–4), 5–18.
- [11] Rulkens, W.M., Grotenhuis, J.T.C. & Soczó, E.R. (1993). Remediation of contaminated soil: state of the art and desirable future developments, in *Proceedings of the Fourth International KfK/TNO Conference on Contaminated Soil*, F. Ahrendt, G.J. Annokkée, R. Bosman & W.J. van den Brink, eds, Kluwer Academic Publishers, Dordrecht.
- [12] Okx, J.P. & Stein, A. (2000). Use of decision trees to value investigation strategies for soil pollution problems, *Environmetrics* **11**, 315–325.
- [13] Broos, M.J., Aarts, L., Van Tooren, C.F. & Stein, A. (1999). Quantification of the effects of spatially varying environmental contaminants into a cost model for soil remediation, *Journal of Environmental Management* **56**, 133–145.
- [14] Stein, A. & Van Oort, P. (2006). Concepts of handling spatio-temporal data quality for cost functions, in *Fundamental of Spatial Data Quality*, R. Devillers & R. Jeansoulin, eds, ISTE, London.

Related Articles

Environmental Health Risk

Environmental Monitoring

Environmental Risk Regulation

ALFRED STEIN AND NORMAN KERLE

Environmental Risk Assessment of Water Pollution

Environmental risk due to water pollution is an ongoing every day problem. Municipal wastewater systems are one of the largest pollution sources to surface waters, however; other major sources include residential and industrial discharges, and agricultural runoff, dumped into rivers directly or by seepage from contaminated groundwater. Threats from chemicals may come from sources such as pesticide use and waste disposal. Air pollution and land pollution also affect water quality (*see Air Pollution Risk; Soil Contamination Risk*).

Water resources are considered as one of the main factors affecting our life, industries, agriculture, and economics. Therefore, we must properly assess the risk of water pollution from these resources. Consequently, measures should be taken in order to protect water sources from pollution. For these reasons, we have to identify its contents and detect the sources that cause these contents to deviate from their safety and healthy levels.

This work presents an illustration of two statistical techniques that can assess the risk of surface water pollution from bacterial and chemical sources, as applied to a sample from Tigris river water after it has passed through the Baghdad metropolitan area. The first is an adaptation of existing methods to lognormal data, while the second is an application of intervention analysis with time series data.

Data Description

The data consisted of samples collected over 48 weeks. At each sample point six measurements were made, three on bacterial variables (coliform, fecal coliform, and plate count), and three on chemical variables (chloride, T-hardness, and conductivity). Samples were taken from two selected locations on Tigris river during 1988. The first was located north of Baghdad, approximately 35 km from city center (upper stream), and the second was situated south of Baghdad, approximately 20 km from city center (downstream). These two locations bracket most of

the pollution sources. Furthermore, to account for the river width and water current speed, measurements were replicated at four points across the river at time t . These measurements were also correlated. The study area brackets major pollution sources such as chemical factories and vegetation farms that use fertilizers. Almost all of these sources are on the west side of the river. These six measurements generated six time series, each of 48 observations (with four readings across the river) in which successive observations were serially dependent and may be nonstationary, or with seasonal effects.

Adapting the Lognormal Model

Suppose D_{kt} is the k th downstream measurement ($k = 1, 2, 3, \dots, K$) across the river observed at time ($t = 1, 2, 3, \dots, T$) on a pollutant content in the water (chemical or bacterial), and U_{kt} denotes the corresponding upper stream measurement. Define the ratio between observations at downstream and upper stream, by $R_{kt} = D_{kt}/U_{kt}$; hence $D_{kt} = R_{kt}U_{kt}$. Considering all major pollution sources on one side of the river (west side) and taking into account the width of the Tigris river along with the water current speed, a recursive interdependency between the four readings across the river can be derived. The ratio R_t can be rewritten as (see [1])

$$\begin{aligned} R_t &= R_{0t} \prod_{k=1}^K L_k = R_{0t} \left[\prod_{k=1}^K L_k^{1/K} \right]^K \\ &= R_{0t} [G_K]^K \quad t = 1, 2, \dots, T \end{aligned} \quad (1)$$

where R_{0t} is the initial value for the ratios, $(L_1, L_2, \dots, L_K)$ is a positive random vector and G_K is the geometric mean (GM) for $L_1, L_2, \dots, L_K$. Therefore, from relations (1) we can deduce that the ratio R_t is proportional to the geometric mean of the ratios at sample points across the river. A significant deviation of the expected value of R_t from unity (>1) indicates an increase in pollution level. Letting $L_k = 1 + \delta_k$, and substituting the Taylor expansion of $\ln(1 + \delta_k)$ in equation (1), $\ln R_t$ can be approximated by

$$\ln R_t = \ln R_{0t} + \sum_{k=1}^K \delta_k \quad (2)$$

By the additive central limit theorem $\ln R_t$ is asymptotically normally distributed, and by definition R_t will have a lognormal distribution, $R_t \sim \Lambda(\gamma, \mu, \sigma)$, where γ is location (threshold) parameter, μ is the mean, and σ is the standard deviation. Using R as a generic symbol for R_t the random variable $Y = \ln(R - \gamma)$ is normal, $N(\mu, \sigma^2)$, $\sigma > 0$, and $R > \gamma$.

The parameters γ, μ, σ of the lognormal distribution will be estimated using “modified moment estimation” (MME) (see [2]). For an ordered sample of size n , the estimating equations are given by

$$\gamma + e^\mu e^{\sigma^2/2} = \bar{r} \quad (3)$$

$$e^{2\mu} e^{\sigma^2} (e^{\sigma^2} - 1) = s^2 \quad (4)$$

$$\gamma + e^\mu \exp(\sigma E Z_{1,n}) = r_1 \quad (5)$$

where $\bar{r}$ and s^2 are the sample mean and variance (unbiased); r_1 is the observed first-order statistics, of a random sample of size n , and $E Z_{1,n}$ is the expected value of the first-order statistic for a sample of size n from $N(0, 1)$.

From equation (4) the unbiased estimator for μ is

$$\hat{\mu} = \ln\{s[e^{\hat{\sigma}^2}(e^{\hat{\sigma}^2} - 1)]\}^{-1/2} \quad (6)$$

and $\hat{\sigma}$ can be found using Table 5 or Figure 2 in [2], by entering the table or the figure with $J(n, \hat{\sigma}) = s^2/(\bar{r} - r_1)$ to read $\hat{\sigma}$. Finally, γ can be estimated from equation (5).

The summary of the coliform data collected from the Tigris river is as follows: $K = 4, n = 48$ pairs of measurements resulted in 48 ratios(r), $\bar{r} = 53.295, s = 113.99, r_1 = 1.51$. Using the MME procedure, $\hat{\sigma} = 1.32, \hat{\gamma} = 0.782, \hat{\mu} = 3.976$. The ratios of coliform bacteria “geometric means” follow a lognormal distribution, $R \sim \Lambda(0.782, 3.976, 1.32)$. Hence $\ln(R - \gamma)$ is normal. Testing $H_0 : E(R - \gamma) = 1$ against $H_1 : E(R - \gamma) > 1$, yield Z score = 9.54, which is highly significant ($p \approx 0$) (see [1]).

Using Intervention Analysis in Time Series

Consider the data $(\Lambda, z_{t-1}, z_t, z_{t+1}, \Lambda)$, which represents a time series over equal time intervals under certain intervention effects. The time series model for Z_t , before intervention (upper stream), can be written as follows:

$$\phi(B)Z_t = \theta(B)a_t \quad (7)$$

where $\phi(B)$ and $\theta(B)$ are polynomials in B , the backshift operator, $(BN_t = N_{t-1})$, and $\{a_t\}$ is sequence of white noises [3]. The generic model after intervention (downstream) can be written as:

$$y_t = f(\alpha, \xi_t, t) + N_t \quad (8)$$

where y_t is some transformation of Z_t ; ξ_t is exogenous variable (intervention), α is the set of unknown parameters, and N_t represents is a noise. The noise N_t can be transformed to a white noise using equation (7). The exogenous variable model ξ_t can be represented by the dynamic model:

$$f(\delta, \Omega, \xi_t, t) = \sum_{j=1}^I \eta_{tj} = \sum_{j=1}^I \left[\frac{\Omega_j(B)}{\delta_j(B)} \right] \xi_{tj} \quad (9)$$

where η_{tj} is the dynamic transformation of ξ_{tj} and $\delta_j(B), \Omega_j(B)$ are polynomials in B of degrees r_j and s_j , respectively, and I represents number of intervention (which is equal to 1 for our analysis). Transforming η_t to ξ_t can be achieved by the linear difference equation:

$$\delta(B)\eta_t = \Omega(B)\xi_t \quad (10)$$

A step function, $S(t, T) = \begin{cases} 0, & t < T \\ 1, & t \geq T \end{cases}$, can adequately represent the intervention variable ξ_t .

An output step change of unknown magnitude would be produced according to

$$\eta_t = \omega BS(t, T) \quad (11)$$

Finally the general intervention model can be written as follows:

$$Z_t = R(B)\xi_t + \phi^{*-1}(B)\theta^*(B)a_t \quad (12)$$

where $(\phi^*(B), \theta^*(B))$ are the estimates of $(\phi(B), \theta(B))$, and $R(B) = \Omega(B)/\delta(B)$. Equation (12) can be expressed in a regression format

$$Y_t = \beta X_t + a_t \quad (13)$$

where $Y_t = \phi^*(B)\theta^{*-1}(B)Z_t, X_t = \phi^*(B)\theta^{*-1}(B)\xi_t$ and $\beta = R(B)$. The least-squares estimate of β is $\beta^* = \sum_{t=1}^n y_t x_t / \sum_{t=1}^n x_t^2$ with variance $V(\beta^*) = \sigma_a^2 (\sum_{t=1}^n x_t^2)^{-1}$. Applying the methodology to the coliform data after transforming it to a stationary series, the Autoregressive Integrated Moving Average (ARIMA (0,1,1)) model showed a best fit to upper

stream series, with estimated noise model $N_t = (1 - 0.71529B)a_t$. For the downstream series, the least-squares estimate for the regression equation (13) is $Y_t = 79247X_t + 23218$, with $R^2 = 63.7\%$, for details (see [1]). The model and parameter tests were highly significant ($p \cong 0$) assessing a high risk of coliform bacteria concentration in Tigris river at the selected region. For test results on the other pollutants in Tigris river (see [1]).

References

- [1] Al-Khalidi, A.S. (2002). Measuring water sources pollution using intervention analysis in time series and log-normal model (a case study on Tigris river), *Environmetrics* **13**, 693–710.
- [2] Crow, E.L. & Shimuzu, K. (1988). *Log-normal Distributions. Theory and Applications*, E.L. Crow & K. Shimuzu, eds, Marcel Dekker, New York.
- [3] Box, G.E.P. & Taio, G.C. (1975). Intervention analysis with applications to economics and environment problems, *Journal of the American Statistical Association* **70**, 70–80.

ABDUL SATTAR AL-KHALIDI

Environmental Risk Regulation

Modern governments adopt laws and regulations to protect environmental quality and human health. The fundamental justification for environmental regulation is to control harmful externalities that arise when an individual, firm, or any other organization releases pollutants to the environment that cause harm to other people (e.g., air pollution from burning fossil fuels causes human death and disease, adverse effects to ecosystems, and alters global climate). Regulation may also be justified as a method to protect individuals and households from their own poor decisions, as when individuals may have insufficient information to make informed decisions (e.g., advisories to limit consumption of fish from contaminated water bodies).

Regulation of *environmental hazard* (see **Environmental Hazard**) involves questions of risk assessment, standard setting, and regulatory mechanism. Risk assessment describes how the probability and severity of harm depends on the level of pollution (see **Environmental Health Risk**). Standard setting is the choice of how much pollution to allow, and the regulatory mechanism is the legal framework, which influences how much each polluter can release.

Risk Assessment

Risk assessment characterizes the possible harms (or benefits) of exposure to a pollutant and their probabilities. Regulatory decision makers are interested in both the average risks in a population and the distribution or variability of risk in the population, including risks to the most highly exposed and most sensitive individuals. Population risks are often described in terms of the expected number of cases (e.g., fatalities or cancers), but this shorthand form of expression may be misleading. For most environmental risks, the identities of the individuals who suffer the adverse outcomes are unknowable both *ex ante* and *ex post*.

Risk assessment is conventionally divided into four components: hazard identification, dose–response assessment, exposure assessment, and risk characterization [1].

Hazard identification involves determining what health effects can be produced by the chemicals, radiation, or other agents, and under what conditions. For example, an agent may cause cancer or other disease, may be acutely lethal, or may cause some form of nonfatal illness, or reproductive anomaly. Hazard identification often relies on short-term tests of laboratory animals or microorganisms and yields a qualitative statement about the possible health consequences of exposure to an agent and the conditions under which these effects may occur (see **Environmental Hazard**; **Environmental Risks**).

Dose–response assessment involves determining how the type, severity, and probability of the health effects depend on exposure to the agent (see **Dose–Response Analysis**). Dose–response assessment is critical to characterizing risk for, as the sixteenth century physician Paracelsus recognized, “All substances are poisons; there is none which is not a poison. The right dose differentiates a poison and a remedy” [2].

For carcinogens, it is conventionally assumed that the type and severity of the cancer that may be caused is independent of dose but the probability of developing cancer is proportional to dose (for small dose levels), which implies a positive probability at all exposure levels above zero. This “linear no-threshold” model was initially motivated by the notion that a single genetic mutation can initiate a process that leads to development of cancer (see **Statistics for Environmental Teratogenesis**). In some cases, there may be a threshold dose below which the probability of harm is zero and above which the probability or severity of harm increases with increasing dose. The threshold model represents a situation in which the body’s immunological or other defense mechanisms can effectively protect against low doses, but are overwhelmed by higher doses of the agent. A third possibility is a “hormetic” dose–response function, in which a low dose of the agent may cause a beneficial effect but larger doses may cause a harmful effect (for example, exposure to sunlight promotes vitamin D production at low dose but causes skin cancer at high dose).

In some cases, risk assessment purports to identify a “safe” dose or exposure. For example, the “reference dose” is “an estimate (with uncertainty spanning perhaps an order of magnitude) of daily exposure to the human population (including sensitive subgroups)

that is likely to be without an appreciable risk of deleterious effects during a lifetime” [3]. In cases where the dose–response relationship has a threshold, the threshold dose satisfies this criterion but in cases with no threshold, identification of a “safe” dose depends on the definition of “safe”.

Exposure assessment involves determining the quantities of the agent to which individuals are exposed, together with exposure conditions such as route and time pattern when these influence the probability of harm. For some agents (e.g., carcinogens), the probability of the adverse effect is treated as a function of cumulative exposure over an individual’s lifetime. For others (e.g., acute toxins), the maximum exposure over a short period may be more relevant.

In some cases, dose is not explicitly considered and risk assessment substitutes an exposure–response or concentration–response function for a dose–response function. For example, assessments of the risk from exposure to air pollutants may describe the probability of adverse health effects as a function of the average daily or peak hourly concentration of the pollutant in ambient air (*see Air Pollution Risk*).

Risk characterization involves aggregating the results of the other three components to describe the risk, i.e., the potential health effects and their probabilities of occurrence, including uncertainty about these matters.

Standard Setting

Setting an appropriate standard for environmental risks requires balancing the benefits of reduced risk against the costs, or other harm incurred in reducing the risk (*see Risk–Benefit Analysis for Environmental Applications*). Any significant reductions (ancillary benefits) and increases (countervailing risks) in other risks should also be considered [4]. This comparison may be explicit, as when standards are based on economic evaluation, or implicit as when standards are based on criteria such as “protection of public health” or “as low as reasonably achievable”.

Benefit–cost analysis (BCA) is a method to account for all the significant consequences of a regulation, quantifying them in monetary units. BCA allows one to rank alternative policies by the expected

value of the net benefits (i.e., the expected difference between benefits and costs) to determine which of these produces the greatest expected gain to society. An alternative approach, cost-effectiveness analysis (CEA), measures some of the consequences in nonmonetary “effectiveness” units (e.g., “lives saved” or deaths prevented from a specific risk) and estimates the ratio of cost per unit effectiveness. CEA allows one to compare the efficiency with which policies produce units of effectiveness but the question of whether the beneficial effects justify the costs incurred requires an independent judgment.

Principles of Economic Evaluation

An environmental regulation will typically benefit some people and harm others. A critical question is how to determine whether the harm caused to some is outweighed by the benefit conferred on others. Economic welfare analysis begins with the assumption that one cannot reliably compare gains or losses in welfare to different people, i.e., if two people develop the same illness, one cannot determine who suffers more. Economic evaluation distinguishes two aspects of a change in social outcomes, efficiency and the distribution of well-being among people. A situation is defined to be *Pareto* or *allocatively efficient*, if it is not possible to change matters in a way that benefits at least one person without harming anyone. Many situations may be efficient but differ in the distribution of well-being in a population. If everyone at the dinner table prefers a larger to a smaller slice of pie, any division that leaves no crumbs is efficient.

Any change in environmental regulation that is a Pareto improvement is efficiency enhancing and there is a presumption that it should be adopted [5]. However, since many potential changes are neither Pareto superior nor Pareto inferior to the *status quo*, regulations are usually evaluated by whether they satisfy the Kaldor–Hicks compensation test: those who benefit from the change could provide monetary compensation to those who are harmed so that everyone prefers the change with compensation to the *status quo*. Whether a change that causes harm to some people yet satisfies the compensation test qualifies as a social improvement can be questioned. The arguments in favor of adopting the test include the following: (a) adopting policies that satisfy the

test expands the “social pie”, which at least admits the possibility that everyone gets a bigger slice; (b) if some redistribution of resources is desired (e.g., to reduce inequality), it may be better achieved through more direct transfers, such as taxation and welfare policies, than by adopting inefficient environmental regulations; (c) if society follows the compensation test routinely, the groups that benefit and are harmed will tend to differ from case to case and everyone may be better off than if decisions are made on some other basis.

BCA provides a method of determining whether a proposed change satisfies the Kaldor–Hicks compensation test. For an individual, benefits are measured as the maximum amount the person would be willing to pay to obtain the changes he views as beneficial and costs as the minimum amount of compensation he would require to accept the changes he views as adverse. Summing benefits and costs across the population determines whether those who gain would be able to compensate those who are harmed and still prefer the change.

Defining and Measuring Benefits and Costs

Economic evaluation attempts to measure the “social” or “resource” benefits and costs of a policy or activity. Some private benefits and costs increase or decrease social welfare, but others are simply transfers within a society. For example, a requirement to install catalytic converters in automobiles is a private cost to new-vehicle purchasers. It is also a social cost, because workers’ time and materials used to produce the catalytic converters cannot be used to produce something else of value. In contrast, a tax on gasoline purchases is not a social cost but a transfer from gasoline consumers to the government, which can use these resources to purchase other goods. (Imposing a gasoline tax may create a social cost by preventing consumers who would be willing to pay the cost of the gasoline, but not the cost plus the tax, from using gasoline.)

The social cost of an action to reduce environmental risk is the value of the resources used to reduce the risk. It may include the cost of control technologies (such as catalytic converters), changes in processes (such as substituting fuel injection for carburetion), or changes in inputs (such as substituting natural gas for petroleum fuel). When multiple technologies are

employed, one may be able to estimate the incremental cost by comparing the prices of the technologies, controlling for other factors using statistical methods. Alternatively, costs may be estimated using engineering models to project the quantities of resources required to produce the new technology and the market prices or other estimates of the social value of the resources. The costs of new technologies that have not yet been produced at commercial scale are necessarily rather speculative.

Benefits are estimated by “revealed-preference” or “stated-preference” methods [6, 7]. Revealed-preference methods are based on the assumption that an individual chooses from a set of available alternatives the one he prefers most. For example, the benefit of reduced mortality risk is often estimated by examining the extra compensation workers receive for more dangerous jobs [8]. Stated-preference methods rely on asking survey respondents what choices they would make in a clearly defined setting. For example, respondents may be asked if they would vote to support or oppose a law reducing air pollution from electric plants that would increase the price of electricity. Revealed-preference methods are often preferred since they are based on consequential choices. Stated-preference methods are more flexible since they are not limited to situations where the actual choices of people can be observed and respondents can be questioned about the attribute of interest, e.g., future changes in environmental risk.

Regulatory Mechanism

Regulations are classified as using “command-and-control” or “economic-incentive” mechanisms. Command-and-control mechanisms specify either the technologies that must be used to limit environmental pollution (e.g., “end-of-pipe controls” such as catalytic converters on automobiles), or the allowable quantity or rate of emissions (e.g., performance standards such as vehicle miles per gallon). Command-and-control instruments cannot be easily tailored to accommodate differences in pollution-control costs between firms and so they may not minimize the social cost of reducing pollution. Technology standards also require the government to select the technology, even though firms may have better information about the costs of alternative methods for reducing pollution.

Economic-incentive mechanisms provide firms with an incentive to account for the external costs of emissions, but allow them flexibility in how emissions are reduced. The primary forms of incentive-based instruments are taxes and tradable permits.

An emissions tax internalizes the external damages associated with a firm's pollution. It provides an incentive for firms to reduce emissions when the tax exceeds the incremental cost of emission reduction, and to not reduce emissions when the tax is less than this cost. If the tax is set equal to the social benefit of reducing a unit of pollution, the socially optimal level of pollution should be achieved.

Under a tradable-emission-permit system, a limited quantity of permits is issued and each firm is required to have a permit for every unit of pollution it releases. This establishes a performance standard but adds the option of interfirm flexibility. Firms may buy and sell permits among themselves. The permit price functions analogously to an emissions tax. Firms that can reduce emissions at a cost lower than the permit price will do so, and either buy fewer permits or sell any excess permits they may have. Firms that cannot reduce emissions at a cost lower than the permit price will rather purchase additional permits to cover their emissions.

Several case studies of environmental risk regulation including the application of risk assessment, BCA, and choice of regulatory mechanism are described in [9].

References

- [1] National Research Council (1983). *Risk Assessment in the Federal Government: Managing the Process*, National Academy Press, Washington, DC.

- [2] Rodricks, J.V. (1992). *Calculated Risks: The Toxicity and Human Health Risks of Chemicals in Our Environment*, Cambridge University Press, Cambridge.
- [3] U.S. Environmental Protection Agency (1995). *The Use of the Benchmark Dose Approach in Health Risk Assessment*, EPA/630/R-94/007, Washington, DC.
- [4] Graham, J.D. & Wiener, J.B. (eds) (1995). *Risk Versus Risk: Tradeoffs in Protecting Health and the Environment*, Harvard University Press, Cambridge.
- [5] Okun, A.M. (1975). *Equality and Efficiency: The Big Tradeoff*, The Brookings Institution, Washington, DC.
- [6] Freeman III, A.M. (2003). *The Measurement of Environmental and Resource Values: Theory and Methods*, 2nd Edition, Resources for the Future, Washington, DC.
- [7] Hammitt, J.K. (2000). Valuing mortality risk: theory and practice, *Environmental Science and Technology* **34**, 1396–1400.
- [8] Viscusi, W.K. & Aldy, J.E. (2003). The value of a statistical life: a critical review of market estimates throughout the world, *Journal of Risk and Uncertainty* **27**, 5–76.
- [9] Morgenstern, R.D. (ed) (1997). *Economic Analyses at EPA: Assessing Regulatory Impact*, Resources for the Future, Washington, DC.

Related Articles

Economic Criteria for Setting Environmental Standards

JAMES K. HAMMITT

Environmental Risks

What Is Meant by the “Environment”?

The term *environment* in the context of environmental health has been defined based on a variety of different dimensions. In its 2006 comparative risk assessment of the environmental contribution to the global burden of disease, the World Health Organization (WHO) acknowledged that the term environment can include not only the physical environment, but can also be interpreted to include the natural, social, and behavioral components. This broad construction of environment is consistent with the definition offered by Last [1], “All that which is external to the human host. Can be divided into the physical, biological, social, cultural, etc., any or all of which can influence the health status of populations...” However, the WHO ultimately offered a more “practical” definition of the environment in the context of risk factors that can be modified to address environmental health risks as follows, “The environment is all the physical, chemical and biological factors external to a person, and all the related behaviors” [2].

The term *environment* can also be defined based on spatial and/or temporal factors. For example, spatial definitions of environment can be characterized (from narrowest to broadest) as personal, micro, local, metropolitan, regional, national, and global. With respect to exposure to airborne contaminants, the environment is often divided between the outdoors and indoors. The environment of an individual or a population can also be defined by the time frame and activities (i.e., time–activity patterns) of that individual or population.

The “environments” affecting health also have an inherently hierarchical structure that extends from the individual to the global levels. Individuals can determine some aspects of their personal environment, but environments at higher levels of organization, e.g., local, regional, and national, are not under their control, even if critical to health status.

One useful conceptual formulation of the structure of environments is the “microenvironmental model”, often applied to air pollution exposure and risk. In this model (1), total personal exposure (E) is the time-averaged pollutant concentration for an individual in

the various places, or microenvironments, where the person spends time. Thus,

$$E = \sum_i c_i t_i \quad (1)$$

where c_i is the pollutant concentration in microenvironment i , and t_i is the time spent there. In some microenvironments, the pollutant concentration might be affected by occupant activities, such as smoking – a source of airborne particles–while in others, concentration could be determined by population-level factors, e.g., traffic.

Environmental Health Risks

Risks to human health from environmental agents can occur from a wide variety of sources that can result in an increased risk of disease or injury *via* various biological or physiological processes, or of premature death. Rodricks [3] has proposed five broad categories of environmental agents that can increase the risk of adverse health outcomes: pathogenic agents; natural and synthetic chemicals; radiation; nutritional substances; and physical objects. Table 1 provides examples of specific types of agents under these broad categories.

For the purposes of this section of the encyclopedia, the focus is on natural and synthetic chemicals that are considered environmental pollutants, i.e., toxic substances introduced into the environment, or naturally occurring substances which have demonstrated an adverse impact on health, and also on radiation.

Environmental pollutants are typically categorized according to media-specific regulatory structures (e.g., air, water, and foods). These pollutants are regulated through limits on acceptable levels in these media, or through limits on the emissions from sources of these pollutants *via* these media into the environment. For example, the United States Environmental Protection Agency (USEPA) and the WHO have established ambient air quality standards and guidelines, respectively, that indicate the concentration of pollutants in the air to levels deemed to represent minimal (though not zero) levels of health risk. Similar types of acceptable pollutant levels have been established for surface and drinking water, as well as pesticide residue in foods.

Table 1 Five categories of environmental agents that can contribute to increased human health risks^{(a)(b)}

Pathogenic agents	Natural and synthetic chemicals	Radiation	Nutritional substances	Physical objects
Many microorganisms (bacteria, fungi, viruses, parasites)	Natural products (chemicals in foods, beverages, plants, animals, insects)	All forms of ionizing and nonionizing radiation (including heat, sound)	Constituents of foods and beverages that are necessary for nutrition	Weapons
	Industrial products and by-products			
	Consumer products			
	Pesticides, agricultural chemicals			
	Medicines, medical and diagnostic devices			
	Chemicals used for clothing, shelter, other physical structures and objects			Machinery/equipment/ tools
	Tobacco and its combustion products			Moving vehicles
	Substances used for fuels and their combustion products			Components of physical structures
	Substances of abuse			Water
	By-products and wastes from the above			Natural formation and products

^(a) Reproduced from [3]. © Heldref Publications, 1994.

^(b) Some subgroups contain many individual agents; naturally occurring chemicals in the diet probably make up the single largest subgroup of individual chemicals. Distinctions are made between pathogenic microorganisms that cause harm by invading the body and growing there, and those that produce chemical toxins outside the body that cause harm when ingested; the latter are in the group of natural chemical products. Similarly, though the constituents of physical objects are all chemicals, the agents of harm are the objects themselves, and the harm they may create (usually some form of physical trauma) is not related to the hazards of their chemical components

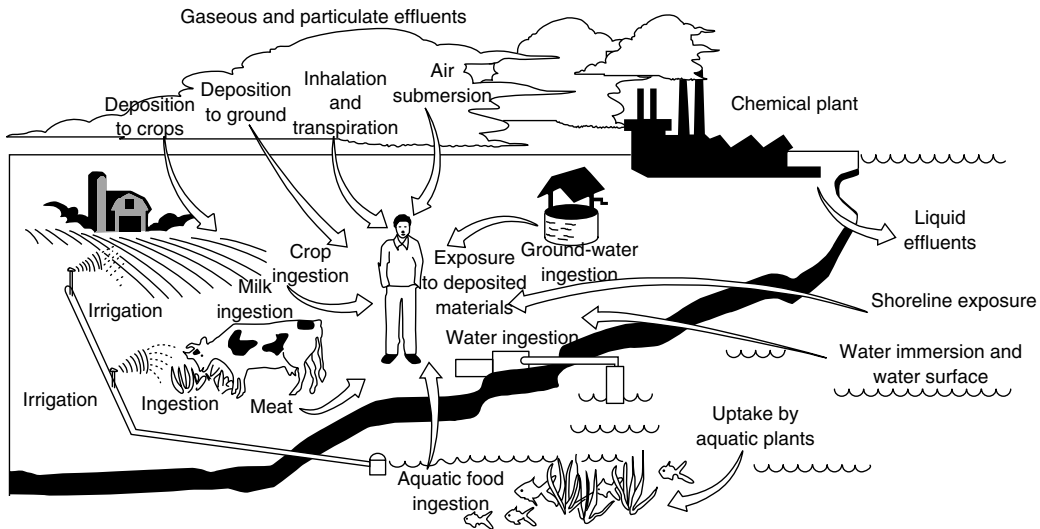


Figure 1 Example of multiple exposure pathways to environmental pollutants [Reproduced from [3]. National Institute of Environmental Health Sciences, 1994.]

Environmental pollutants can also be categorized according to the ubiquity of exposure and type of health outcome associated with exposure. For example, under the structure established by the US Clean Air Act, pollutants associated with cancer, neurotoxic or reproductive health effects, but with more limited population exposure, are classified and regulated as “hazardous air pollutants”, while exposure to “criteria air pollutants” is presumed to be more widespread, but typically not associated with more severe health outcomes, such as cancer. As discussed in the case study that follows, airborne particulate matter is an example of an exception to this classification approach, as it is classified as a criteria pollutant given its widespread exposure profile, even though it is associated with lung cancer and premature death.

In developed countries, increasing attention has been focused on environmental health risks associated with relatively low levels of exposure to well-studied pollutants such as airborne particulate matter and environmental tobacco smoke, as well as pollutants of emerging concern that have potential neurological effects, or those that affect the endocrine system and are associated with potentially adverse reproductive effects. Developing countries typically face environmental health risks associated with significantly higher levels of environmental

contaminants resulting from pollutant emissions from rapid industrialization, as well as home heating and cooking, and rudimentary sanitary practices.

Pathways and Routes of Exposure to Environmental Contaminants

Humans are exposed to environmental contaminants from diverse sources through various *pathways*. Pathways are the physical course taken by a contaminant from the source to the exterior of the exposed person or organism. These pathways include air, water, food, and soil contact. The contaminants are then transferred from the exterior to the interior of the exposed person or organism through several possible *routes* of exposure: inhalation (*via* air); ingestion (*via* water or food consumption); or dermal absorption (*via* water or soil contact). Figure 1 provides an illustration of the sequence from sources of pollution to human exposures pathways and potential routes of exposure, finally resulting in adverse health effects.

Exposure to some pollutants occurs *via* only one pathway. Carbon monoxide, a gas formed *via* incomplete combustion, pollutes outdoor and indoor air and is inhaled; breathing is the only pathway. By contrast, exposure to lead might occur through

inhaled air, water contaminated from lead-containing pipes, or ingestion of lead-contaminated dust.

Environmental and Occupational Health Risks

Environmental health risks can be distinguished from the risks to workers from pollutants in occupational settings in three significant categories: (a) the magnitude and susceptibility to disease of the potentially exposed populations, (b) the number of contaminants that populations are exposed to, and (c) the levels of exposure to the pollutants and the magnitude of excess risk resulting from such exposure.

The size of the populations potentially exposed to widespread environmental contaminants is typically two to three orders of magnitude larger than that of workers exposed to occupational pollutants. Whereas the size of worker populations potentially exposed to dangerous levels of specific contaminants is often in range of tens or hundreds of thousands, a population exposed to unhealthy levels of contaminants in the environment are typically in the range of tens or hundreds of millions. For example, the USEPA estimates that 133 million people in the United States (out of a total population of 281 million) lived in areas where monitored air quality in 2001 violated at least one health-based national air quality standard [4]. By comparison, the US National Institute of Occupational Safety and Health (NIOSH) estimates based on a survey taken in the early 1980s (the most recent data available) that approximately 2.5 million US workers were potentially exposed to the chemical 1,1,1-trichloroethane (TCE), one of the most ubiquitous chemicals used in industrial applications, though approximately 70% of workers exposed to TCE in occupational settings received no protection from these exposures [5, 6].

In addition to major differences in the size of exposed populations, there are significant differences in the susceptibility to disease or other adverse health outcomes from occupationally exposed populations as compared to the population exposed to environmental contaminants. Among those exposed to environmental contaminants are particularly susceptible subpopulations such as infants and children, the elderly, and those with preexisting disease, all of whom are usually not exposed to occupational pollutants. Workers in occupational settings

are generally considered healthier than the general population, and results of epidemiologic studies on the effects of workplace contaminants must account for this “healthy worker” effect before findings can be extended to the general population.

There are important differences in the scope of environmental *versus* occupational contaminants that pose a potential health risk to exposed populations. The general population is exposed to virtually thousands of pollutants and chemicals that potentially may pose a health risk, while the number of chemicals of concern in occupational settings is much more limited in scope. However, the levels of pollutant exposures and resulting health risks are typically orders of magnitude higher for occupational settings than in the general environment. Thus the health risk to any individual is typically much higher for exposure to occupational contaminants, while the population risk is generally higher from environmental pollutants owing to the large numbers of potentially exposed individuals.

Risk can be characterized at the individual and population levels. For individuals, either absolute or relative risk indicators may be applied. For populations, burden estimates are made that give the proportion of disease attributable to the environmental agent [7]. The most widely applied measure is the population attributable risk (PAR), which is described in equation (2):

$$PAR = \frac{P_E(RR - 1)}{1 + P_E(RR - 1)} \quad (2)$$

where P_E is the population prevalence of exposure and RR is the relative risk associated with exposure. The formula is useful for examining the interplay between patterns of exposure, risk to the exposed, and population risk. An exposure with very high relative risk as might occur for an occupational exposure has little impact on PAR if P_E is low. On the other hand, an exposure that is common, i.e., P_E is high, may contribute substantially to PAR, even if RR is low.

Environmental Health risks Case Study – The 1952 London Smog Episode

Perhaps one of the most dramatic examples of the public health consequences of environmental pollution occurred in London, England during December

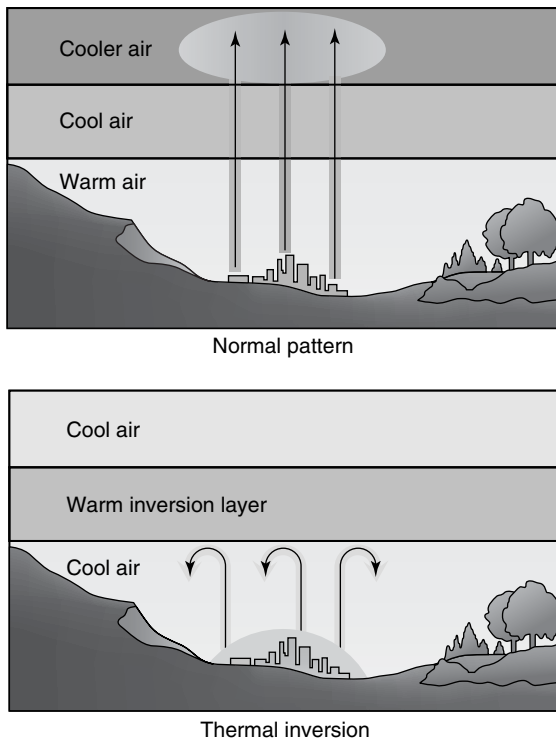


Figure 2 Schematic of inversion and normal atmospheric conditions

5–9, 1952. The combination of a thermal inversion (Figure 2) and the increased burning of coal for heating resulted in a dramatic increase in the levels of particulate matter, sulfur dioxide, and other combustion-related air pollutants during this 5-day period. The air pollution smoke and fog (thus the term *smog*) experienced in this episode reduced visibility to near zero in some areas, and increased air pollution levels by up to a factor of 10.

The health consequences of this air pollution episode mirrored the rise in air pollution. Widespread cattle deaths were the first indication that these high levels of pollution would have major human health consequences. The daily death rate in London tripled over the course of this event, rising from about 300 to a high of 900 deaths per day [8]. The populations at greatest risk from exposure to these high air pollution levels were the elderly and those with preexisting respiratory disease. For example, mortality for persons with bronchitis and pneumonia

increased sevenfold as a result of this episode. In total, an estimated 4000 deaths were attributed to the immediate effects of the episode, and up to another 8000 deaths in the first quarter of 1953 have been linked to the air pollution exposures during this episode [9].

While current levels of air pollution in developed countries are now many times lower than those in the 1930s–1960s, and air pollution episodes such as the 1952 London fog are no longer experienced in the United States, Europe, and other developed countries, relatively high air pollution levels are still experienced by populations living in developing countries experiencing rapid industrial growth. Additionally, recent evidence from scientific studies indicate that even relatively low levels of air pollution that are more typical of air quality experienced in the United States, Europe, and other developed countries pose a substantial health risk. In the United States, an estimated 60 000 premature deaths each year from cardiopulmonary disease and lung cancer have been associated with exposure to outdoor particulate matter pollution [10].

References

- [1] Last, J.M. (2001). *A Dictionary of Epidemiology*, 4th Edition, Oxford University Press, New York.
- [2] Prüss-Üstün, A. & Corvalán, C. (2006). *Preventing Disease through Healthy Environments: Towards an Estimate of the Environmental Burden of Disease*, World Health Organization, Geneva.
- [3] Rodricks, J.V. (1994). Risk assessment, the environment, and public health, *Environmental Health Perspectives* **102**, 258–264.
- [4] U.S. Environmental Protection Agency (2002). *Latest Findings on National Air Quality: 2001 Status and Trends*, Research Triangle Park.
- [5] Centers for Disease Control and Prevention. *National Occupational Exposure Survey (1981–1983): Agents Rank-Ordered by the Estimated Total Number of Employees Potentially Exposed to Each Agent*, available at <http://www.cdc.gov/noes/noes3/empl0001.html> (accessed Oct 2007).
- [6] Centers for Disease Control and Prevention. *National Occupational Exposure Survey (1981–1983): Estimated Percentages of Controlled and Uncontrolled Potential Exposures to Specific Agents by Occupation within 2-Digit Standard Industrial Classification (SIC)*, available at <http://www.cdc.gov/noes/noes5/46970ctr.html> (accessed Oct 2007).

- [7] Gordis, L. (2000). *Epidemiology*, 2nd Edition, W.B. Saunders, Philadelphia.
- [8] London Ministry of Health (1954). *Mortality and Morbidity during the London Fog of December 1952*, Reports on Public Health and Medical Subjects No.95, London.
- [9] Bell, M.L. & Davis, D.L. (2001). Reassessment of the lethal London fog of 1952: novel indicators of acute and chronic consequences of acute exposure to air pollution, *Environmental Health Perspectives* **109**, 389–394.
- [10] Shprentz, D.S., Bryner, G.C. & Shprentz, J.S. (1996). *Breath Taking: Premature Mortality due to Particle Air Pollution in 239 American Cities*, National Resources Defense Council (NRDC), New York.

JONATHAN M. SAMET AND RONALD H. WHITE

Environmental Security

Environmental Security Defined

Environmental security has emerged as an increasingly important concern of governments and their defense establishments due to several trends that have the potential to threaten stability [1, 2]. These potential threat issues include: population growth; water scarcities and allocation; groundwater depletion; globalization; significant degradation of arable land; loss of tropical forests; greenhouse gas emissions and climate change; exacerbation of biological species loss; rapid urbanization; increasingly serious urban air and water quality problems; and increasing frequency of hurricanes, earthquakes, floods, and other natural disasters [3, 4].

Several definitions of environmental security exist and demonstrate that after more than two decades of discussion the concept of environmental security still has no widely agreed upon formulation [5, 6]. Examples include the following:

“Environmental security (ecological security or a myriad of other terms) reflects the ability of a nation or a society to withstand environmental asset scarcity, environmental risks or adverse changes, or environment-related tensions or conflicts.” [7];

“Science-based case studies, which meld physical science with the discipline of political economy, are a suitable vehicle for forecasting future conflicts derived in some measure from environmental degradation” [8];

“[T]hose actions and policies that provide safety from environmental dangers caused by natural or human processes due to ignorance, accident, mismanagement or intentional design, and originating within or across national borders.” [9];

“Environmental Security is a state of the target group, either individual, collective or national, being systematically protected from environmental risks caused by inappropriate ecological process due to ignorance, accident, mismanagement or design.” [10];

“Environmental security is protectedness of natural environment and vital interests of citizens, society, the state from internal and external impacts, adverse processes and trends in development that threaten human health, biodiversity and sustainable functioning of ecosystems, and survival of humankind.” [10];

“Environmental security is the state of protection of vital interests of the individual, society, natural environment from threats resulting from anthropogenic and natural impacts on the environment.” [10]

The discipline of “environmental security” is neither a pure security issue nor an environmental issue [7]. However, environmental issues are often security concerns because, even without directly causing open conflict, they can result in environmental perturbations or triggers that can destabilize the status quo and result in a loss of regional, national, and local political, social, economic, and personal security [4].

Environmental security concerns can be grouped into three general categories [10]: (a) security of the environment, which is a good in itself, (b) security from environmental change that can create societal instability and conflict, and (c) security from environmental change (e.g., water scarcity, air pollution, etc.) that would threaten the material well-being of individuals [10]. Common elements of environmental security definitions include public safety from environmental dangers caused by natural or human processes due to ignorance, accident, mismanagement, or design; amelioration of natural resource scarcity; maintenance of a healthy environment; amelioration of environmental degradation; and, prevention of social disorder and conflict (promotion of social stability) [11].

Environmental Security and Risk Management

Application of risk assessment in the field of environmental security is a challenging task. There are many compelling reasons to use the current risk assessment/management paradigm for environmental security applications [6]. These include a variety of acceptable/unacceptable risk criteria at different regulatory levels or under specific statutes representing local control of risk decisions, existing law, experience with the system, and proven flexibility in the risk-management decision-making phase. The implicit acceptable risk-management process is an administratively and legally established process with known methods to determine whether acceptable risks have been exceeded. In such cases, it is important that the acceptable risk level is looked at as a point of departure for consideration, and not as a bright line.

Even though the risk assessment paradigm successfully used by many agencies may be generally applicable to environmental security, its application requires incorporating an uncertainty in basic knowledge that is much larger than the uncertainty for current applications. Tools that are currently used for uncertainty analysis (such as probabilistic modeling, predictive structure–activity analysis, etc.) may not be easily applied to environmental security. An additional challenge is the balancing of multiple relevant decision factors, including the environmental and societal benefits, as well as risks associated with the specific policy option.

Additional consideration could be incorporated into the commonly used risk assessment paradigm through the use of comparative risk assessment [12]. Comparative risk analysis (CRA, *see Comparative Risk Assessment*) has been most commonly applied within the realm of policy analysis. Programmatic CRA has helped to characterize regional and national environmental priorities by comparing the multidimensional risks associated with policy alternatives [13]. At smaller scales, CRAs often have specific objectives within the broader goal of evaluating and comparing possible alternatives and their risks [14]. CRA studies have compared interrelated risks involving specific policy choices, such as chemical versus microbial disease risks in drinking water. Central to CRA is the construction of a two-dimensional decision matrix that contains project alternatives' scores on various criteria. However, CRA lacks a structured method for combining performance on criteria to identify an optimal project alternative [15].

Nevertheless, adding limited number of additional factors and its qualitative evaluation may not be enough for environmental security applications. Risk management requires balancing scientific findings with multifaceted input from many stakeholders with different values and objectives. In such instances, systematic decision analysis tools [16] are an appropriate method to solve complex technical and behavioral issues [15]. Given the uncertainty in all aspects related to environmental security, the structured, transparent, and justifiable tools offered by multicriteria decision analysis (MCDA) for quantifying both scientific and decision makers' values and views, as well as for developing a system of performance metrics consistent with regulatory requirements, may be especially valuable for this emerging field. Even though MCDA tools may not be able to resolve

contradictions in current risk estimates, application of MCDA tools can help bring further enlightenment by making trade-offs explicit. The development and implementation of such a framework will make clear what information should be collected to support decisions. MCDA tools, coupled with the value of information analysis and adaptive management, could provide a good foundation for both bringing together multiple information sources to assess environmental security risks and also for developing justifiable and transparent regulatory decisions [17, 18].

Framework for Incorporating Risk Assessment and Decision Analysis into Environmental Security Management

A systematic decision framework (Figure 1, from [15]) could be used as a generalized road map to environmental security decision-making process, including emerging threats [19]. Having the right combination of people is the first essential element in the decision process. The activity and involvement levels of three basic groups of people (environmental managers and decision makers, scientists and engineers, and stakeholders) are symbolized in Figure 1 by dark lines for direct involvement and dotted lines for less direct involvement. While the actual membership and function of these three groups may overlap or vary, the roles of each are essential in maximizing the utility of human input into the decision process. Each group has its own way of viewing the world, its own method of envisioning solutions, and its own societal responsibility. Policy and decision makers spend most of their effort defining the problem's context and the overall constraints of the decision. In addition, they may have responsibility for the final policy selection and implementation. Stakeholders may provide input in defining the problem, but they contribute the most input toward helping formulate a performance criteria and making value judgments for weighting the various success criteria. Depending on the problem and regulatory context, stakeholders may have some responsibility in ranking and selecting the final option. Scientists and engineers have the most focused role in that they provide the measurements or estimations of the desired criteria that determine the success of various alternatives. While they may take a secondary role as stakeholders or decision makers, their primary role is to provide the technical input necessary for the decision process.

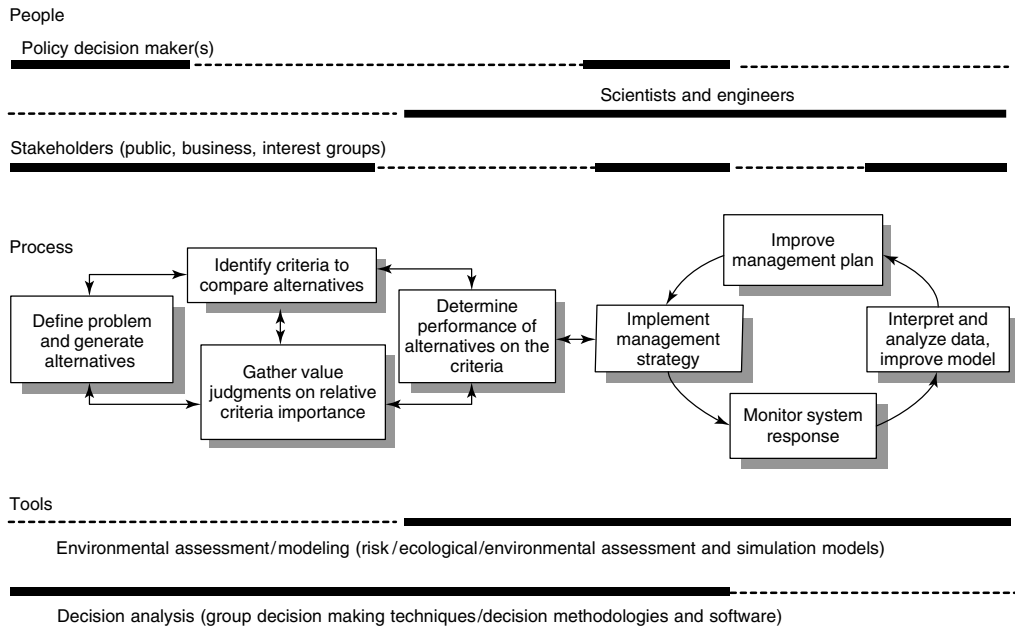


Figure 1 Adaptive decision framework. Solid lines represent direct involvement for people or utilization of tools; dashed lines represent less direct involvement or utilization [Reproduced from [15]. © Pergamon Press, 2006.]

The process depicted in Figure 1 follows two basic themes: (a) generating management alternatives, success criteria, and value judgments, and (b) ranking the alternatives by applying value weights. The first part of the process generates and defines choices, performance levels, and preferences. The latter section methodically prunes nonfeasible alternatives by first applying screening mechanisms (for example, overall cost, technical feasibility, or general societal acceptance) and then ranking in detail the remaining options by decision analytical techniques (Analytical Hierarchy Process (AHP), Multi-attribute Utility Theory (MAUT), or outranking) that use the various criteria levels generated by environmental tools, monitoring, or stakeholder surveys. While it is reasonable to expect that the process may vary in specific details among regulatory programs and project types, emphasis should be given to designing an adaptive management structure. In contrast to traditional “command and control” approaches [17], adaptive management uses adaptive learning as a means for incorporating changes in decision priorities, and new knowledge from monitoring programs into strategy selection.

The tools used within group decision making and scientific research are essential elements of the

overall decision process. As with group involvement, the applicability of the tools is symbolized in Figure 1 by solid lines (direct or high utility) and dotted lines (indirect or low utility). Decision analysis tools help to generate and map stakeholder preferences as well as individual value judgments into organized structures that can be linked with other technical tools from risk analysis, modeling, monitoring, and cost estimations. Decision analysis software can also provide useful graphical techniques and visualization methods to express the gathered information in understandable formats. Linkov *et al.* [18] and Yatsalo *et al.* [20] present an example of MCDA assessments for two sediment management case studies utilizing four different MCDA software packages. Figure 2 presents just one example output. Criterium Decision Plus software [21] was used to implement a Simple Multi-attribute Rating Technique (SMART)/MAUT approach to quantitatively incorporate stakeholder value judgments on criteria along with technical measures for specific criteria weighting. Figure 2 shows an example of ranking of four sediment management alternatives. The wetland restoration alternative is the most preferable because of its environmental and ecological benefits. An upland disposal is ranked

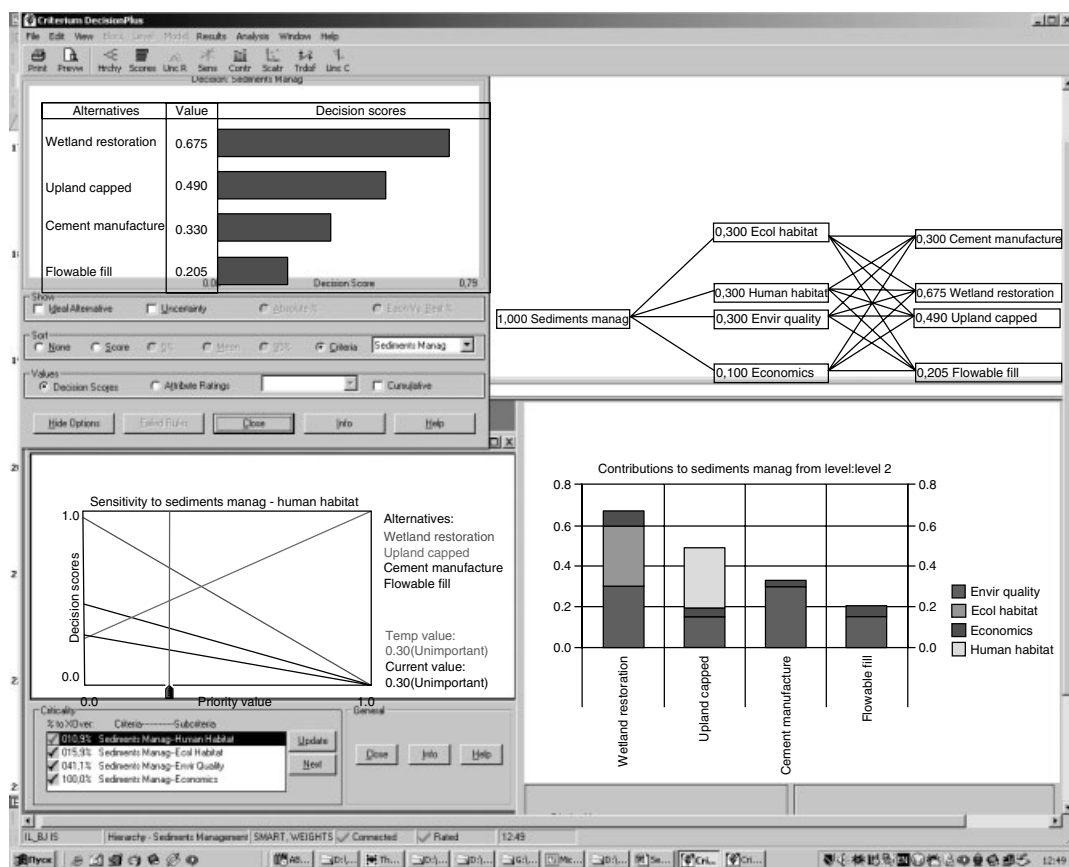


Figure 2 Example of MCDA software output. Ranking of sediment management alternatives and sensitivity analysis using Criterion Decision Plus software [Reproduced from [22]. © Springer Publishing, 2007.]

second due to its relatively low cost and minimal human health risks.

When changes occur in the requirements or the decision process, decision analysis tools can respond efficiently to reprocess and iterate with the new inputs. The framework depicted in Figure 1 provides a focused role for the detailed scientific and engineering efforts invested in experimentation, environmental monitoring, and modeling that provides rigorous and defensible details for evaluating criteria of performances under various alternatives. This integration of decision tools and scientific and engineering tools allows each to have a unique and valuable role in the decision process without attempting to apply either type of tool beyond its intended scope.

The result of the entire process is a comprehensive, structured process for selecting the optimal

alternative in any given situation, drawing from stakeholder preferences and value judgments, as well as scientific modeling and risk analysis. This structured process would be of great benefit to decision making in environmental security management, where there is currently no structured approach for making justifiable and transparent decisions with explicit trade-offs between social and technical factors. The MCDA framework links heterogeneous information on causes, effects, and risks for different environmental policy options with decision criteria and weightings elicited from decision makers, allowing visualization and quantification of the trade-offs involved in the decision-making process. The proposed framework can also be used to prioritize research and information gathering activities and thus can be useful for the value of information analysis.

References

- [1] Morel, B. & Linkov, I. (2006). *Environmental Security: The Role of Risk Assessment*, Springer, Amsterdam, p. 326.
- [2] Linkov, I., Kiker, G. & Wenning, R. (2007). *Environmental Security in Harbors and Coastal Areas: Management Using Comparative Risk Assessment and Multi-Criteria Decision Analysis*, Springer, Amsterdam.
- [3] National Intelligence Council (2000). *Global Trends 2015: A Dialogue about the Future with Nongovernment Experts*. December 2000, http://www.odci.gov/nic/PDF_GIF_global/globaltrend2015.pdf.
- [4] Schwartz, P. & Randall, D. (2003). *Abrupt Climate Change Scenario and Its Implications for United States National Security (October 2003)*. GBN: Global Business Network, <http://www.gbn.com/ArticleDisplayServlet.srv?aid=26231>.
- [5] Belluck, D.A., Hull, R.N., Benjamin, S.L., Alcorn, J. & Linkov, I. (2006). Environmental security, critical infrastructure and risk assessment: definitions and current trends, in *Environmental Security and Environmental Management: The Role of Risk Assessment*, B. Morel & I. Linkov, eds, Springer, Amsterdam.
- [6] Belluck, D.A., Hull, R.N., Benjamin, S.L., Alcorn, J. & Linkov, I. (2006). Are standard risk acceptability criteria applicable to critical infrastructure based on environmental security needs, in *Environmental Security and Environmental Management: The Role of Risk Assessment*, B. Morel & I. Linkov, eds, Springer, Amsterdam.
- [7] Chalecki, E.L. (2006). *Environmental Security: A Case Study of Climate Change*. Pacific Institute for Studies in Development, Environment, and Security, http://www.pacinst.org/environment_and_security/env_security_and_climate_change.pdf.
- [8] McNeil, F. (2000). *Making Sense of Environmental Security*. North-South Agenda. Paper. Thirty-Nine, <http://www.miami.edu/nsc/publications/pub-ap-pdf/39AP.pdf>.
- [9] Cheremisnoff, N.P. (2002). *Environmental Security: The Need for International Policies*. Pollution Engineering, http://www.pollutionengineering.com/CDA/Article-Information/features/BNP.Features_Item/0,6649,103962,00.html (accessed May 2002).
- [10] AC/UNU Millennium Project (2006). *Emerging International Definitions, Perceptions, and Policy Considerations*. *Environmental Security Study*, <http://www.acunu.org/millennium/es-2def.html>.
- [11] Glenn, J.C., Gordon, T.J. & Perelet, R. (1998). In *Defining Environmental Security: Implications for the U.S. Army*, Molly Landholm, ed, Army Environmental Policy Institute, AEPI-IFP-1298.
- [12] Linkov, I. & Ramadan, A. (2004). *Comparative Risk Assessment and Environmental Decision Making*, Kluwer, Amsterdam, p. 436.
- [13] Andrews, C.J., Apul, D.S. & Linkov, I. (2004). Comparative risk assessment: past experience, current trends and future directions, in *Comparative Risk Assessment and Environmental Decision Making*, I. Linkov & A. Ramadan, eds, Kluwer, Amsterdam.
- [14] Bridges, T., Kiker, G., Cura, J., Apul, D. & Linkov, I. (2005). Towards using comparative risk assessment to manage contaminated sediments, in *Strategic Management of Marine Ecosystems*, E. Levner, I. Linkov & J.M. Proth, eds, Springer, Amsterdam.
- [15] Linkov, I., Satterstrom, K., Kiker, G., Batchelor, C. & Bridges, T. (2006). From comparative risk assessment to multi-criteria decision analysis and adaptive management: recent developments and applications, *Environment International* **32**, 1072–1093.
- [16] Figueira, J., Greco, S. & Ehrgott, M. (2005). *Multiple Criteria Decision Analysis: State of the Art Surveys*, Springer, New York.
- [17] Linkov, I., Satterstrom, K., Seager, T.P., Kiker, G., Bridges, T., Belluck, D. & Meyer, A. (2006). Multi-criteria decision analysis: comprehensive decision analysis tool for risk management of contaminated sediments, *Risk Analysis* **26**, 61–78.
- [18] Linkov, I., Satterstrom, K., Kiker, G., Bridges, T., Benjamin, S. & Belluck, D. (2006). From optimization to adaptation: shifting paradigms in environmental management and their application to remedial decisions, *Integrated Environmental Assessment and Management* **2**, 92–98.
- [19] Linkov, I., Satterstrom, K., Steevens, J., Ferguson, E. & Pleus, R. (2007). Multi-criteria decision analysis and nanotechnology, *Journal of Nanoparticle Research* **9**, 543–554.
- [20] Yatsalo, B., Kiker, G., Kim, J., Bridges, T., Seager, T., Gardner, K., Satterstrom, K. & Linkov, I. (2007). Application of multi-criteria decision analysis tools for management of contaminated sediments, *Integrated Environmental Assessment and Management* in press.
- [21] InfoHarvest (2004). *Criterion Decision Plus Software Version 3.0*, <http://www.infoharvest.com>.
- [22] Linkov, I., Satterstrom, F.K., Yatsalo, B., Tkachuk, A., Kiker, G., Kim, J., Bridges, T.S., Seager, T.P., Gardner, K. (2007). Comparative assessment of several multi-criteria decision analysis tools for management of contaminated sediments, in *Environmental Security in Harbors and Coastal Areas: Management Using Comparative Risk Assessment and Multi-Criteria Decision Analysis*, I. Linkov, G. Kiker & R. Wenning, eds, Springer.

Related Articles

Environmental Monitoring

Environmental Risk Regulation

Statistics for Environmental Justice

Environmental Tobacco Smoke

Background

Exposure to environmental tobacco smoke (ETS), also referred to as passive smoking, involuntary smoking, or secondhand smoke, has long been considered harmful. The US Surgeon General's report in 1972 [1] was one of the first governmental reports to recognize ETS as a health effect. The authors noted that animals exposed to tobacco smoke in laboratory experiments had a higher risk of respiratory cancers and there was evidence to suggest that acute respiratory illnesses were more common in children with parents who smoked.

Biological Plausibility

Approximately 85% of ambient tobacco smoke comes from the burning end of a lit cigarette between puffs (sidestream smoke). The rest is largely that exhaled by the smoker (mainstream smoke). Cigarette smoke contains over 4000 chemicals and exposure to ETS involves breathing in the same chemicals as active smoking but at a lower concentration. Active smoking is a cause of many fatal disorders including cancer, cardiovascular disease, and respiratory disease, as well as reproductive effects and chronic disorders such as cataracts, stomach ulcers, and hip fractures [2, 3].

Because of the dose–response relationship (*see Dose–Response Analysis*) between active smoking and each disorder, it is entirely plausible that a relatively little exposure (i.e., from ETS) will be associated with some risk of developing the same disorders. Examining markers of tobacco smoke intake in the urine, saliva, and blood of passive smokers supports this hypothesis. Nicotine is a substance that is, for practical purposes, only found in the environment in tobacco smoke. Nicotine and cotinine (a by-product of nicotine) are, therefore, useful markers of tobacco smoke inhalation and are sensitive enough to detect low concentrations. Studies show that these markers are two to three times greater in nonsmokers who report exposure to ETS compared to those who report no exposure [4], and that the concentration

increases with increasing exposure (Figure 1). Studies also show that cotinine and nicotine levels in passive smokers are, on average, about 1% of those seen in active smokers [4]. This is a nonnegligible level of exposure, which is expected to be associated with a nonnegligible level of risk. These biochemical studies confirm that tobacco smoke is inhaled and metabolized by nonsmokers in the same way as active smokers.

Health Effects in Adults

The first epidemiological studies, reported in 1981, focused on lung cancer [6, 7]. The two studies were based on lifelong nonsmokers and compared the incident rates of lung cancer in people married to smokers to those married to nonsmokers. Both showed an increased risk of lung cancer. Spousal exposure is a good measure of ETS because it is easily defined, validated (people married to smokers have higher concentrations of nicotine and cotinine), and it is a relatively stable measure since people tend to remain married for several years.

Many studies of ETS showed an increase in risk for various disorders, but most were not large enough to detect statistically significant associations on their own. Meta-analyses (*see Meta-Analysis in Nonclinical Risk Assessment; Meta-Analysis in Clinical Risk Assessment*) are a useful way of combining the quantitative information from all relevant studies to provide a more precise estimate of risk than any study on its own. A meta-analysis of 46 studies of ETS and lung cancer (Table 1) yields a 24% increase in risk. Adjustment for bias and potential confounding does not have a material effect on the estimate [8]. Other studies have looked at coronary (or ischemic) heart disease, chronic obstructive pulmonary disease (COPD), and stroke (Table 1). They too show a 25–30% excess risk. Although the main public health focus of ETS has been on lung cancer, the effect on the number of deaths from heart disease is greater because it is much more common than lung cancer in nonsmokers.

In the United States, in 2005 it was estimated that ETS caused 3400 lung cancer deaths among nonsmokers and 46 000 deaths from coronary heart disease – equivalent to about 5 deaths every hour [12]. In 24 countries of the European Union, about 19 240 nonsmokers were estimated to have died

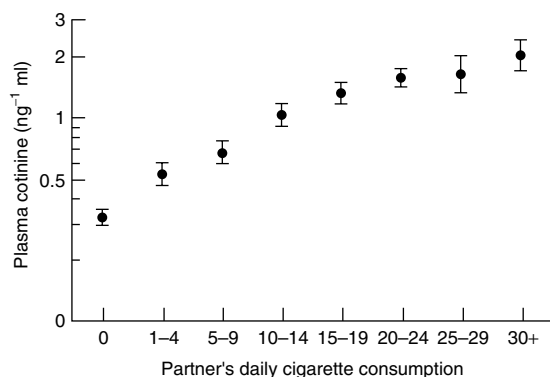


Figure 1 Mean plasma cotinine (geometric mean and 95% confidence interval) in 8170 married or cohabiting nonsmokers according to the cigarette consumption of their partner [Reproduced from [5]. © BMJ Publishing Group, 2001.]

in 2002 from four disorders: lung cancer (1550), coronary heart disease (10 240), COPD (1170), and stroke (6280). This is equivalent to about two deaths every hour [13]. It should, however, be noted that these estimates from the United States and Europe are approximate and based on various assumptions associated with the prevalence of passive smoke exposure, the risk of developing each disorder, which were then used to apportion the total number of deaths in the population among nonsmokers and active smokers.

Although most of the evidence has been based on people who have never smoked, the effect of ETS on former smokers should not be ignored. The risk of death or chronic disease in smokers who quit rarely becomes the same as that of a never-smoker, and therefore the risk due to ETS when added to the risk associated with past smoking habits is an important consideration for ex-smokers. For example, the age-standardized rate of dying from lung cancer is 0.68

per 1000 among male ex-smokers and 0.17 per 1000 in never-smokers – a relative risk of 4 (0.68/0.17) [14]. Assuming ETS increases an ex-smoker's risk by 24% (as it does in never-smokers) the resulting death rate is 0.84 per 1000 – a relative risk of 5.

Health Effects in Children

A range of disorders in children and infants are caused by exposure to ETS. The main source of exposure investigated in studies has been parental smoking. Table 2 shows the odds ratio of having certain respiratory diseases or sudden infant death [2, 15]. The risk of a child suffering from lower respiratory illness (LRI) is 59% greater if either parent smoked compared to having two nonsmoking parents (odds ratio of 1.59). Other respiratory effects caused by ETS include middle ear effusion, asthma, cough, and wheeze. An even stronger association was found with sudden infant death, where infants were twice as likely to die early if their mother smoked. Consideration of confounding factors had little effect on the estimates of excess risk. In Table 2, the increase in risk tends to be higher with maternal smoking than paternal smoking reflecting the longer time spent with the mother during childhood, further supporting a causal link. The adverse effects on children are often overlooked when considering the public health impact of ETS, yet there is a clear effect on the health and quality of life during childhood if exposed to tobacco smoke at that time.

The Case for Banning Smoking in Public Places

Several countries already have some form of policy on restricting smoking in various environments,

Table 1 Health effects in adults caused by ETS (summary results from published meta-analyses)

Disorder	Number of studies in the meta-analysis	Total number of people with the disorder in the meta-analysis	Percentage increase in risk (95% CI) due to ETS
Lung cancer [9]	46	6257	24 (14–34)
Coronary heart disease [10]	19	6600	30 (22–38)
COPD [11]	8	1171	25 (10–43)
Stroke [4]	3	1147	27 (10–46)

CI, confidence interval

Table 2 Health effects in children caused by ETS (summary results from published meta-analyses)^(a)

Disorder	From meta-analysis	Either parent smoked	Mother smoked	Father smoked
Lower respiratory illness (LRI)	Number of studies	38	41	18
	Odds ratio (95% CI)	1.59 (1.47–1.73)	1.72 (1.59–1.86)	1.31 (1.19–1.43)
Hospitalization for LRI, bronchitis, or pneumonia	Number of studies	9	11	7
	Odds ratio (95% CI)	1.73 (1.31–2.28)	1.49 (1.29–1.73)	1.31 (0.98–1.76)
Recurrent otitis media	Number of studies	9	5	3
	Odds ratio (95% CI)	1.37 (1.10–1.70)	1.37 (1.19–1.59)	0.90 (0.70–1.15)
Middle ear effusion	Number of studies	6	–	–
	Odds ratio (95% CI)	1.33 (1.12–1.58)	–	–
Asthma	Number of studies	31	21	12
	Odds ratio (95% CI)	1.23 (1.14–1.33)	1.33 (1.24–1.43)	1.07 (0.97–1.18)
Wheeze	Number of studies	45	27	14
	Odds ratio (95% CI)	1.26 (1.20–1.33)	1.28 (1.21–1.35)	1.13 (1.08–1.20)
Cough	Number of studies	39	16	10
	Odds ratio (95% CI)	1.35 (1.27–1.43)	1.34 (1.17–1.54)	1.22 (1.12–1.32)
Sudden infant death	Number of studies	–	18	–
	Odds ratio (95% CI)	–	2.13 (1.86–2.43)	–

^(a) Based on meta-analyses first published in reference [15] and updated in reference [2]

CI, confidence interval

including the United Kingdom, California, New York, Ireland, Australia, and Italy. Many workplaces either do not allow it at all or have separate smoking rooms. In enclosed public places, while it is important to reduce or eliminate exposure to ETS among customers, the main justification behind a ban is the protection of workers. People who work in bars, clubs, and restaurants are often exposed to high levels of tobacco smoke for longer periods of time. Restrictions on smoking in these workplaces should, therefore, be in line with levels of protection in other areas of employment. The excess risk associated with ETS exposure in the hospitality industry is completely avoidable.

References

- [1] US Department of Health and Human Services (1972). *The Health Consequences of Smoking: A Report of the Surgeon General*, DHEW Publication No. (HSM) 72–7516, Public Health Service, Office of the Surgeon General.
- [2] US Department of Health and Human Services (2006). *The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General*, Centers for Disease Control and Prevention, Office on Smoking and Health.
- [3] Vineis, P., Alavanja, M., Buffler, P., Fontham, E., Franceschi, S., Gao, Y.T., Gupta, P.C., Hackshaw, A., Matos, E., Samet, J., Sitas, F., Smith, J., Stayner, L., Straif, K., Thun, M.J., Wichmann, H.E., Wu, A.H., Zaridze, Z., Peto, R. & Doll, R. (2004). Tobacco and cancer: recent epidemiological evidence, *Journal of the National Cancer Institute* **96**(2), 99–106.
- [4] Britton, J. & Edwards, R. (eds) (2005). The health effects of environmental tobacco smoke, in *Environmental Tobacco Smoke. A Report of the Tobacco Advisory Group of the Royal College of Physicians*, Royal College of Physicians.
- [5] Jarvis, M.J., Feyerabend, C., Bryant, A., Hedges, B. & Primatesta, P. (2001). Passive smoking in the home: plasma cotinine concentrations in non-smokers with smoking partners, *Tobacco Control* **10**, 368–374.
- [6] Garfinkel, L. (1981). Time trends in lung cancer mortality among nonsmokers and a note on passive smoking, *Journal of the National Cancer Institute* **66**, 1061–1066.
- [7] Hiriyama, T. (1981). Nonsmoking wives of heavy smokers have a higher risk of lung cancer: a study from Japan, *British Medical Journal* **282**, 183–185.
- [8] Hackshaw, A.K., Law, M.R. & Wald, N.J. (1997). The accumulated evidence on lung cancer and environmental tobacco smoke, *British Medical Journal* **315**, 980–988.

- [9] IARC (2004). *Monographs on the Evaluation of Carcinogenic Risks to Humans, Volume 83, Tobacco Smoke and Involuntary Smoking*, WHO/IARC, Lyon.
- [10] Law, M.R., Morris, J.K. & Wald, N.J. (1997). Environmental tobacco smoke and ischaemic heart disease: an evaluation of the evidence, *British Medical Journal* **315**, 973–980.
- [11] Law, M.R. & Hackshaw, A.K. (1996). Environmental tobacco smoke, *British Medical Bulletin* **52**(1), 22–34.
- [12] California Environmental Protection Agency (2005). *Proposed Identification of Environmental Tobacco Smoke as a Toxic Air Contaminant. Part B: Health Effects*, Office of Environmental Health Hazard Assessment, Sacramento.
- [13] Jamrozik, K. (2006). An estimate of deaths attributable to passive smoking in Europe, in *Lifting the Smoke-screen: 10 Reasons for A Smoke Free Europe*, ERSJ Ltd (<http://www.ersnet.org>).
- [14] Doll, R., Peto, R., Boreham, J. & Sutherland, I. (2004). Mortality in relation to smoking: 50 years' observation on male British doctors, *British Medical Journal* **328**, 1519–1527.
- [15] Cook, D.G. & Strachan, D.P. (1999). Health effects of passive smoking: 10, *Thorax* **54**, 357–366.

Related Articles

Air Pollution Risk

Risk and the Media

Stakeholder Participation in Risk Management Decision Making

ALLAN HACKSHAW

Epidemiology as Legal Evidence

In tort cases concerned with diseases resulting from exposure to a toxic chemical or drug, epidemiologic studies are used to assist courts in determining whether the disease of a particular person, typically the plaintiff, was a result of his or her exposure. This may seem puzzling to scientists, because whenever there is a natural or background rate of an illness one cannot be certain that its manifestation in a specific individual who was exposed to a toxic agent actually arose from that exposure. Indeed, the probability of causation in a specific individual is nonidentifiable [1]. The standard of proof that courts utilize in civil cases, however, is the preponderance of the evidence or the “more likely than not” criterion. Thus, scientific evidence that a particular agent can cause a specific disease or set of related diseases in the general population supports an individual’s claim that his or her disease came from their exposure. Conversely, scientific studies indicating no increased risk of a specific disease amongst exposed individuals are relied on by defendants, typically producers of the chemical or drug, to support the safety of their product. Studies relied on by experts testifying in cases are scrutinized intensely by opposing counsel for possible biases due to omitted covariates and measurement error or possible failure to consider the latency period before exposure and manifestation of the disease or inadequate power. Thus, it is important for scientists to be thoroughly prepared to discuss the relevant peer-reviewed literature before testifying. Similar questions of causation (*see* **Causality/Causation**) arise in cases alleging harm from exposure to hazardous wastes; although the issue in these cases is often whether the exposure was sufficient in magnitude and duration to cause the disease (*see* **Environmental Health Risk; Environmental Hazard**).

Epidemiologic studies are also used to determine eligibility for workers’ compensation, where the issue is whether the employee’s disease arose from exposure to an agent in the course of employment [2, p. 831], in regulatory hearings to determine safe exposure levels in the workplace (*see* **Occupational Cohort Studies**), and have even been submitted as evidence in criminal cases [3, p. 153]. We emphasize scientific evidence in tort law, which includes product

liability and mass chemical exposure cases, because it is the major area of the law utilizing epidemiologic studies as evidence.

Tort Law

Tort law generally concerns suits for wrongful injury that do not arise from a contract between the parties. Thus, remedies to compensate for injuries from a wide variety of accidents resulting from someone’s negligence, e.g., professional malpractice, assault and battery, environmentally induced injury, and fraud can be obtained by a successful plaintiff. Product liability is a special area of tort law dealing with the obligations of manufacturers of products to consumers who may suffer personal injury arising from the use of the product.

In any tort claim the plaintiff needs to establish a *prima facie* case by showing that the defendant has a legal duty of care due to the plaintiff and that the defendant breached that duty. In addition, a plaintiff needs to show that (a) he/she suffered an injury and that the defendant’s failure to fulfill its duty of care was the (b) factual and (c) legal cause of the injury in question. The law also recognizes defenses that relieve the defendant of liability. The two most prominent ones in tort suits are contributory negligence by the plaintiff and statutes of limitations, which bar suits that are brought after a specified period of time has elapsed from either the time of the injury or the time when the relationship between the injury and the use of the product was known to the plaintiff [4]. In some jurisdictions, especially in Europe [5, p. 834], if the injury results from a defect arising from the product’s compliance with a mandatory legal provision at the time it was put on the market, then the manufacturer is not liable. There are substantial differences between jurisdictions as to whether a plaintiff’s contributory negligence totally absolves the defendant from liability, reduces it in proportion to the relative fault of the parties, or has no effect on the liability of a defendant whose contribution to the injury was small. In the United States, the plaintiff’s fault is rarely a complete bar to recovery when the defendant’s negligence had a significant role. Similarly, the effective starting date of the limitations period varies among nations and among the states in the United States.

When reading actual legal cases that rely on scientific evidence, one needs to be aware of the relevant

legal rules. For example, although the epidemiologic evidence linking the appearance of a rare form of vaginal cancer in a young woman to her mother's use of diethylstilbestrol during pregnancy is quite strong [6], some states barred plaintiffs from suing because the statute of limitations had expired. Since the cancers were recognized only when the young women passed puberty, typically in the late teens or early twenties, a number of injured women could not receive compensation. Other states, however, interpreted the limitations period as beginning at the time the plaintiff should have been aware of the connection. In Europe, the European Economic Community (EEC) directive of 1985 provides for a 10-year statute of limitations and allows plaintiffs to file claims within 3 years after *discovering* the relationship. Markesinis [5] summarizes the directive and the relevant English and German laws.

In fact epidemiologic evidence is most useful in resolving the issue of cause, i.e., whether exposure to the product made by the manufacturer or chemicals spilled onto one's land by a nearby company can cause the injury suffered by the plaintiff. An alternate formulation of the factual cause issue is whether exposure increases the probability of contracting the disease in question to an appreciable degree. Case-control studies (*see Case-Control Studies*) were used for this purpose in the litigation surrounding Rely and other highly absorbent tampons [2, p. 840]. Within a year or two after these products were introduced, the incidence of toxic shock syndrome (TSS) among women who were menstruating at the time of the illness began to rise sharply. Several studies, cited in Gastwirth [2, p. 918], indicated that the estimated relative risk (*see Relative Risk*) of contracting the disease for users of these tampons was at least 10, which was statistically significant.

In light of the sharp decline in the incidence of TSS after the major brand, Rely, was taken off the market, the causal relationship seems well established and plaintiffs successfully used the studies to establish that their disease was most likely a result of using the product. When only one case-control study, however, indicates an association between exposure and a disease, courts are less receptive. Inskip [7] describes the problems that arose in a British case concerning radiation exposure of workers and leukemia in their children, and Fienberg and Kaye [8] discuss general issues concerning the information provided by clusters of cases.

There is a rough rule relating the magnitude of the relative risk, R , of a disease related to exposure and the legal standard of preponderance of the evidence, i.e., at least half of the cases occurring amongst individuals exposed to the product in question should be attributable to exposure. As the attributable risk is $(R - 1)/R$, this is equivalent to requiring a relative risk of at least 2.0. While a substantial literature discusses this requirement (see [9, pp. 1050–1054], [2, Chapters 13 and 14], [10], and [11, pp. 167–170] for discussion and references), some courts have been reluctant to adopt it formally [12], since it would allow the public to be exposed to agents with relative risks just below 2.0 without recourse. The author found the lowest value of R accepted by a court to be 1.5, in a case concerning the health effects of asbestos exposure. Courts usually require that the estimated R be statistically significantly greater than 1.0 and have required a confidence interval for R but also consider the role of other error rates [11, pp. 153–154]. When a decision must be based on sparse evidence, courts implicitly consider the power of a test and may not strictly adhere to significance at the 0.05 level.

The relative risk estimated from typical case-control studies is taken as an average for the overall population. Courts also consider the special circumstances of individual cases and have combined knowledge of the prior health of a plaintiff, the time sequence of the relevant events, the time and duration of exposure, as well as the latency period of the disease, with epidemiologic evidence to decide whether or not exposure was the legal cause of a particular plaintiff's disease. The different verdicts concerning the manufacturer in the cases of Vioxx, a Cox-2 pain reliever shown to increase the risk of cardiovascular problems, may be due to the substantial variability in the other risk factors and general health of the plaintiffs.

So far, our discussion has dealt with the criteria for factual causality where an injury has already occurred. In some cases concerning exposure to a toxic chemical, plaintiffs have asked for medical monitoring, such as periodic individual exams or a follow-up study. As this is a new development, a specific minimal value of R has not been established. Recently, the Supreme Court of Missouri in *Meyer v. Fluor Corp.* No. SC8771 (2007) allowed a class action on behalf of children exposed to lead from a smelting facility to proceed. The opinion

sided with jurisdictions that have held that a plaintiff can obtain damages for medical monitoring when they have significantly increased risk of contracting a particular disease as a consequence of their exposure.

In product liability law, a subclass of tort, in addition to negligence claims, sometimes one can assert that the manufacturer is subject to strict liability [13, 14]. In strict liability the test is whether the product is unreasonably dangerous, not whether the manufacturer exercised appropriate care in producing the product. Epidemiologic studies indicating a substantial increased risk of a disease can be used to demonstrate that the product is “unreasonably dangerous” from the viewpoint of the consumer.

An increasing number of product liability cases concerns the manufacturer’s duty to warn of dangers that were either known to the manufacturer or could reasonably have been foreseen at the time the product was marketed. In the United States, producers are also expected to keep abreast of developments after the product is sold and to issue a warning and possibly recall the product if postmarketing studies show an increased risk of serious disease or injury.

One rationale underlying the duty to warn is informed consent [15, p. 209] (*see Randomized Controlled Trials*). Because asbestos was linked to lung cancer by a major study [16] published in the 1960s, the plaintiff in *Borel v. Fibreboard Paper Products Corp.*, 493 F. 2d 1076 (5th Cir. 1973) prevailed on his warning claim. The opinion observed that a duty to warn arises whenever a reasonable person would want to be informed of the risk in order to decide whether to be exposed to it.

The time when the risk is known or knowable to the manufacturer is relevant. In *Young v. Key Pharmaceuticals*, 922 P. 2d 59 (Wash. 1996), the plaintiff alleged that the defendant should have warned that its asthma drug increased the risk of a seizure. The firm argued that the studies existing in 1979, when the child was injured, were not clinically reliable. Even though subsequent research confirmed those early studies that suggested an increased risk, the court found that the defendant did not have a duty to warn in 1979.

The reverse situation may have occurred in the *Wells* case, 788 F. 2d 741 (11th Cir. 1986). At the time the mother of the plaintiff used the spermicide made by the defendant, two studies had shown

an increased risk of limb defects and the court found the firm liable for failing to warn. Subsequent studies, which still may not be definitive, did not confirm the earlier ones, and in a later case the defendant was found not to be liable. While this seems inconsistent from a scientific point of view, from a legal perspective both decisions may be reasonable because the information available at the two times differed. A review of the studies and reports discussed in the *Wells* and *Young* decisions is given in Gastwirth [17] where an ethical issue in conducting follow-up studies of the relationship of spermicide use while pregnant and birth defects is noted. Once a woman is pregnant, asking her to use any product that might harm the child is unethical, so one cannot conduct a controlled clinical trial, randomly assigning females to continue to use or not use a spermicide.

The legal issues involved in warning cases are quite involved and differ among the states in the United States. Some states have a learned intermediary rule, which absolves the manufacturer of liability if it warns the medical community and the treating doctor would not have changed his/her decision to prescribe the drug. The decision in *Madsen v. American Home Products Corp.* (E.D. Mo. No. 02–1835, 2007) noted that although the firm knew about the risk it did not adequately warn consumers or the medical community about the risk of heart valve disease associated with two diet drugs as reported in studies in 1993 and case reports in 1994–1996 showed the firm knew about the risk. However, the plaintiff needed to show that the warning would have affected either the doctor’s prescription or her decision to take them. As the doctor continued to prescribe the drug after the medical community was notified of the risk in mid-1997 and as the plaintiff never read any material about the drugs, an earlier warning would not have prevented her illness. In *Petty v. U.S.*, 740 F.2d 1428 (1984), which concerned an illness contracted in the context of a mass immunization where there was no learned intermediary, the producer, however, had the responsibility to warn the ultimate consumer.

The increased media advertising of drugs directly to consumers has generated complaints of misleading advertising, especially when the drug is marketed for an off-label use, i.e., for treating a different ailment than it was approved for by the authorities. In 1996, Zyprexa was approved for treating

schizophrenia and bipolar disorder and the producer apparently informed the psychiatric community of risks associated with its use. Recently, several other states including Montana have filed suits accusing of the manufacturer of misleading advertising as it marketed the drug for treating other problems, e.g., anxiety and depression rather than psychotic conditions it was approved for without adequately informing patients or their primary care physicians of known increased risks for strokes, heart attacks, and pancreatic disorders.

Government Regulation

Epidemiologic studies are used by regulatory agencies such as the Food and Drug Administration (FDA) and Occupational Safety and Health Administration (OSHA) to get manufacturers to recall harmful products or give an appropriate warning. Indeed, the manufacturer of Rely tampons recalled the product after the fourth case-control study linked it to TSS. More recently, case-control studies supported a warning campaign.

In 1982, after a fourth study indicated an association between aspirin use and Reye's syndrome, the FDA proposed a warning label on aspirin containers. The industry challenged the original studies, and the Office of Management and Budget (OMB) asked the FDA [18, 19] to wait for another study. The industry suggested that caretakers of cases would be under stress and might guess aspirin, especially if they had heard of an association, so two new control groups (children hospitalized for other reasons and children who went to an emergency room) were included in the follow-up study [20]. The odds ratios (OR) (*see Odds and Odds Ratio*) for cases compared with each of these two control groups were about 50, far exceeding those of the school ($OR = 9.5$) and neighborhood controls ($OR = 12.6$).

In late 1984 the government, aware of these results, asked for a voluntary warning campaign; a warning was mandatory as of June 1986. The following are the Reye's syndrome cases and fatalities from 1978 to 1989: 1978 (236, 68); 1979 (389, 124); 1980 (555, 128); 1981 (297, 89); 1982 (213, 75); 1983 (198, 61); 1984 (204, 53); 1985 (93, 29); 1986 (101, 27); 1987 (36, 10); 1988 (25, 11); and 1989 (25, 11). The cases are graphed in Figure 1. Notice the sharp decline between 1983-1984 and 1985-1986, which resulted from the warning campaign.

Criteria for Admissibility of Testimony Relying on Epidemiologic Studies as Evidence

Courts are concerned with the reliability of scientific evidence, especially as it is believed that lay people may give substantial weight to scientific evidence. In the United States, the *Daubert* decision, 113 US 2786 (1993), set forth criteria that courts may use to screen scientific evidence before it goes to a jury. The case concerned whether a drug, Bendectin, prescribed for morning sickness caused birth defects, especially in the limbs. Related cases and the studies are described at length in Green [21]. The *Daubert* decision replaced the *Frye* 293 F. 1013 (DC Cir. 1923) standard, which stated that the methodology used by an expert should be "generally accepted" in the field by the criteria in the Federal Rules of Evidence. The court gave the trial judge a gate-keeping role to ensure that scientific evidence is reliable. Now judges must examine the methodology used and inquire as to whether experts are basing their testimony on peer-reviewed studies and methods of analysis before admitting the evidence at trial. Several commentators, e.g., Kassirer and Cecil [22] have noted that there is substantial variability in the way lower courts apply the *Daubert* criteria when they evaluate expert reports and prospective testimony.

The US Supreme Court decision in *Daubert* remanded the case for reconsideration under the new guidelines for scientific evidence. The lower court, 43 F. 3d (9th Cir. 1995), decided that the expert's testimony did not satisfy the *Daubert* guidelines for admissibility in part because the plaintiff's expert never submitted the meta-analysis (*see Meta-Analysis in Clinical Risk Assessment*) of several studies, which was claimed to indicate an increased relative risk, for peer review. Similarly, in *Rosen v. Ciba-Geigy*, 78 F. 3d 316 (7th Cir. 1996), the court excluded expert testimony that a man's smoking while wearing a nicotine patch for 3 days caused a heart attack. The appeals court said that the expert's opinion lacked the scientific support required by *Daubert* because no study supported the alleged link between short-term use of the patch and heart disease caused by a sudden nicotine overdose. The *Rosen* opinion notes that the trial judge is not to do science but to ensure that when scientists testify in court they adhere to the same standards of intellectual rigor they use in their professional work. If they

do so and their evidence is relevant to an issue in the case, then their testimony is admissible, even though the methods used are not yet accepted as canonical in their branch of science.

In two opinions that followed *Daubert*, *Joiner v. General Electric*, 522 U.S. 136 (1997), and *Kumho Tire Co. v. Carmichael*, 119 S.Ct. 1167 (1999), the Court expanded the trial judge's role in screening expert testimony for reliability. Now, testimony relying on studies from social science and technical or engineering experience will be subject to review by the judge before the expert is allowed to testify. The *Kumho* opinion noted that the factors mentioned in *Daubert* (e.g., whether the theory or technique on which the testimony is based has been tested, whether it has been subject to peer review and publication, the known or potential error rate) were only a guideline rather than criteria to be strictly applied to prospective

expert testimony. In particular, the circumstances of the particular case will have a major role. Commentators [23–26]) have discussed its implications as well as cases where the circuit courts (covering different regions of the United States) have disagreed in their evaluations of similar evidence. Fienberg *et al.* [27] and Loue [28] discuss the reviewing process, noting some important factors for judges to consider.

The *Chapin v. A & L Parts, Inc.* (Ct. of Appeals, MI 2007) opinion provides an instructive discussion of the factors courts use in assessing the reliability of expert testimony on scientific studies. Although a number of epidemiologic studies had not found an increased risk of mesothelioma in workers handling brake parts due to possible exposure to asbestos, plaintiff's expert noted that the populations studied included workers who had less exposure than the plaintiff (who ground brake linings in his job as a

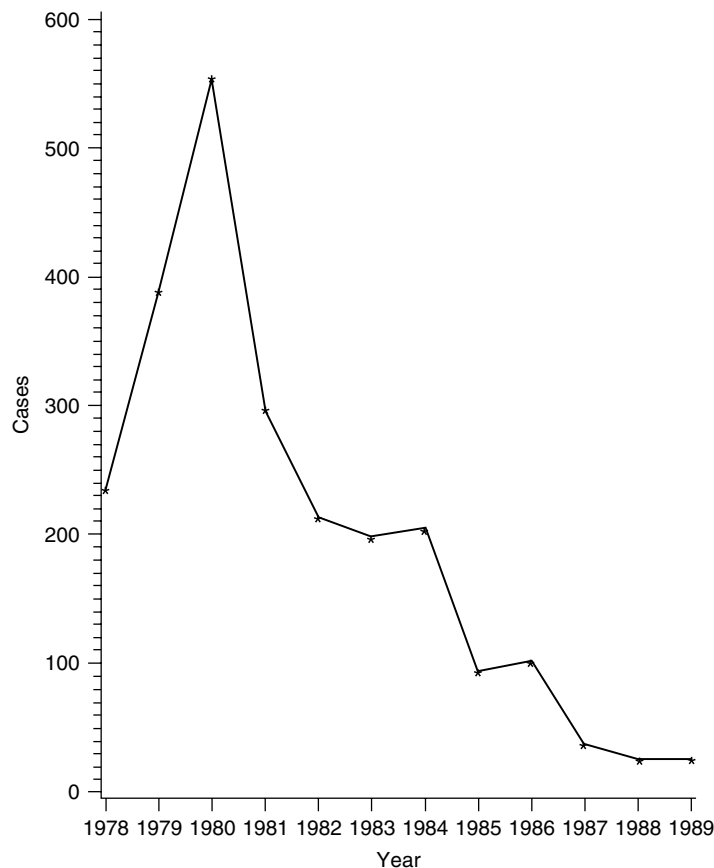


Figure 1 The number of cases of Reye's syndrome for the years 1978–1989

brake mechanic). He also noted that some studies had not allowed sufficient time (latency period) for the disease to manifest itself, while others were of limited sample size and lacked power to detect a moderate increased relative risk. Because the other criteria in the guidelines suggested by Sir Bradford Hill [29] were met and the disease is known to be highly related to asbestos exposure, he asserted that it was more likely that the plaintiff's case was caused by asbestos exposure. The defense expert apparently stated that causation could only be established by well-controlled epidemiological evidence and that case reports and other studies, e.g., animal and toxicological, were insufficient. Usually courts give much greater weight to epidemiologic studies of human populations; however, the plaintiff's expert provided a sound explanation of their deficiencies. This suggests that statistical methods for assessing the potential impact of omitted variables and other problems that inevitably occur with observational studies (Rosenbaum [30]) may have a more important role in the future. Usually these methods are used to explore whether a statistical association suggesting an increased risk of an illness can be plausibly explained by an omitted variable or other flaw; however, they can also be used to examine whether another risk factor that is more prevalent in the control group could mask a true increased risk (Yu and Gastwirth [31]).

In contrast to *Chapin*, when an expert simply relies on the temporal relationship between exposure to a drug or chemical and the subsequent development of an impairment, which may be caused by other factors, the testimony was deemed inadmissible. This occurred in *Salden v. Matrixx Initiatives* (E.D. Mich. No. 06–10277, 2007) where the expert had not performed any tests nor referred to other tests or studies published by other scientists.

Recently, in *Knight v. Kirby Inland Marine Inc.* (5th Cir. 2007), the court upheld a lower court's decision to exclude expert testimony in a maritime toxic tort suit because the court concluded that the studies relied upon by the expert failed to give an adequate basis for the opinion that the types of chemicals that the plaintiffs were exposed to in their marine employment can cause their particular injuries in the general population. The trial judge concluded that many of the studies the expert relied on had focused on several chemical exposures rather than on benzene, the chemical the plaintiffs had been

exposed to, and that some of the other studies were not statistically significant. The appellate opinion also reviews other relevant legal cases concerning the broad leeway trial judges have in performing their gate-keeping role.

Courts have reached different conclusions concerning the admissibility of the method of differential diagnosis, where medical experts conclude that a disease was caused by a particular exposure by eliminating other potential causes. The cases concerning the drug Parlodel and its relationship to stroke, discussed in [17], illustrate the problem. After studies showed that the drug could cause ischemic strokes, some plaintiffs offered expert testimony that these studies showed the drug could cause hemorrhagic strokes too. The *Rider v. Sandoz*, 295 F.3d 1194 (11th Cir. 2002) opinion upheld a lower court's rejection of this extrapolation. At the same time it cited favorably *Globetti v. Sandoz* (111 F. Supp. N.D. Ala 2001), which admitted testimony based on differential diagnosis, and also stated that epidemiologic studies are not an absolute requirement. While no human studies had been carried out, animal studies had indicated a risk. The expert was allowed to utilize this information in a differential diagnosis. Thus, the trial judge's assessment of the care and thoroughness with which a differential diagnosis or other scientific study has been carried out by a prospective expert as well as whether the expert has considered all other relevant evidence that is available at the time will be a major factor in deciding whether the testimony is admissible.

References

- [1] Robins, J. & Greenland, S. (1989). The probability of causation under a stochastic model for individual risk, *Biometrics* **45**, 1125–1138.
- [2] Gastwirth, J.L. (1988). *Statistical Reasoning in Law and Public Policy*, Academic Press, San Diego.
- [3] Finkelstein, M.O. & Levin, B. (1990). *Statistics for Lawyers*, Springer-Verlag, New York.
- [4] Green, M.D. (1988). The paradox of statutes of limitations in toxic substances litigation, *California Law Review* **76**, 965–1014.
- [5] Markesinis, B. (1994). *German Tort Law*, 3rd Edition, Clarendon Press, Oxford.
- [6] Apfel, R.J. & Fisher, S.M. (1984). *To Do No Harm; DES and the Dilemmas of Modern Medicine*, Yale University Press, New Haven.
- [7] Inskip, H.M. (1996). *Reay and Hope versus British Nuclear Fuels plc: issues faced when a research project*

- formed the basis of litigation, *Journal of the Royal Statistical Society, Series A* **159**, 41–47.
- [8] Fienberg, S.E. & Kaye, D.H. (1991). Legal and statistical aspects of some mysterious clusters, *Journal of the Royal Statistical Society* **154**, 61–174.
- [9] Rubinfeld, D.L. (1985). Econometrics in the courtroom, *Columbia Law Review* **85**, 1048–1097.
- [10] Thompson, M.M. (1992). Causal inference in epidemiology: implications for toxic tort litigation, *North Carolina Law Review* **71**, 247–291.
- [11] Green, M.D., Freedman, D.M. & Gordis, L. (2000). Reference guide on epidemiology, *Reference Manual on Scientific Evidence*, Federal Judicial Center, Washington, DC, pp. 122–178.
- [12] Carruthers, R.S. & Goldstein, B.D. (2001). Relative risk greater than two in proof of causation in toxic tort litigation, *Jurimetrics* **41**, 195–209.
- [13] Markesinis, B. & Deakin, S.F. (1994). *Tort Law*, 3rd Edition, Clarendon Press, Oxford.
- [14] Robertson, D.W., Powers Jr, W. & Anderson, D.A. (1988). *Cases and Materials on Torts*, West, St Paul.
- [15] Phillips, J.J. (1988). *Products Liability*, 3rd Edition, West, St Paul.
- [16] Selikoff, I.J., Hammond, E.C. & Churg, J. (1964). Asbestos exposure, smoking and neoplasia, *Journal of the American Medical Association* **188**, 22–26.
- [17] Gastwirth, J.L. (2003). *The Need for Careful Evaluation of Epidemiologic Evidence in Product Liability Cases: A Reexamination of Wells v. Ortho and Key Pharmaceuticals*.
- [18] Novick, J. (1987). Use of epidemiological studies to prove legal causation: aspirin and Reye's syndrome, a case in point, *Tort and Insurance Law Journal* **23**, 536–557.
- [19] Schwartz, T.M. (1988). The role of federal safety regulations in products liability actions, *Vanderbilt Law Review* **41**, 1121–1169.
- [20] US Public Health Service (1985). Public health service study on Reye's syndrome and medications, *New England Journal of Medicine* **313**, 847–849.
- [21] Green, M.D. (1996). *Bendectin and Birth Defects*, University of Pennsylvania Press, Philadelphia.
- [22] Kassirer, J.P. & Cecil, J.C. (2002). Inconsistency in evidentiary standards for medical testimony: disorder in the court, *Journal of the American Medical Association* **288**, 1382–1387.
- [23] Berger, M.A. (2000). The Supreme Court's trilogy on the admissibility of expert testimony, *Reference Manual on Scientific Evidence*, Federal Judicial Center, Washington, DC, pp. 1–38.
- [24] Faigman, D.L. (2000). The law's scientific revolution: reflections and ruminations on the law's use of experts in year seven of the revolution, *Washington and Lee Law Review* **57**, 661–684.
- [25] Hall, M.A. (1999). Applying *Daubert* to medical causation testimony by clinical physicians, *Toxics Law Reporter* **14**, 543–552.
- [26] Sacks, M.J. (2000). Banishing *Ipse Dixit*: the impact of *Kumho Tire* on forensic identification science, *Washington and Lee Law Review* **57**, 879–900.
- [27] Fienberg, S.E., Krislov, S.H. & Straf, M.L. (1995). Understanding and evaluating scientific evidence in litigation, *Jurimetrics* **36**, 1–32.
- [28] Loue, S. (2000). Epidemiological causation in the legal context: substance and procedures, in *Statistical Science in the Courtroom*, J.L. Gastwirth, ed, Springer, New York, pp. 263–280.
- [29] Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**, 295–300.
- [30] Rosenbaum, P.R. (2002). *Observational Studies*, 2nd Edition, Springer, New York.
- [31] Yu, B. & Gastwirth, J.L. (2003). The 'reverse' Cornfield inequality and its use in the analysis of epidemiologic data, *Statistics in Medicine* **22**, 3383–3401.

JOSEPH L. GASTWIRTH

Equity-Linked Life Insurance

In equity-linked life insurance contracts, the benefits are directly linked to the value of an investment portfolio. This portfolio is typically composed of units of one or more mutual funds (also called *unit trusts*, in some countries), and is kept separate from the other assets of the insurance company. That is why in Canada it is usually referred to as *segregated fund*. It can be directly managed by the insurance company, although an independent management is very common. In particular, even if the term *equity-linked* clearly applies to an investment portfolio mainly composed of equities, it can also be interpreted in a broad sense and concerns all life insurance contracts in which benefits are not (totally) expressed in units of the usual domestic currency but in units of a given asset that can be, e.g., a stock, a stock index, a mutual fund, a foreign currency, a commodity, etc. The same contracts are also referred to as *unit-linked*, or *variable life* in the United States.

Then the distinguishing feature of equity-linked contracts with respect to traditional nonparticipating policies is that the cash value of benefits is stochastic (see **Stochastic Control for Insurance Companies**). Also the premiums could be expressed in units of the reference asset, although this is rather unusual. Premiums, net of insurance and expense charges, are deemed to be invested in the reference asset, i.e., used to buy units of this asset that are credited to a separate account called *unit account* of the policyholder. Note that it is not necessary that such investments are actually made, although this could be required by regulations. In any case the number of units deemed to be acquired up to a given time defines the reference portfolio to which benefits are linked. Partial withdrawals from the unit account are sometimes allowed, as well as early surrender, which however is usually penalized.

Equity-linked policies are generally characterized by a high level of financial risk. This risk can be totally charged to the policyholder, in *pure equity-linked contracts*, or it can be shared between the policyholder and the insurance company, in *guaranteed equity-linked contracts*. In the first case the insurance company acts as a mere financial intermediary, and bears neither investment nor mortality risk

if it actually invests the premiums (net of expense charges) in the reference asset. Minimum guarantees are often offered in case of death and sometimes also in case of survival. The minimum guarantee is usually a fixed amount, possibly time dependent. This amount can be related to the total premiums paid or to the part of them deemed to have been allocated in the reference fund. In particular it can be equal to a percentage, ≤ 100 , of these premiums, with or without accrued interests at a minimum interest rate guaranteed. Both in case of death and in case of survival, the guarantees can operate only when benefits become due (*terminal guarantees*) or year by year (*annual* or *cliquet guarantees*). Obviously, annual guarantees are much riskier than terminal ones from the insurance company's point of view, and hence they could require a too high cost. There are also other kinds of guarantees more or less "exotic", e.g., they can contain a *reset* feature, etc.

Compared with traditional life insurance contracts, the main advantage of equity-linked policies from the policyholder's point of view resides in the fact that he/she can directly determine his/her investment strategy, i.e., choose the particular assets to which the policy is being linked and, often, change them during the life of the contract (*switching option*). Moreover, there could be also a certain flexibility in the choice of the periodic premium amounts, as in *Universal Life Insurance*. Last, but not least, there is more transparency about the returns on the policyholder's account as well as about the charging structure, and there are no elements of uncertainty introduced by the discretionary powers of the insurance company in setting bonuses. Compared with a direct investment in a mutual fund, the policyholder can benefit from a more favorable tax treatment, in particular tax relief is often available on the premiums (see **Longevity Risk and Life Annuities**).

The Development of the Unit Account

Consider an equity-linked contract issued at time 0. Assume that this contract is a whole life assurance, or an endowment policy, or a deferred annuity with a death benefit. In the last two cases, we denote the maturity date or the end of the accumulation period, respectively, by T . Assume that the contract is paid by a sequence of fixed (or flexible) premiums $\{P_t\}$, due at predetermined dates t (e.g., beginning of each

year, or beginning of each month, until death or until time T , whichever comes first). Note that in case of single premium contracts, P_0 represents the single premium and $P_t = 0$ for any $t > 0$. Assume, at least for the moment, that the contract is a pure equity-linked one; hence, it offers no kind of guarantees and the value of the unit account of the policyholder will be totally paid at death or maturity, or converted into a life annuity at the end of the accumulation period according to the current market conditions.

As already said, only a part of the premium P_t is deemed to be invested in the reference asset. If we denote the allocation rate by α_t (≤ 1), the amount $\alpha_t P_t$ is used to buy units of this asset that are credited to the unit account. These units are bought at the *offer price*, which at any time t is about 5% greater than the corresponding *bid price*, i.e., the price at which units are converted (“sold back”) in order to make benefit payments or expense deductions. Then the difference between offer and bid price, called *bid-offer spread*, compensates the insurance company for transaction costs incurred in buying and selling units of the reference asset. The nonallocated premium $(1 - \alpha_t)P_t$ is used to recover expenses, in particular initial and premium collection expenses.

Possible income distributions from the assets composing the reference portfolio are usually reinvested. This implies the purchase of additional units at the offer price, which are credited to the unit account. The cash value of this account at any time t , also called *face value*, is given by the number of units multiplied for the bid price.

At regular intervals, e.g., every month, charges for administrative expenses are deducted from the unit account. This implies the cancelation of a certain number of units at the bid price in order to recover such expenses, which are usually expressed as a fixed amount, possibly time dependent (e.g., linked to a price index). Moreover, also a fund management charge, expressed as a percentage of the face value of the unit account (commonly in the range 0.25–1%), is deducted by canceling units at the bid price.

All unallocated premiums, as well as all charges deducted from the unit account or deriving from the bid-offer spread, constitute another reserve, called *nonunit account* (or *sterling account*, in the United Kingdom), which could also be negative (see **Premium Calculation and Insurance Pricing**). Note

that the value of this account is not so transparent to the policyholder as that of the unit account. In fact it behaves like the reserves backing traditional, nonlinked, business.

Till now we have considered a pure equity-linked contract that works exactly as a financial investment plan. Then, as already said, the insurance company bears no kind of risk (provided that charges for expenses are appropriate). Instead if we consider a guaranteed contract, in which the death benefit is fixed or there is a minimum guarantee at death or maturity, there are additional charges, called *insurance* or *risk charges*, that increase the nonunit reserve. These charges are given by the risk premium for (possible) *nonunit benefits* that will be deducted from the nonunit reserve, when due. In case of periodic premiums the insurance charges imply a decrease of the allocatable amount $\alpha_t P_t$ (if sufficient), while in case of paid-up policies it is very common to make regular deductions from the unit account, as for administrative and fund management charges, i.e., to transfer funds from the unit to the nonunit account.

The nonunit death benefit is given by the *sum at risk*, i.e., by the difference between the death benefit and the face value of the unit account. This difference could also be negative when the death benefit is fixed, and in this case the mortality “charge” could involve a transfer of funds from the nonunit to the unit account. However, if there is a minimum death guarantee, the sum at risk is never negative and can be seen as the payoff of a put option with underlying variable being the face value of the unit account and exercise price being the minimum amount guaranteed (see **Optimal Stopping and Dynamic Programming**).

Insurance charges are also applied when there are minimum guarantees in case of survival. As far as non-unit survival benefits are concerned, note that they are not necessarily related to the presence of survival guarantees. In fact they can arise also if no benefits are actually paid, and in this case they imply a transfer of funds from the nonunit to the unit account. This happens, e.g., when a cliquet guarantee becomes effective, because the face value of the unit account at the end of each year cannot be less than the corresponding annual guarantee settled at the beginning of the year.

To sum up, we represent, in what follows, the evolution of the unit account in the generic time interval $[t, t + 1]$ (during the accumulation period, if

a variable annuity is dealt with). We assume that there are no movements of units between t and $t + 1$ and that premiums (if any) are collected (and allocated) at the beginning of the interval, as well as insurance and administrative charges, while fund management charges and (possible) income distributions or transfers of funds due to nonunit survival benefits are made at the end of the interval. Of course, if there are no premiums, or transfers, or income distributions, the corresponding variables are set equal to 0. In particular, as for insurance charges, we assume that (part of) administrative charges are deducted from the unit account, at the bid price, only if the amount deemed to be allocated in the reference fund $\alpha_t P_t$ is not sufficient, otherwise they are directly deducted from this amount before the purchase of new units at the offer price.

Besides the variables α_t and P_t already introduced, we denote by n_τ the number of units of the unit account at time τ ($\tau = t, t + 1$), after deductions for fund management charges, transfer of funds for nonunit survival benefits, income distribution and consequent reinvestment, and before payment and premium allocation, as well as before insurance and administrative charges; F_{t+1} is the face value of the unit account at time $t + 1$, O_τ (respectively B_τ) the offer (bid) unit price at time τ ($\tau = t, t + 1$), I_t the insurance charge, A_t the charge for administrative expenses, f_{t+1} the rate of fund management charge, D_{t+1} the distributed income, and S_{t+1} the nonunit survival benefits transferred from the nonunit to the unit account.

We then have,

$$\begin{aligned} n_{t+1} = n_t + & \frac{\max\{\alpha_t P_t - I_t - A_t, 0\}}{O_t} \\ & - \frac{\max\{I_t + A_t - \alpha_t P_t, 0\}}{B_t} \\ & - f_{t+1} n_{t+1} + \frac{S_{t+1}}{B_{t+1}} + \frac{D_{t+1}}{O_{t+1}} \end{aligned} \quad (1)$$

from which

$$\begin{aligned} n_{t+1} = & \frac{1}{1 + f_{t+1}} \left[n_t + \frac{\max\{\alpha_t P_t - I_t - A_t, 0\}}{O_t} \right. \\ & - \frac{\max\{I_t + A_t - \alpha_t P_t, 0\}}{B_t} \\ & \left. + \frac{S_{t+1}}{B_{t+1}} + \frac{D_{t+1}}{O_{t+1}} \right] \end{aligned} \quad (2)$$

and

$$F_{t+1} = n_{t+1} B_{t+1} \quad (3)$$

Note that the number of units bought with the nonunit survival benefits is credited at the bid price, because the amount of such benefits has to correspondingly increase the face value of the unit account, computed at the bid price.

The face value of the unit account is then entirely paid in case of death or survival at maturity, or its units are transformed into units of a paid-up annuity at the end of the accumulation period. In case of surrender it is usual to pay back this value after the application of a surrender penalty. Any other payments deriving from minimum guarantee provisions are instead made by the nonunit account.

Hedging Strategies for Nonunit Benefits

We have seen, in the previous section, that if the unit account is not only a notional portfolio but the real investment reserve backing the policy under scrutiny, then the (residual) liabilities that the insurance company has to hedge are given by the nonunit benefits. These benefits are null in pure equity-linked contracts, while they are always non-negative in guaranteed (uncapped) contracts.

To fix the ideas, consider the endowment policy and assume that in case of death during a given year of contract, the benefit is paid at the end of the year. Assume that the contract provides some terminal minimum guarantees at death and/or survival at maturity. We denote the minimum amount guaranteed in case of death during the t th year of contract by G_t^D , $t = 1, 2, \dots, T$, and the minimum amount guaranteed in case of survival at maturity by G_T^S . These amounts are assumed to be fixed at inception. In particular, if $G_t^D = 0$ for any t there are no minimum death guarantees, while if $G_T^S = 0$ there are no survival guarantees.

Then the stream of liabilities of the insurance company due to nonunit death benefits, $\{L_t^D\}_{t=1}^T$, is given by

$$\begin{aligned} L_t^D = & \max\{G_t^D - F_t, 0\} \mathbb{I}_{[t-1 < T_x \leq t]}, \\ & t = 1, 2, \dots, T \end{aligned} \quad (4)$$

where T_x denotes the remaining lifetime of the insured (with entry age x) and $\mathbb{I}_{\{\mathcal{E}\}}$ the indicator function of the event $\mathcal{E}$, equal to 1 if $\mathcal{E}$ is true and 0 otherwise. Moreover, since we are dealing with terminal guarantees, there is only one nonunit survival benefit that will be paid in case of survival at maturity; hence, the corresponding liability is given by

$$L_T^S = \max\{G_T^S - F_T, 0\} \mathbb{I}_{\{T_x > T\}} \quad (5)$$

All these liabilities are clearly dependent on both mortality and financial risk. In particular, they can be seen as the payoff of a (single) put option with underlying variable being the face value of the unit account (as already said), time-dependent exercise price (and possibly also survival-dependent, if $G_T^D \neq G_T^S$), and stochastic exercise date. Taking into account that this date is not chosen by the option holder, as in the case of American options, but is triggered by an event, the death of the insured, which can be reasonably assumed to be independent of the behavior of financial markets, such an option is something in between European and American styles; that is why Milevsky and Posner called it *Titanic option* (see [1]).

However, if the mortality risk is diversified away because the insurance company is able to sell enough contracts in order to eliminate the fluctuations between actual and expected mortality rates, this Titanic option can be replaced by a portfolio of European options with different maturities $t = 1, 2, \dots, T$ and payoff stream given by

$$\begin{cases} \tilde{L}_t^D = \max\{G_t^D - F_t, 0\} {}_{t-1|}q_x, & t = 1, 2, \dots, T \\ \tilde{L}_T^S = \max\{G_T^S - F_T, 0\} {}_T p_x \end{cases} \quad (6)$$

where the standard actuarial symbol ${}_{t-1|}q_x$ denotes the probability that the insured dies during the t th year of contract, and ${}_T p_x$ the probability that he/she is still alive at the maturity date. Then the simplest approach to hedge the liabilities since inception is undoubtedly that of transferring them to a third party, i.e., to buy the put options in the market, if traded, or from a reinsurer. However, due to the long-term nature of life insurance business, it is very unlikely that these options are traded. When this approach is not viable either because no options on the reference fund are available, even if issued by a reinsurer, or because they would require a too

high cost, the insurance company has to directly hedge its liabilities with investments in the available assets, typically bonds and units of the reference portfolio.

We observe that even if the analysis of equity-linked life insurance policies from an “actuarial” point of view was a “hot” topic that attracted the interest of many actuaries since the issue of the first contracts in the beginning of the second half of the twentieth century, Boyle, Brennan, and Schwartz were the first who recognized that a guaranteed benefit can be decomposed in terms of European options, once the mortality risk is diversified away, in their pioneering work of the mid-1970s (see [2–5]), and applied to them on the then recent results on option pricing theory initiated by Black, Merton, and Scholes at the beginning of the same decade (see [6, 7]). In particular, Boyle, Brennan, and Schwartz analyzed an equity-linked endowment policy (which we are considering here) in an idealized framework in which insurance and financial markets are perfectly competitive and frictionless, free of arbitrage opportunities, and trading takes place in continuous time. Hence they did not take into account transaction costs and consequent bid–ask spread, at least in a first moment, as well as expenses and relative charges. Moreover, they assumed no dividend distributions from the assets composing the reference portfolio.

In such a framework, characterized by complete financial markets, the payoff of any contingent claim and hence also that of European put options can be perfectly replicated by means of a dynamic self-financing hedging strategy in riskless bonds and underlying asset. This strategy, also called *risk-neutral strategy*, involves continuous rebalancing of the hedging portfolio at no cost (and no proceeds). In particular, the hedging of a put option requires to shortsell a certain quantity of underlying asset given by the partial derivative (“sensitivity”) of the option price with respect to the price of the same asset. Then, since the risk-neutral strategy allows to perfectly replicate the final payoff of the option without requiring additional money (apart from that required to set it up), to prevent arbitrage opportunities the price of the option must be given by the initial cost of the strategy. It can be shown that this price is expressible as the expectation, under a suitable probability measure called *risk-neutral measure*, of

the final payoff of the option discounted with the rate of return on risk-free assets. To sum up, in this framework the “fair” insurance charges are given by the expectation, under the risk-neutral measure, of all nonunit liabilities $\{\tilde{L}_t^D\}_{t=1}^T$ and $\tilde{L}_T^S$ discounted with the risk-free rate (see **Statistical Arbitrage**).

Unfortunately, the risk-neutral strategy just described requires a continuous rebalancing of the hedging portfolio, but trading in the “real world” is not time-continuous and, moreover, the real world is affected by transaction costs that could make very frequent rebalancing prohibitive. Then there is a trade-off between the frequency of rebalancing in a discrete setting (that determines the “hedging error”) and transaction costs. For this reason Brennan and Schwartz in [4, 5] analyze discrete approximations of the perfect (risk-neutral) hedging strategy by taking transaction costs into account and assuming that rebalancing takes place either at regular intervals or when the hedging error (absolute or relative) exceeds a given threshold. We remark that since the hedging of nonunit benefits would require the short sale of the reference asset, in order to reduce transaction costs it is better to hedge the total liabilities (unit + nonunit part) as a whole and, hence, keep investments in the reference asset less than those defining the unit account of the policyholder.

Another approach to hedge the nonunit benefits, called sometimes the *actuarial approach*, consists in projecting, by means of a stochastic simulation model, the future liabilities of the insurance company and keeping investments in risk-free assets sufficient to meet them with a “high” probability. The initial cost required by this value-at-risk (VaR)-based approach, pioneered in [8], is generally higher than that required by the risk-neutral one, but there is a high probability that the value of the hedging portfolio turns out to be greater than the liabilities when the contract expires, and hence that money can be recovered back by the insurance company. However, if the reference portfolio exhibits an extremely bad performance, this approach can lead to high losses. For further details see [9, 10].

So far we have assumed that the mortality risk is diversified away. If this is not the case, i.e., if the insurance company does not sell enough contracts and/or there is a systematic part of mortality risk that affects all policies in the same direction, then we cannot replace the Titanic option with a portfolio of European options and hence reduce

ourselves to hedge purely financial contingent claims in a complete market. In this case the insurance company has to hedge an *integrated risk*, depending on both mortality and financial uncertainty, in a typically incomplete market, where not all claims are replicable. This implies the choice of a particular *risk-minimizing* hedging strategy, which can be followed also in case of diversifiable mortality risk when the financial market is however incomplete. For an analysis of these risk-minimizing strategies, we refer to the work by Møller (see [11–14]).

To conclude, we observe that the same philosophy presented here with reference to the hedging of terminal guarantees can be applied to the case of cliquet guarantees. In fact, also in this case the nonunit benefits can be expressed as the payoff of put options, although with exercise price not known at inception but only at the beginning of each year of contract. Then it is more appropriate to set up suitable hedging strategies for them at the beginning of each year, rather than directly at inception. This is also in line with the practice (described in the previous section) of making regular deductions from the unit account for insurance charges even in the case of single premium contracts.

References

- [1] Milevsky, M.A. & Posner, S.E. (2001). The titanic option: valuation of the guaranteed minimum death benefit in variable annuities and mutual funds, *The Journal of Risk and Insurance* **68**(1), 93–128.
- [2] Brennan, M.J. & Schwartz, E.S. (1976). The pricing of equity-linked life insurance policies with an asset value guarantee, *Journal of Financial Economics* **3**, 195–213.
- [3] Boyle, P.P. & Schwartz, E.S. (1977). Equilibrium prices of guarantees under equity-linked contracts, *The Journal of Risk and Insurance* **44**(4), 639–660.
- [4] Brennan, M.J. & Schwartz, E.S. (1979). Alternative investment strategies for the issuers of equity-linked life insurance policies with an asset value guarantee, *Journal of Business* **52**(1), 63–93.
- [5] Brennan, M.J. & Schwartz, E.S. (1979). *Pricing and Investment Strategies for Equity-Linked Life Insurance*, The S.S. Huebner Foundation for Insurance Education, Wharton School, University of Pennsylvania, Philadelphia.
- [6] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**(3), 637–654.
- [7] Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.

- [8] Maturity Guarantees Working Party of the Institute and Faculty of Actuaries (1980). Report of the maturity guarantees working party, *Journal of the Institute of Actuaries* **107**, 103–231.
- [9] Boyle, P.P. & Hardy, M. (1997). Reserving for maturity guarantees: two approaches, *Insurance: Mathematics and Economics* **21**(2), 113–127.
- [10] Hardy, M. (2003). *Investment Guarantees: Modeling and Risk Management for Equity-Linked Life Insurance*, John Wiley & Sons, New York.
- [11] Møller, T. (1998). Risk-minimizing hedging strategies for unit-linked life insurance contracts, *ASTIN Bulletin* **28**(1), 17–47.
- [12] Møller, T. (2001). Hedging equity-linked life insurance contracts, *North American Actuarial Journal* **5**(2), 79–95.
- [13] Møller, T. (2001). Risk-minimizing hedging strategies for insurance payment processes, *Finance and Stochastics* **5**, 419–446.
- [14] Møller, T. & Steffensen, M. (2007). *Market Valuation Methods in Life and Pension Insurance*. Cambridge University Press, Cambridge.

ANNA R. BACINELLO

Estimation of Mortality Rates from Insurance Data

Mortality rates are instrumental in life insurance. These rates form the basis for pricing of insurance products and determination of reserves for a life insurance company. They also can be used to determine other functions within a life table (*see Insurance Applications of Life Tables*), such as the life expectancy (*see Longevity Risk and Life Annuities*) and survival rates. Consequently the determination of these rates from data in life insurance studies plays a very important role in the actuarial profession.

Within the Society of Actuaries in the United States of America, intercompany data on insurance policies issued to standard lives (policies issued to lives that were judged as “healthy” when the insurance policy was issued) resulted in the 1965–1970 [1] and the 1975–1980 basic mortality tables [2]. More recent studies [3] are based on the Society of Actuaries 1995–2000 mortality study.

This paper is split into four parts. The first part defines the mortality rates that are pertinent in life insurance. The second part discusses the estimation of these mortality rates from intercompany insurance data. The third part discusses the methods used in determining rates for ages not available, called *interpolation*, and the smoothing of rates to take into account general mortality trends. Taken together, these methods are referred to as *graduation*. The fourth part compares the two general graduation methods and discusses the advantages and disadvantages of each approach.

Definition of Mortality Rates

The mortality rate, for a given age (x), is defined as the probability that a life now aged (x) will die within 1 year. The actuarial convention is to denote this mortality rate as q_x [4, p. 53]. It can be theoretically calculated or practically estimated by taking the ratio of the number of people dying between ages (x) and ($x + 1$) in a year and dividing by the number of people who are alive at age (x) at the beginning of the year. Associated with the mortality rate is the

force of mortality, which is also called the *hazard rate* in reliability theory [4, p. 55]. This force of mortality, denoted as μ_x , can be interpreted as the instantaneous annualized death rate for a life aged x . The relationship between μ_x and q_x is

$$q_x = 1 - \exp\left(-\int_0^1 (\mu_{x+t}) dt\right) \quad (1)$$

The mortality rates, q_x , are generally broken down into various categories, such as gender and smoking status.

Another important variable in the mortality rate, which is pertinent in insurance investigations on standard lives, is the time since the particular life now aged x was insured, which is called the *duration*. When this variable is included in the mortality rate, the corresponding mortality rate is called the *select mortality rate* and is defined as follows:

$$q_{[x-t]+t} = \text{probability that a life now aged} \\ x \text{ who was insured} \\ t \text{ years ago will die within 1 year} \quad (2)$$

where x is the attained age and t is the duration. The age at issue would be $x - t$.

In general, the mortality rate will be an increasing function of attained age for fixed duration for ages above 30. For ages below 30, the mortality rates [5, p. 84] decrease sharply at the younger ages because of the infant mortality rate, remain fairly constant until the early teens, fall between ages 18 and 25 because of the accidental death rate, and then continue to increase to the end of the life table.

The mortality rate is an increasing function of duration for fixed attained age. The latter trend is true in the case of medically issued standard insurance when a life has to provide evidence of insurability by undergoing a medical examination. Hence a life now aged 35 who was insured 5 years ago at age 30 is subject to a higher mortality risk than a life now aged 35 who was just insured. This implies that

$$q_{[30]+5} \geq q_{[35]} \quad (3)$$

After a certain selection period, the effects of selection wear off and the increase in mortality as a function of duration becomes negligible. At that point,

$$q_{[x-t]+t} = q_x \quad (4)$$

For example, the 1975–1980 and the 1965–1970 basic tables assume a 15-year select period [1, 2]. In equation (4) the mortality rate q_x is called the *ultimate rate*. If a table of mortality rates does not include select rates, the mortality rate q_x is called the *combined rate*.

Estimation of Mortality Rates

The traditional method of estimation of mortality rates is to base the estimate on the usual binomial maximum-likelihood estimate, namely,

$$\hat{q} = \frac{X}{n} \quad (5)$$

where X is the number of annual deaths for that particular mortality class, n is the number of lives which are still alive in the beginning of the period for that particular mortality class, and $\hat{q}$ is the estimated mortality rate for that particular mortality class.

The data usually come in different groups by duration and issue age. For the construction of the 1965–1970 and the 1975–1980 basic tables, the data were in varying age groups by issue age (0, 1, 2–4, 5–9, and in 5-year classes up to 65–69, with 70 and over being the last class) and by duration (from 0 to 14).

Unfortunately, the number of lives that are still alive at the beginning of the period are not available, because insured lives are included in the study at all times. Consequently the binomial maximum-likelihood estimate does not apply. The traditional method is to interpret n in equation (5) as the number of exposures to death. For example, if an insured becomes part of the study halfway through a period, he will be counted as 0.5 of an exposure (see **Censoring**). These methods are discussed in [6–8]. An alternative method was developed by Kaplan and Meier, known as the *Kaplan Meier estimate* [9], which has been used to estimate survival probabilities in biostatistics applications (see **Hazard and Hazard Ratio**). The traditional actuarial method is equivalent to the Kaplan Meier estimate under certain assumptions [5, p. 323].

The result of this process will provide mortality rates for different classifications of issue ages and durations, but not for every integral attained age. To determine the select and ultimate rates by age and duration, as given by equations (2) and (4), the process of mortality rate graduation is required.

Graduation of Mortality Rates

Graduation is the process of taking the estimated mortality rates that are available in different age classes, and by a process of interpolation and smoothing, producing rates for all integral ages and durations. The purpose of interpolation is to provide estimates for $q_{[x-t]+t}$ and q_x for all integral ages from the grouped data estimates obtained from the insurance data. The purpose of smoothing is to take into account known mortality trends that may not apply to the estimates because of random fluctuations and differences among company experience that have contributed to the data.

There are two general methods for accomplishing this process. One method is to use smoothing techniques based on numerical methods. This area has a very rich history in the actuarial literature. Classical methods of graduation were discussed in Miller [10] and more recently by London [11]. One of the first known methods is the Whittaker–Henderson method [12, 13]. This method defines an objective function, which is a weighted average of the measure of fit (as measured by the sums of the squares of the differences between the smoothed mortality rates and the estimated mortality rates) and a measure of smoothness (as measured by third differences of the smoothed mortality rates). The Whittaker–Henderson method then selects the smoothed rates that minimize this objective function. This method was used to graduate the 1975–1980 basic tables [2]. More recent research on Whittaker–Henderson graduation methods by Lowrie [14] takes into account mixed differences and prior information. It also contains a very comprehensive bibliography of the development of Whittaker–Henderson graduation methods. A method of graduation based on kernel methods is discussed by Gavin *et al.* [15]. Bayesian graduation methods were discussed by Kimeldorf and Jones [16] and Hickman and Miller [17]. In this method, Bayesian ideas (see **Bayes’ Theorem and Updating of Belief**) were used to take into account prior information on the structure of the mortality curve, and this prior information was modified by using estimates based on data obtained from current mortality studies. This method was used in the construction of the 1965–1970 basic tables with the prior information based on the 1955–1960 basic tables [1]. Further research on this method has continued [18, 19].

More recent methods for interpolation and smoothing are based on the use of splines, which was developed by Schoenberg [20]. These methods have largely replaced the older methods because the solutions are based on linear systems of equations that could be easily solved on the computer, see [5, p. 485]. For an example involving cubic splines, see [5, pp. 485–514].

The other method of smoothing is based on using mathematical laws of mortality and fitting the data to these mathematical laws using regression least square methods or maximum-likelihood methods. This method was traditionally known as *graduation by mathematical formula*. Two such laws are those of Gompertz and Makeham, see [4, p. 78]. These laws relate the force of mortality to the attained age as follows:

$$\text{Gompertz Law: } \mu_x = Bc^x \quad (6)$$

$$\text{Makeham's Law: } \mu_x = A + Bc^x \quad (7)$$

Gompertz law assumes that the force of mortality increases at a constant rate per age. Makeham's law is a modification of Gompertz's law and is an attempt to take into account the smaller progression of mortality rates at the younger ages. Equation (1) can be used to express the mortality rates as a function of the Makeham and Gompertz parameters. Carriere [21] presented a great discussion and bibliography of the Gompertz law and other laws that have been proposed.

Several attempts have been made to modify these laws to model the selection effect. Tenenbein and Vanderhoof [22] developed several three- and four-parameter Gompertz curves to model the select and ultimate rates for attained ages above 30 and applied this method to fit the data upon which the 1965–1970 and 1975–1980 basic tables are based. Carriere [23] presented an 11-parameter model that applies for all attained ages and durations and used it to fit the data upon which the 1975–1980 basic tables are based.

Comparison of the Graduation Methods

The two methods of graduation are numerical methods and graduation based on mathematical formula. Each method has its advantages and disadvantages. The advantage of the numerical methods is that it will always form a set of smooth rates that provide a good

fit of the data. Considering the complexity of the mortality rates by age, this is a strong advantage. Its big disadvantage is that it cannot be used for extrapolation for ages that are not in the data set. For example, this method of graduation would be unreliable to estimate mortality rates at the higher ages [5, p. 488]. This is due to the fact that graduation by numerical methods is a purely mechanical method, which does not take into account any biological considerations.

Graduation by mathematical formula provides a method that will produce graduated mortality rates that are naturally smooth. The formulae are determined by taking into account the biological considerations. If the models work, then there is a greater justification for extrapolating beyond the ages in the data set. The main drawback is that it is difficult to produce a formula that works over the entire age range. For this latter reason, graduation by mathematical formula has not been widely used.

On the other hand, when it is desired to project mortality trends into the future, or to estimate mortality rates at the higher ages, or to estimate mortality patterns when data is not plentiful, mathematical models are very useful. Lee and Carter [24] and Lee [25] presented a mathematical model for forecasting mortality, and estimated the parameters using least-squares methods. The North American Actuarial Journal devoted almost an entire issue discussing the problem of forecasting mortality change and its impact on Social Security [26]. Wong-Fillipow and Haberman [27] adapted the Carriere model [23] to estimate mortality rates for lives over 80 by altering the parameters. Mathematical models for estimating mortality patterns at the higher ages are discussed by Buettner [28] and Watts *et al.* [29]. A logistic regression model was proposed for modeling pension plan mortality data by Vinsonhaler *et al.* [30].

References

- [1] Committee on Mortality under Ordinary Insurance and Annuities (1974). 1965–1970 basic tables, *Transactions of the Society of Actuaries* **3**, 199–224.
- [2] Committee on Mortality under Ordinary Insurance and Annuities (1982). 1975–1980 basic tables, *Transactions of the Society of Actuaries* 55–81.
- [3] Rhodes, T. (2004). SOA 1995–2000 mortality study, *Society of Actuaries Newsletter of the Individual Life Insurance and Annuity Product Development Section* **59**, 13–15.

- [4] Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A. & Nesbitt, C.A. (1997). *Actuarial Mathematics*, Society of Actuaries, Schaumburg.
- [5] Klugman, S.A., Panjer, H.H. & Willmot, G.E. (2004). *Loss Models: From Data to Decisions*, Wiley-Interscience, Hoboken.
- [6] Batten, R.W. (1978). *Mortality Table Construction*, Prentice Hall, Englewood Cliffs.
- [7] Gershenson, H. (1961). *Measurement of Mortality*, Society of Actuaries, Chicago.
- [8] London, D. (1997). *Survival Models and their Estimation*, ACTEX Publications, Winsted.
- [9] Kaplan, E. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.
- [10] Miller, M.D. (1949). *Elements of Graduation*, The Actuarial Society of America and the Institute of Actuaries, Philadelphia.
- [11] London, D. (1985). *Graduation: The Revision of Estimates*, ACTEX Publications, Winsted.
- [12] Henderson, R. (1938). *Mathematical Theory of Graduation*, Actuarial Society of America, New York.
- [13] Spoerl, C.A. (1938). The Whittaker-Henderson graduation formula, *Transactions of the Actuarial Society of America* **38**, 403–462.
- [14] Lowrie, W.B. (1993). Multidimensional Whittaker-Henderson graduation with constraints and mixed differences, *Transactions of the Society of Actuaries* **45**, 215–256.
- [15] Gavin, J., Haberman, S. & Verrall, R. (1995). Graduation by kernel and adaptive kernel methods with a boundary correction, *Transactions of the Society of Actuaries* **47**, 173–210.
- [16] Kimeldorf, G.S. & Jones, D.A. (1967). Bayesian graduation, *Transactions of the Society of Actuaries* **19**, 66–112.
- [17] Hickman, J.C. & Miller, R.B. (1977). Notes on bayesian graduation, *Transactions of the Society of Actuaries* **29**, 7–49.
- [18] Broffitt, J.D. (1988). Increasing and increasing convex bayesian graduation, *Transactions of the Society of Actuaries* **40**, 115–148.
- [19] Carlin, B.P. (1992). A simple Monte Carlo approach to bayesian graduation, *Transactions of the Society of Actuaries* **44**, 55–76.
- [20] Schoenberg, I. (1964). Spline functions and the problem of graduation, *Proceedings of the National Academy of Science* **52**, 947–950.
- [21] Carriere, J.F. (1992). Parametric models for life tables, *Transactions of the Society of Actuaries* **44**, 77–99.
- [22] Tenenbein, A. & Vanderhoof, I.T. (1980). New mathematical laws of select and ultimate mortality, *Transactions of the Society of Actuaries* **32**, 119–183.
- [23] Carriere, J.F. (1994). A select and ultimate parametric model, *Transactions of the Society of Actuaries* **46**, 75–98.
- [24] Lee, R. & Carter, C. (1992). Modeling and forecasting the time series of US mortality, *Journal of the American Statistical Association* **87**, 659–671.
- [25] Lee, R. (2000). The Lee-Carter method for forecasting mortality with various extensions and applications, *North American Actuarial Journal* **4**(1), 80–93.
- [26] Society of Actuaries Seminar (1998). Impact of mortality improvement on social security: Canada, Mexico, and the United States, *North American Actuarial Journal* **2**(4), 10–138.
- [27] Wong-Fupuy, C. & Haberman, S. (2004). Projecting mortality trends: recent developments in the United Kingdom and the United States, *North American Actuarial Journal* **8**(2), 56–83.
- [28] Buettner, T. (2002). Approaches and experiences in projecting mortality patterns for the oldest-old, *North American Actuarial Journal* **6**(3), 14–29.
- [29] Watts, K.A., Dupuis, K.J. & Jones, B.L. (2006). An extreme value analysis of advanced age mortality data, *North American Actuarial Journal* **10**(4), 162–178.
- [30] Vinsonhaler, C., Ravishankar, N., Vadiveloo, J. & Rasoanaivo, G. (2001). Multivariate analysis of pension plan mortality data, *North American Actuarial Journal* **5**(2), 126–138.

Related Articles

Fair Value of Insurance Liabilities

Individual Risk Models

Longevity Risk and Life Annuities

EDWARD L. MELNICK AND AARON TENENBEIN

Ethical Issues in Using Statistics, Statistical Methods, and Statistical Sources in Work Related to Homeland Security

This article addresses the role of ethics (*see Risk in Credit Granting and Lending Decisions; Credit Scoring; Syndromic Surveillance*) in work on statistics and homeland security (*see Managing Infrastructure Reliability, Safety, and Security; Game Theoretic Methods; Sampling and Inspection for Monitoring Threats to Homeland Security*), particularly focusing on issues relevant to the use of data mining (*see Change Point Analysis; Early Warning Systems (EWSs) for Predicting Financial Crisis; Privacy Protection in an Era of Data Mining and Record Keeping*) and risk assessment in counterterrorism applications. However, many of the principles and issues raised apply broadly to other statistical applications in the field related to homeland security and, indeed, to other applications as well.

Any consideration of ethical issues in any scientific or public policy context must begin with a clear understanding that ethics, science, and law while they often largely overlap and reinforce one another, sometimes diverge or even are in conflict. At the extreme, the law may be used to justify ethically flawed behavior or the perceived needs of science may be used to excuse serious ethical misconduct.

Historically, these conflicts have arisen most frequently in times of national crisis and so are particularly relevant to threat assessment generally and in the context of homeland security. In such pressing circumstances, the explicit consideration of ethical issues is often turned into a consideration of whether or not a given action is legal, or in conformity with permitted administrative arrangements, or simply in accord with the policies and priorities of the current government. In addition, because statisticians also properly perceive themselves as engaged in scientific work, they sometimes consider that the beneficence of their scientific work, particularly when apparently legal, excuses them from further ethical examination.

These practices and views, however, are short-sighted and in the extreme contrary to the laws

of many countries as well as international law (for example, national and international laws relating to genocide and crimes against humanity). It may be recalled that among the defenses offered by those charged with medical experiments, and the related collection of information and anthropological materials, in the so-called doctors' trial at Nuremberg after World War II were the defenses that (a) their actions were legal under current national law and (b) they were engaged in scientific work for an important and beneficent purpose. These defenses were explicitly rejected by the Nuremberg Tribunal when it found the defendants guilty of crimes against humanity. Subsequently, many national and international ethical statements, research policies and regulations, as well as related laws have adopted the same perspective.

In considering the use of statistics and statistical sources and methods in risk assessment related to the investigation of terrorism and homeland security, three different sets of ethical considerations arise: (a) those related to the validity of estimates made and conclusions drawn; (b) those arising when the data sources used were originally collected for statistical purposes under a pledge of statistical confidentiality; and (c) those related to the ethical character of the intermediary and final uses that the data, estimates, and analysis are intended to serve.

We take up each of these sets of issues in turn, referring to them as validity, statistical confidentiality, and responsibility, respectively, while recognizing that the terms validity, statistical confidentiality, and responsibility are incomplete descriptors of all the issues to which they refer. In doing so, we draw on the ethical standards adopted by the American Statistical Association (ASA) [1], the Association for Computing Machinery (ACM) [2], and the International Statistical Institute (ISI) [3], three diverse but relevant professional organizations.

Validity

Not only do statisticians and others using data have a scientific responsibility to use appropriate methods in data collection, analysis, estimation, and presentation, but they have an ethical responsibility to do so as well (*see Expert Judgment; Repeated Measures Analyses*). For example, a number of provisions of the ASA's Ethical Guidelines for Statistical Practice directly address issues of the validity of risk

assessments, particularly in the context of applications related to homeland security:

Guard against the possibility that a predisposition by investigators or data providers might predetermine the analytic result. Employ data selection or sampling methods and analytic approaches that are designed to assure valid analyses ... Assure that adequate statistical and subject-matter expertise are both applied to any planned study. If this criterion is not met initially, it is important to add the missing expertise before completing the study design ... Use only statistical methodologies suitable to the data and to obtaining valid results. For example, address the multiple potentially confounding factors in observational studies, and use due caution in drawing causal inferences ... The fact that a procedure is automated does not ensure its correctness or appropriateness; it is also necessary to understand the theory, the data, and the methods used in each statistical study. This goal is served best when a competent statistical practitioner is included early in the research design, preferably in the planning stage [1, Sections II.A.2, .4, .5, and .7].

And in reporting the results of any analysis or estimation process:

Report statistical and substantive assumptions made in the study ... Account for all data considered in a study and explain the sample(s) actually used. Report the sources and assessed adequacy of the data. Report the data cleaning and screening procedures used, including any imputation. Clearly and fully report the steps taken to guard validity. Address the suitability of the analytic methods and their inherent assumptions relative to the circumstances of the specific study. Identify the computer routines used to implement the analytic methods. Where appropriate, address potential confounding variables not included in the study ... Report the limits of statistical inference of the study and possible sources of error. For example, disclose any significant failure to follow through fully on an agreed sampling or analytic plan and explain any resulting adverse consequences [1, Sections II.C.2, .5, .6, .7, .8, .9, and .12].

While these principles seem only to be sound science, they have a heavy ethical importance as well, particularly in circumstances of risk assessment when the perceived threats are seen as so great that efforts are made to set aside sound scientific and technical cautions.

Confidentiality

Confidentiality issues play little role in ethical considerations for many kinds of risk assessment applications. However, in those applications that involve the use of data sets referring to human populations (for example, population censuses or population registration systems) or that involve efforts at individual or group surveillance of human populations, confidentiality (see **Public Health Surveillance**) issues may present important ethical as well as legal and policy challenges.

In order to promote as complete and accurate reporting as possible, both government and private agencies engaged in the collection of data for statistical purposes give assurances that the information provided will not be used to harm the data providers. A key element of this protection is the assurance that the information so obtained will be kept confidential (that is, not shared with other government agencies with intelligence, police, surveillance, and other non-statistical objectives) and only be used to generate statistics. The following are the words of the United Nation's Fundamental Principles of Official Statistics in this regard:

Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes [4, Principle 6].

Both the ISI's Declaration of Professional Ethics [3, Clause 4] and the ASA's Ethical Guidelines for Statistical Practice [1, Section II.D] underscore the principle of statistical confidentiality, which is embodied in the laws of many countries.

The ISI Declaration also places this principle in a useful context:

Statisticians should be aware of the intrusive potential of some of their work. They have no special entitlement to study all phenomena. The advancement of knowledge and the pursuit of information are not themselves sufficient justifications for overriding other social and cultural values [3, Clause 4.1].

Statisticians and others attempting to assess the impact of violations of statistical confidentiality often mistakenly focus their attention only on the technical indicator, "risk of disclosure", rather than also examining its ethically more-informed counterpart "the harm arising, given the disclosure". For example,

obtaining a list of persons aged 35–44 years from a confidential census file is generally less likely to generate grave harms than obtaining a list of persons from the same source showing specific ancestry, ethnicity, race, religion categories, etc.

It should be noted, however, that most laws protecting statistical confidentiality only safeguard against the misuse of microdata (that is, the individual record for each covered unit in the population). Other kinds of safeguards, including ethics, must be relied on to protect against the misuse of mesodata (that is, small area tabulations that may be used to target vulnerable population subgroups for human rights abuses) [5].

As a practical matter, there is often little to be gained by attempting to circumvent statistical confidentiality protections. Population data systems (for example, population censuses) used to gather data for statistics are usually very poor surveillance tools. The definitions and concepts used are often at variance with commonly understood meanings and the reporting errors common to any large-scale population data collection operation will give rise to many false positives and false negatives for any given population group of interest, seriously compromising the quality of any resulting risk assessments.

Responsibility

While some statisticians and other technical personnel involved in the work on risk assessment may at first consider that the ethical character of the intermediary and final uses of the data, estimates, and analysis on which they work are beyond their concern (see, for example, the discussion in Sigma Xi [6, pp. 52–54]), they should understand that legally and morally they are considered responsible for their actions under widely accepted current standards. This was the judgment of the Nuremberg Tribunal [7, p. 267] and is reflected in a number of relevant ethical norms. For example, Association for Computing Machinery's Code of Ethics and Professional Conduct begins with the following "moral imperative":

Contribute to society and human well-being. This principle ... affirms an obligation to protect fundamental human rights and to respect the diversity of all cultures ... Well-intended actions, including those that accomplish assigned duties, may lead to harm unexpectedly. In such an event the responsible

person or persons are obligated to undo or mitigate the negative consequences as much as possible [2, Sections 1.1 and 1.2].

The ASA's ethics guidelines are equally unambiguous:

Statistical tools and methods, like many other technologies, can be employed either for social good or for evil. The professionalism encouraged by these guidelines is predicated on their use in socially responsible pursuits by morally responsible societies, governments, and employers. Where the end purpose of a statistical application is itself morally reprehensible, statistical professionalism ceases to have ethical worth [1, Section 1.B].

Of course, decisions about the moral worth of a given application may require thoughtful discernment. However, the moral worth must be examined and not be simply assumed.

Promoting Ethical Work

Clearly a first, important step for sound statistically based work on risk assessment in the context of homeland security and counterterrorism (*see Counterterrorism; Sampling and Inspection for Monitoring Threats to Homeland Security*) (that is, work that is valid, ethical, and responsible) is the development of written statistical standards. Such standards are important not only in promoting sound work by having agreed-on protocols for reaching conclusions where quantitative data play a major role, but also by providing an opportunity for both statisticians and others to agree on such protocols independently from an actual crisis-laid situation and the tensions and emotions that decisions in such circumstances frequently evoke.

Such standards do not mean that mistakes will not be made either because the standards are ignored or that the standards adopted were defective in some essential respect. (For example, they were insufficiently specific or structured on incomplete or faulty assumptions). One way of improving the quality of such standards is to provide as wide a scrutiny of these standards as possible at various stages in their development. Despite their possible shortcomings, such agreed-on written standards provide some authority that quantitative analysts can cite in describing the basis for their conclusions in the face of those holding alternative views. It is also one way

of promoting the use of the best statistical methods in work on risk assessment.

Second, those engaged in risk assessment in any field, particularly those involved in work related to homeland security and counterterrorism, should routinely participate in ethical reviews of their work. To be effective, such reviews must also involve at least some persons with no intellectual, career, institutional, or personal stake in any specific outcome of the review. In this regard, many of the major past ethical failures have arisen in situations where none of those involved recognized that what they were doing posed an ethical problem (see, for example, [8, 9]). Accordingly, ethical reviews and discussions of ethics are one of the best preventive measures to take against serious ethical problems.

References

- [1] American Statistical Association (ASA) (1999). *Ethical Guidelines for Statistical Practice*, Alexandria, at <http://www.amstat.org/profession/index.cfm?fuseaction=ethicalstatistics>.
- [2] Association for Computing Machinery (ACM) (1992). *ACM Code of Ethics and Professional Conduct*, at <http://www.acm.org/constitution/code.html>, last updated 2003.
- [3] International Statistical Institute (ISI) (1986). Declaration of professional ethics for statisticians, *International Statistical Review* **54**, 227–247, at <http://www.cbs.nl/isi/ethics.htm>.
- [4] United Nations Economic and Social Council (1994). *Report of the Special Session of the Statistical Commission*, New York, April 11–15, E/1994/29, at <http://unstats.un.org/unsd/methods/statorg/FP-English.htm>.
- [5] Seltzer, W. (2005). Official statistics and statistical ethics: selected issues, *Proceedings of the 55th Session of the International Statistical Institute*, Sydney, April 5–12, 2005, Session # IPM 81, International Statistical Institute, The Hague, CD disk, at <http://www.uwm.edu/~margo/govstat/integrity.htm>.
- [6] Sigma Xi (1999). *The Responsible Researcher: Paths and Pitfalls*, Research Triangle Park, at <http://www.sigmaxi.org/programs/ethics/ResResearcher.pdf>.
- [7] Caplan, A.L. (1992). The doctor's trial and analogies to the Holocaust in contemporary bioethical debates, in *The Nazi Doctors and the Nuremberg Code: Human Rights in Human Experimentation*, G.J. Annas & M. Grodin, eds, Oxford University Press, New York, pp. 258–275.
- [8] Jones, J.H. (1981). *Bad Blood: The Tuskegee Syphilis Experiment*, The Free Press, New York.
- [9] Annas, G.J. & Grodin, M. (eds) (1992). *The Nazi Doctors and the Nuremberg Code: Human Rights in Human Experimentation*, Oxford University Press, New York.

Related Articles

Counterterrorism

WILLIAM SELTZER AND MARGO ANDERSON

Evaluation of Risk Communication Efforts

Evaluation is a critical, yet often overlooked, component of any risk communication effort. Evaluation has been defined as a “purposeful effort to determine effectiveness” [1]. Without a systematic plan for evaluation, the communicator has no way of knowing if risk communication activities have reached the intended audience, have been communicated effectively, or have inspired behavior change or other actions.

Increasing interest in and application of risk communication in the last 25 years (*see Risk and the Media*) has been accompanied by the awareness of the importance of evaluating risk communication efforts [2–7]. Information and guidance has been provided on the use of evaluation in generalized risk communication program development [1, 3], message development [4], efficacy of different communication models for evaluation [8–10], assessment of agency effectiveness [11, 12], public participation programs and citizen advisory committees [13, 14], and community-based research [15]. This paper will look at the compelling reasons to conduct evaluation, the various types of evaluation, some established key factors in successful evaluation, and some of the barriers to conducting evaluation.

Why Evaluate?

The overall goal for conducting an evaluation is to improve communication activities. However, evaluation serves many purposes in risk communication efforts. Chelimsky [16] grouped different evaluation purposes into three general perspectives. Evaluation for *accountability* involves the measurement of results or efficiency, and usually involves the provision of information to decision makers to determine if the anticipated changes did or did not occur. Evaluation for *development* is done to improve institutional performance. Evaluation works both prospectively and retrospectively to help determine communication agendas, improve communication strategies, develop indicators of institutional effectiveness and responsiveness, audience response to the agency and programs, and whether resources are being expended

wisely. Finally, evaluation for *knowledge* is used to acquire a more profound understanding and explanation of communication processes and mechanisms. This type of evaluation contributes to the collective wisdom communicators may draw upon in continuously perfecting their craft.

Within these three perspectives, well-planned and effective evaluation of risk communication efforts serves many specific and valuable purposes:

To ensure the right problem is being addressed

Problem formulation is the most critical stage of any communication effort. However, like many problem-solving exercises, communication strategies frequently fail because of the following [17]:

- They are solving the wrong problem
For example, trying to develop a communication program to alleviate concerns about a potential health risk when the community is actually concerned about lack of control in decisions about the imposed risk or does not rate the problem as a priority.
- They are stating the question so that it cannot be answered
For example, trying to solve the question “How can we communicate this risk so that people will understand and accept our decision” is doomed to failure – communicating a predetermined decision, particularly in a controversial issue, will not ensure public acceptance, regardless of the communication strategy employed.
- They are solving a solution
As in the previous example, a one-way communication of an already determined risk-management solution does not address the risk problem – this can only be achieved through reciprocal communication and involvement of those potentially affected.
- They are stating questions too generically
For example, many agencies initiate a communication program with the general goal of improving communication, then cannot evaluate success because of the lack of clear, measurable, and relevant objectives.
- They are trying to get agreement on the answer before there is agreement on the question

For example, trying to obtain public consensus on options to reduce the risk from a new industrial facility, when the real question from the public's perspective is whether it should even be sited in that location.

Effective risk communication requires that all interested and affected parties be involved in the early stages of problem formulation and solution development for environmental risk problems [18, 19]. Periodic evaluation to check that the issues of concern to all parties are being addressed ensures that the process will be meaningful and resources will be expended appropriately.

To ensure the needs of all interested and affected parties are being met

Does the process have the "right participation", i.e., have those interested and affected parties who should be involved in the process been correctly identified? Has the target audience for the communication been correctly identified? Has the audience been correctly characterized? Evaluation is also important to ascertain if the process has the "participation right". Is the process responsive to the needs of the interested and affected parties? Have their information needs, viewpoints, and concerns been adequately represented? Have they been adequately consulted? If the participation is wrong and the needs of the audience are not being met, the communication program cannot be successful [18].

To ensure that the program objectives are being met

Objectives should be evaluated throughout the communication program to assess the following:

1. Are the objectives appropriate?
2. Will the plans meet the objectives?
3. Have the objectives been achieved?

To increase institutional understanding of and support for communication efforts

Communication, if conducted successfully and in an open and inclusive manner, should hopefully result in a productive means of addressing conflict and outrage. Unfortunately, many agencies do not readily recognize the value of programs that are "avoiding" the creation of problems. Evaluation provides the opportunity to impress upon senior managers the

value of well-planned and executed communication programs [1, 20].

To increase public understanding and support for communication efforts

Communication participants and audiences frequently fail to recognize the importance and relevance of the communication program until required to examine the process and outcomes.

To enhance and promote public participation in the risk process

Incorporating public participation into a process that is seen by many (public and risk managers alike) as being strictly based on scientific concerns is often problematic. Evaluation can serve to illustrate the value of participation to all skeptics.

To make optimum use of limited resources

Evaluation at the beginning and during the communication effort allows for programs to be modified before limited resources have been inappropriately expended.

To provide evidence of the need for additional funds or other resources

Evaluation is a valuable tool to assess the adequacy of resources (funds, personnel, and time) required to implement the communication program and keep it on track [1, 20].

To avoid making the same mistakes in future communication efforts

Evaluation ensures that continuous learning occurs throughout the process and that each subsequent program is improved accordingly [1].

Types of Evaluation

Evaluation should occur throughout the communication process. Three types of evaluation are generally recognized as being applicable to communication programs: formative evaluation, process evaluation, and outcome evaluation.

Formative Evaluation

Formative evaluation occurs early in the program, as objectives are selected, audience information needs

identified, and the communication strategy planned. Formative evaluation assesses the appropriateness of potential objectives, and the strengths and weaknesses of alternative communication strategies. Evaluation at this stage permits the necessary revisions and modifications to be made before program planning efforts are initiated.

Formative evaluation should determine if the problem analysis is sufficient, if the plans will meet the objectives, and if the plans will meet the audience needs [20]. Formative evaluations can also be used to pretest potential message materials for clarity, tone, and comprehensiveness [21]. While formative evaluation does not guarantee the success of a communication program, it does minimize the possibility that a program will fail owing to developmental flaws [22].

The type of formative evaluation used depends on whether it is conducted for problem analysis, audience assessment or communication strategy design. The scope of the communication program and available resources will also dictate the type of evaluation conducted. Formative evaluations can take the form of a simple pretest of materials, group “brainstorming” or discussions, surveys or focus groups, and personal interviews.

Process Evaluation

Process evaluation occurs during the implementation of the communication strategy, and is used to assess if the program is progressing as planned. Process evaluation can be used to modify the communication during implementation of the process, thus ensuring the resources are being effectively used. This type of evaluation is used to assess whether [22]

- activities are on track and on time;
- the target audience is being reached and understands the information;
- some strategies appear to be more successful than others;
- some aspects of the program need more attention, alteration or replacement;
- resource expenditures are acceptable and within budget; and
- if the budget was realistic to meet the stated problem and objectives.

The type of process evaluation employed will depend on the scope of the program and the available

resources. Process evaluation can range from routine record keeping of activities to a program checklist or even a management audit. Informal evaluation mechanisms, such as audience feedback or communicator observations, may also be useful in some cases. The most effective and efficient means of conducting a process evaluation is to make it a regular part of communication program planning and implementation. Process evaluation should be an ongoing activity in both new and established environment and health communication programs.

Outcome and Impact Evaluation

Also sometimes termed summative and/or impact evaluation, outcome evaluation occurs at the end of the program. It is used to assess to what extent the objectives were achieved, and the short- and/or long-term impacts of the communication program. The typical outcomes that are assessed in communication programs include changes in awareness, knowledge, behaviors, and beliefs and attitudes among target audiences [20]. Outcome evaluation is most effective when the program has clear and measurable goals and consistent implementation [1].

Outcome evaluation techniques can include focus groups, interviews, or questionnaires administered in person, by mail or by telephone [22]. Evaluation can also be done through activity assessments (e.g., demographics of callers to a hotline) or print media review (e.g., monitoring of content of articles appearing in the media) [1].

Impact evaluation is sometimes distinguished from other forms of outcome evaluation based on the scope and long-range timeframe of the impacts. This type of evaluation may be used to assess effectiveness of a program after a certain time period has elapsed. While an outcome evaluation can assess the audience’s perceived change in awareness or understanding immediately after the communication event, an impact evaluation can assess if that change was sustained over a period of time and affected practice or behavior. Impact evaluation can also assess the extent to which a program contributed to long-term, more global changes in health status. As these changes are the result of a multitude of factors, it is difficult to directly relate them to a specific program or activity. Information obtained from an impact evaluation may include such things as changes

in morbidity or mortality, or long-term maintenance of changed behavior [21].

Key Factors for Successful Evaluation

Evaluation cannot be “tacked on” to the end of a risk communication program. Effective and meaningful evaluations must be a carefully considered and planned component of risk communication activities. There are several key factors that should be considered in planning for evaluation [20, 23].

Organizational prioritization and commitment of resources

Decision makers within organizations must be committed to evaluation throughout the lifespan of a program (including conceptualization, implementation, and refinement). They must also be willing to commit the necessary staff time, budget, and other resources to evaluation as an integral part of the program. Evaluation cannot be meaningful unless appropriate resources are available to adequately plan and implement evaluation as an integral part of the communication effort.

Evaluation planning

Programs should be planned in advance of the communication program. Planning should draw upon multiple resources, including all interested and affected parties, conceptual theories, and literature reviews.

Specifying clear, measurable objectives

One of the most common mistakes in planning risk communication programs is failure to consider the specific outcomes to be achieved. If you do not know what you wanted to achieve how can you possibly decide if you have been effective? Program objectives should reflect the outcomes that will be measured during the outcome evaluation.

Measuring target audience characteristics at the outset to set the baseline for evaluating changes

Frequently, communication programs are designed to effect a change in audience awareness, knowledge, understanding, or behavior. However, it is impossible to assess if changes in knowledge or behavior have occurred if there is no baseline information on prior knowledge or behavior against which such comparisons can be made.

Developing clear plans specifying activities, audiences, staff involvement, budget, etc. to provide a basis for comparison

Process evaluations conducted during program implementation must also be compared against planned activities and intended resource allocations. However, these plans must remain flexible if process evaluation is to be effective.

Including appropriate data-gathering activities throughout the implementation of the program

This will ensure that the information necessary for evaluation is available when needed. Process and outcome evaluation are often limited by a lack of specific data required to conduct the evaluation.

Staff training and education

Communicators are not necessarily conversant in evaluation concepts or methodology. Training programs should be supplemented by specific discussions on planning, data analysis, interpretation, and dissemination for each specific program.

Reporting of evaluation research

Frequently, evaluation efforts are not reported, or are only reported in government reports or “grey literature” where they are not readily available as a means of learning from the experiences of others. In addition, program planners require simple and usable information from evaluations if these are to be successful in modifying ongoing programs or changing future efforts. The interpretation and presentation of evaluation efforts need to reflect this simplicity and direct approach.

Barriers to Risk Communication Evaluation

The factors that contribute to successful evaluation are also those that most frequently prove to be the most difficult obstacles. Internal organizational characteristics create the most common constraints to undertaking the tasks necessary for optimal evaluation. Management perceptions regarding the value of evaluation, and the level of support they subsequently provide for well-designed evaluation activities, are major barriers. Limited resources, such as funds and staff time, greatly restrict evaluation activities. Programs are also sometimes designed and implemented

under very restrictive timelines, thus precluding the necessary planning for evaluation.

Difficulties are often encountered in gathering the necessary information from the audience(s) on precommunication attributes and postcommunication assessment of changes. Audiences may be apathetic about evaluation efforts, or may try to please by providing the answers they think are desired. Evaluation efforts may also inadvertently become platforms for expressing dissatisfaction with other issues related to the risk or agency involved. In some jurisdictions, there may even be policies limiting the ability to gather information from the public.

External factors may create constraints to evaluation efforts. In communication programs with multiple partners, the logistics of evaluation are often difficult to coordinate successfully. There may also be difficulties in defining or establishing consensus between agencies regarding the objectives of the program [22].

The evaluation tools available to practitioners and their ability to employ those tools may limit evaluation implementation and effectiveness. Deciding on appropriate methodologies (such as a written questionnaire, individual interviews, or communicator observations) and then designing appropriate measures for that tool require that the communicator have both evaluation training and previous experience in diverse situations. It is also often difficult to separate the effects of program influences from other influences on the target audience in “real-world” situations [22].

Finally, despite best intentions, evaluation results are often not incorporated optimally into program planning and implementation. Some of the reasons suggested for lack of incorporation are as follows [23]:

- organizational inertia (organizations are slow to change);
- methodological weakness (poorly conducted studies have limited credibility);
- design irrelevance (evaluations conducted without input from program planners, decision makers, or the community); and
- lack of active or appropriate dissemination (evaluation results are either not distributed or not tailored to the needs of those in the program).

Summary

While it is acknowledged that evaluation plays a critical role for the communication components of all risk-management processes, it is not consistently applied in risk communication efforts. We need to learn more effectively from our efforts and apply these lessons to future communications, or we will be doomed to forever repeating the same mistakes.

References

- [1] Regan, M.J. & Desvousges, W.H. (1990). *Communicating Environmental Risks: A Guide to Practical Evaluations*, EPA 230-01-91-001, U.S. Environmental Protection Agency, Washington, DC.
- [2] U.S. National Research Council. (1989). *Improving Risk Communication*, National Academy Press, Washington, DC.
- [3] Fisher, A., Pavolva, M. & Covello, V. (eds) (1991). *Evaluation and Effective Risk Communications Workshop Proceedings*, Pub. No. EPA/600/9-90/054, U.S. Environmental Protection Agency, Washington, DC.
- [4] Weinstein, N. & Sandman, P.M. (1993). Some criteria for evaluating risk messages, *Risk Analysis* **13**(1), 103–114.
- [5] Chess, C., Salomone, K.L., Hance, B.J. & Saville, A. (1995). Results of a national symposium on risk communication: next steps for government agencies, *Risk Analysis* **15**(2), 115–125.
- [6] Chess, C., Salomone, K.L. & Hance, B.J. (1995). Improving risk communication in government: research priorities, *Risk Analysis* **15**(2), 127–135.
- [7] Gerrard, S. (1999). Learning from experience: the need for systematic evaluation methods, in *Risk Communication and Public Health*, P. Bennet & K. Calman, eds, Oxford University Press, Oxford, pp. 254–266.
- [8] Niewöhner, J., Cox, P., Gerrard, S. & Pidgeon, N. (2004). Evaluating the efficacy of a mental models approach for improving occupational chemical risk protection, *Risk Analysis* **24**(2), 349–361.
- [9] MacGregor, D.G., Slovic, P. & Morgan, M.G. (1994). Perception of risks from electromagnetic fields: a psychometric evaluation of a risk-communication approach, *Risk Analysis* **14**(5), 815–828.
- [10] Bostrom, A., Atman, C.J., Fischhoff, B. & Morgan, M.G. (1994). Evaluating risk communications: completing and correcting mental models of hazardous processes, part II, *Risk Analysis* **14**(5), 789–798.
- [11] Tinker, T.L., Collins, C.M., King, H.S. & Hoover, M.D. (2000). Assessing risk communication effectiveness: perspectives of agency practitioners, *Journal of Hazardous Materials* **B73**, 117–127.
- [12] Johnson, B.B. & Chess, C. (2006). Evaluating public responses to environmental trend indicators, *Science Communication* **28**(1), 64–92.

- [13] Chess, C. (2000). Evaluating environmental public participation: methodological questions, *Journal of Environmental Planning and Management* **43**(6), 769–784.
- [14] Lynn, F.M. & Busenberg, G.J. (1995). Citizen advisory committees and environmental policy: what we know, what's left to discover, *Risk Analysis* **15**(2), 147–162.
- [15] Gibson, N., Gibson, G. & Macaulay, A.C. (2001). Community-based research: negotiating agendas and evaluating outcomes, in *The Nature of Qualitative Evidence*, J. Morse, J. Swanson & A.J. Kuzel, eds, Sage Publications, Thousand Oaks, pp. 161–186.
- [16] Chelimsky, E. (1997). The coming transformation in evaluation, in *Evaluation for the 21st Century: A Handbook*, E. Chelimsky & W.R. Shadish, eds, Sage Publications, Thousand Oaks, Chapter 1, pp. 1–26.
- [17] Bardwell, L.V. (1991). Problem-framing: a perspective on environmental problem-solving, *Environmental Management* **15**, 603–612.
- [18] U.S. National Research Council (1996). *Understanding Risk: Informing Decisions in a Democratic Society*, National Academy Press, Washington, DC.
- [19] U.S. Presidential/Congressional Commission on Risk Assessment and Risk Management (1997). *Framework for Environmental Health Risk Management*, Final Report, Vols. 1 and 2, Washington, DC.
- [20] U.S. Environmental Protection Agency (1995). *Guidance for Assessing Chemical Contaminant Data for Use in Fish Advisories*, EPA 823-R-95-001, Volume IV; *Risk Communication*, Office of Water, Washington, DC.
- [21] U.S. National Cancer Institute (1992). *Making Health Communication Programs Work: A Planner's Guide*, NIH Publication 92–1493, National Cancer Institute, Washington, DC.
- [22] Arkin, E. (1991). Evaluation for risk communicators, in *Evaluation and Effective Risk Communications Workshop Proceedings*, Pub. No. EPA/600/9-90/054, A. Fisher, M. Pavolva & V. Covello, eds, U.S. Environmental Protection Agency, Washington, DC, pp. 17–18.
- [23] Flora, J.A. (1991). Integrating evaluation into the development and design of risk communication programs, in *Evaluation and Effective Risk Communications Workshop Proceedings*, Pub. No. EPA/600/9-90/054, A. Fisher, M. Pavolva & V. Covello, eds, U.S. Environmental Protection Agency, Washington, DC, pp. 33–40.

Related Articles

Stakeholder Participation in Risk Management Decision Making

CYNTHIA G. JARDINE

Experience Feedback

Experience feedback (*see Reliability Data; Syndromic Surveillance; No Fault Found*) is the process by which information on the results of an activity is fed back to decision makers as new input to modify and improve subsequent activities. It plays a central role in the management systems for the prevention of accidents, incidents, and errors, since these events are symptoms of underlying weaknesses and indicate scope for improvements within the system where they occur [1].

Seen from this perspective, a risk assessment is a process by which knowledge gained from experience of historic industrial systems is transferred to a risk assessment team and used as a basis for recommendations on needs of risk-reducing measures in a system being analyzed (analysis object). The aim is to identify and mitigate possible future accidents in order to reduce the risk to an acceptable level. The object of the analysis may be an existing industrial system, a new system, or a system being modified. The necessary transfer of knowledge is accomplished through different channels, (Figure 1) [2].

Explicit and Tacit Knowledge

Figure 1 distinguishes between explicit and tacit knowledge, and both are experience based [3]. Explicit knowledge is rational and can be expressed in text, numbers, and formulas. It includes theoretical approaches, problem-solving heuristics, manuals, and databases. Tacit knowledge, on the other hand, is personal, subjective, and context related and is not directly suited for documentation. It includes value systems, images, intuition, and mental models and know-how. Some channels are dominated by the transfer of explicit knowledge, and here records of accident and error data play an important role. Other channels involve participation by personnel at the sharp end in the risk assessment team. This makes it possible to tap their tacit knowledge gained through day-to-day experiences with incidents and errors.

Experience Databases and Experience Carriers

Experience databases are representations of explicit knowledge and play an important role in quantitative

risk assessments. They are used in the identification of hazards and in the modeling and estimations of frequency and consequences of unwanted events (Figure 2). Examples of experience databases are as under:

- incidents and accidents databases with data from investigation reports of relevant systems;
- reliability databases with failure data for hardware and software for relevant systems;
- human reliability databases with data on human errors in different types of tasks;
- exposure databases with data on the extent of the operations for which accident, incident, and failure data is collected (e.g., number of hours of operation, number of times an activity has been carried out);
- data on escalation probabilities, e.g., likelihood of ignition after gas leakage;
- site specific population density statistics for the surroundings of the analysis object in question;
- site specific meteorology data.

Experience carriers represent explicit knowledge at a higher level of synthesis than experience databases. Here experience from operations and accidents on historic industrial systems and from earlier risk assessments have been summarized and codified into standardized work processes, decision criteria, analytic models etc. in a way that makes sense to the risk analyst. Examples of experience carriers directly used in risk assessments are as follows (Figure 2):

- *Regulatory requirements to risk assessments*

Examples of these are the provisions in the EU Directive on the control of major-accident hazards (Seveso Directive), the US OSHA Standard 1910.119 on the process safety management of highly hazardous chemicals, and the UK Offshore Installations (Safety Case) Regulations [4–6]. Work on the EU Directive was initiated after a major chemical accident in Seveso, Italy in 1976. The recommendations in Lord Cullen's investigation report on the Piper Alpha disaster in the North Sea in 1988 were followed-up in the UK Safety Case Regulations [7].

- *Standards and handbooks on principles and practice in risk assessments*

They include knowledge of relevance to the risk analyst community such as analytic models, computer

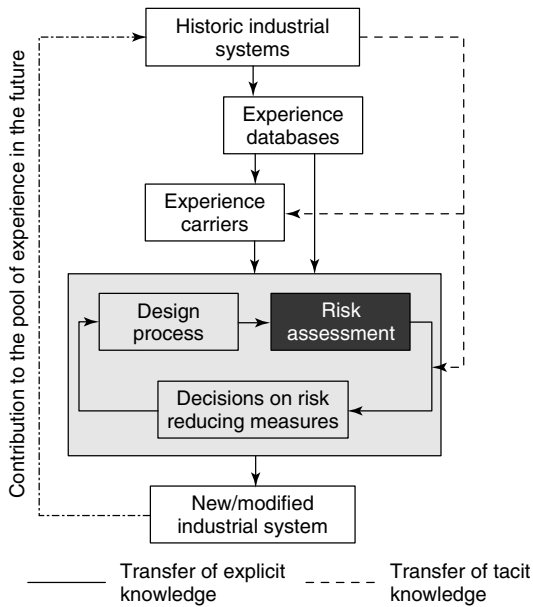


Figure 1 Channels for transfer of experienced-based knowledge in risk assessments

codes for simulation, risk assessment methods, and experience checklists. One example is the Norsok standard Z-013 on risk and emergency preparedness

analysis used by the Norwegian offshore industry [8]. Another example is the IEC standard 61508 on functional safety of electrical/electronic/programmable electronic safety-related systems, used to specify requirements to safety instrument systems using a risk-based approach [9]. These standards are typically developed by expert panels with extensive risk assessment experience. They give detailed accounts on how to execute risk assessments.

- *Decision criteria regarding highest tolerable risk*

A risk level above this limit will trigger requirements to implement risk-reducing measures. The limit is usually based on statistics from historical systems on frequencies and consequences of accidents. Such statistics gives indications of risk levels that have been possible to accomplish in the past and stakeholders' tolerance to accidental losses. Similarly, there are decision criteria on the lowest risk level that is reasonably practicable, below which further risk reduction is deemed unnecessary.

- *Phenomenological knowledge about accident scenarios, failure mechanisms, human tolerance limits, performance barriers*

This type of knowledge is typically documented in scientific and professional articles, textbooks, and

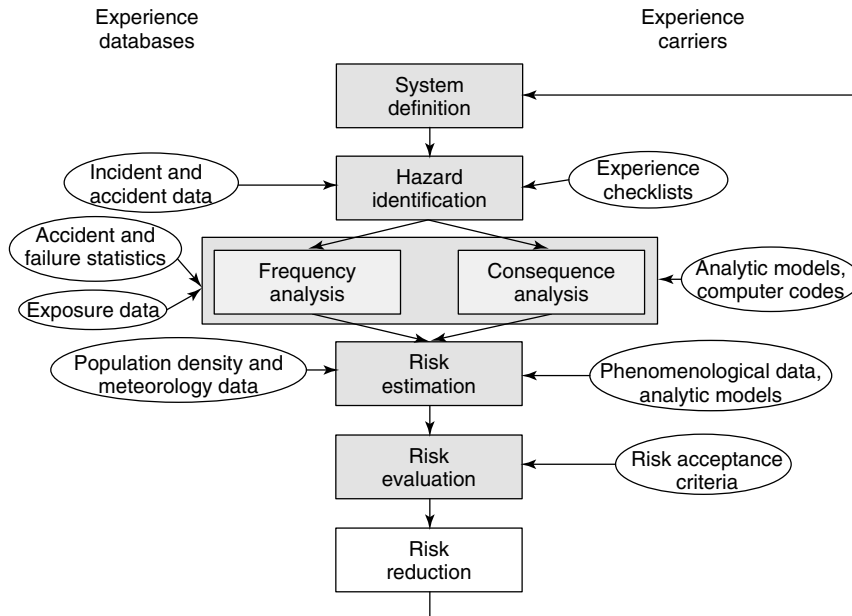


Figure 2 Use of explicit knowledge in risk assessments

handbooks. Since major accidents are rare events, experience data from in-depth accident investigations on the latter part of the accident scenario determining the extent of the consequences and efficiency of barriers is often insufficient. Knowledge, here, is more often based on laboratory and field experiments and simulations of physical processes and on expert judgment.

In the follow-up of the risk assessments, experience carriers documenting risk-reducing measures such as regulations, standards, and guidelines are used to reduce the risk to a tolerable level.

Update with New Experience

For the risk assessment results to be valid, the experience basis needs to be kept updated. The generic incident and failure rates used in risk assessments are based on statistics from the operation of industrial systems in the past [10]. By aggregating data from systems with similar characteristics to the analysis object, a generic average for this type of system is achieved. The question arises on how representative this generic average is for the new industrial system under study with respect to state of the art in technology, operations, and maintenance and relevant accident risks. The historic incident and failure experience has already been used to develop preventive measures, and some of the accident risks represented in the data may not be valid for the analysis object. There is a trade-off between the needs

of representative data on the one hand, and sufficient data with respect to statistical uncertainty on the other. Since accidents are rare events, this trade-off is especially critical for accident frequencies. Trend analysis and expert judgment are sometimes used to establish new frequency estimates to compensate for the inadequacies in experience update.

Updates of regulations and standards often involve bureaucratic and long-lasting procedures. Companies may prefer to develop their own experience-based best practice as a supplement to these documents [11]. The feedback loop here is usually shorter and less delayed than is possible in the update of regulations and standards.

Nonaka's theory of organizational knowledge creation may be used to illustrate how the company's risk analyst community interacts to develop such best practice [12]. The risk analyst's individual knowledge is experience based and acquired through participation in projects where risk assessments are conducted (Figure 3). For a best practice to develop, the individual and mainly tacit knowledge needs to be communicated and discussed within a community of colleagues. This discussion results in the development of shared knowledge that is made explicit and documented. This new knowledge has to be justified to become best practice. This is a management task and takes place in the functional organization. Here the proposed best practice is judged in terms of efficiency and effectiveness and compliance with the legal environment and industry standard. The new best practice is then implemented in the functional

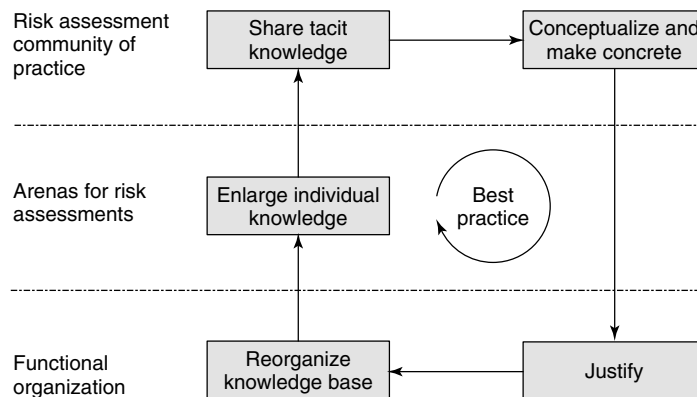


Figure 3 Model of the organizational knowledge-creation process [Reproduced from [10]. © Taylor & Francis Group, 2006.]

organization's knowledge base for use in future risk assessments.

Risk Assessment as an Arena for Experience Exchange

Similar experience exchange and learning processes take place within the risk assessment team. Dependent on the scope and aim of the risk assessment, the team needs to include internal stakeholders with an adequate variety of knowledge and values to promote validity and reliability of results, learning among team members, and acceptance of decisions made by the team. The assessment team could, for example, be made up of systems designers, risk analysts, and personnel at the sharp end with operational experience from similar systems. Operational experience is especially important in judging the relevance of assumptions in the risk assessment and in modeling the early part of the accident sequence, where production disturbances and human and technical errors play an important role. Biases and misconceptions may be avoided, if the group members challenge propositions and discuss alternative interpretations.

Documentation of the Experience Basis

Risk assessments are decision support tools used in judgments of needs of risk-reducing measures. For the decision makers to trust the validity and reliability of the results, the documentation of the assessment needs to be comprehensive, well structured, and transparent [13]. It is thus necessary to document the sources of the experience input to the analysis and their relevance. This includes the databases on accidents and failures that have been used and the experience represented in the analysis team.

Example: Experience Feedback in Environmental Risk Assessment

The oil companies on the Norwegian Continental Shelf carry out environmental risk assessments as part of the permitting process for drilling operations. They apply the results to verify that the acceptance criteria for environmental risk owing to oil spills are met, and to select and dimension safety and emergency preparedness measures. A community of

practice made up of environmental, drilling, and risk assessment experts from Oil Companies and Consultants has cooperated in developing guidelines for the environmental risk assessments [14]. The work on this experience carrier started as a result of new safety legislation in the beginning of the 1990s. It has evolved through different revisions based on experience from previous risk assessments, and represents the recognized best practice of oil spill risk assessment in the Norwegian oil industry. The guidelines standardize input data, models, analysis methods, and risk acceptance criteria, and make it possible to compare results between companies.

When work on environmental risk assessments started, there was a general lack of phenomenological data, e.g., on the distribution of sea birds in open sea. The first generation of data and models used in the risk assessments had a low fidelity compared to the observed variations in distributions in time and space. Improved databases and new simulation models, developed through an interaction between research and risk assessment practice, now map these variations with sufficient fidelity.

Another example of the results of the knowledge-creation process is the development of so-called key parameters for environmental damage. In the absence of species-specific data, the community of practice, has decided to develop and use standardized key parameters for the expected restitution time for a population of sea birds or mammals following a reduction in the population due to acute discharge of oil to sea. An acute discharge resulting in a reduction of the population by up to 5%, for example, is expected to give a restitution time of less than 1 year in 50% of the events and 1–3 years in the remaining events.

Blowout data is required to carry out environmental risk assessments of offshore drilling operations. To meet this need, SINTEF, in Trondheim, has established an Offshore Blowout Database [15]. Input data on blowout incidents and exposure (number of drilled wells, etc.) are derived from public sources such as the US Mineral Management Service, the Health and Safety Executive in the United Kingdom, and the Petroleum Safety Authority in Norway. Complementary information on blowout incidents has been secured, where possible, through direct contacts with the Oil Companies or drilling Contractors in question. The next step has been to develop recommended

blowout frequencies as basis value in risk assessments of well operations of North Sea standard [16]. The recommended frequencies are updated regularly on the basis of new experience. The following measures have been introduced to ensure a sufficiently large database for frequency estimates, while, at the same time, ensuring that the recommended frequencies are representative to North Sea operations:

- The data used for the frequency estimates are limited to the North Sea, the US Gulf of Mexico, and the Canadian East Continental shelf, where drilling operations are carried out according to a similar safety standard.
- “Irrelevant data”, i.e., incidents that are not relevant with respect to equipment, requirements, or practices in the North Sea, are omitted.
- Weighted averages for the last 20 years are used, where newer data are given a higher weight in the calculations of frequencies.

Acknowledgment

Thanks are due to Jan Lund, Scandpower and Erik Odgaard, Norsk Hydro, for comments on earlier versions of the manuscript to this article. Thanks are also due to Jon Rytter Hasle, Norsk Hydro, Odd Willy Brude, DnV, and Per Holand, ExproSoft for valuable input to the article.

References

- [1] Kjellén, U. (2000). *Prevention of Accidents through Experience Feedback*, Taylor & Francis, London.
- [2] Kjellén, U. (2002). Transfer of experience from the users to design to improve safety in offshore oil and gas production, in *System Safety – Challenges and Pitfalls of Intervention*, B. Wilpert & B. Fahlbruch, eds, Elsevier, Oxford, pp. 207–224.
- [3] Nonaka, I. & Takeuchi, H. (1995). *The Knowledge-Creating Company*, Oxford University Press, New York.
- [4] European Council (1982/96). *Directive 96/82/EC on the Control of Major-accident Hazards (Seveso II Directive)*, Official Journal of the European Union, Brussels.
- [5] OSHA (1998). *Process Safety Management of Highly Hazardous Chemicals*, OSHA Standard 1910.119, Washington.
- [6] Crown Copyright (1992, 2005). *The Offshore Installations (Safety Case) Regulations*, Statutory Instrument 1992 No. 2885/2005 No. 3117, London.
- [7] Cullen, W.D. (1990). *The Public Inquiry into the Piper Alpha Disaster*, HMSO, London.
- [8] Standard Norway (2001). *Risk and Emergency Preparedness Analysis*, Norsk Standard Z-013, Oslo.
- [9] International Electrotechnical Commission (1998). *Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems*, Standard IEC 61508, Geneva.
- [10] Modarres, M. (2006). *Risk Analysis in Engineering: Techniques, Tools and Trends*, Taylor & Francis, Boca Raton.
- [11] Kjellén, U. (2004). Improving knowledge sharing and learning in an organisation of safety, health and environmental project engineers, in *How to Manage Experience Sharing: from Organisational Surprises to Organisational Knowledge*, J.H.E. Andriessen & B. Fahlbruch, eds, Elsevier, Oxford, pp. 69–92.
- [12] Von Krogh, G., Ichijo, K. & Nonaka, I. (2000). *Enabling Knowledge Creation*, Oxford University Press, Oxford.
- [13] Suokas, J. & Rouhiainen, V. (1993). *Quality Management of Safety and Risk Analysis*, Elsevier, Amsterdam.
- [14] The Norwegian Oil Industry Association (2005). *Environmental Risk – Guidelines for the Execution of Environmental Risk Assessments for Petroleum Activities on the Norwegian Continental Shelf (in Norwegian)*, The Norwegian Oil Industry Association, Stavanger.
- [15] Holand, P. (2006). Blowout and well release characteristics and frequencies, 2006, Report No. STF50 F06112, SINTEF Technology and Society, Trondheim.
- [16] Solberg, A. (2007). Blowout and well release frequencies – based on SINTEF Offshore Blowout Database, 2006, Report No. 80.005.003/2007/R3, Scandpower Risk Management, Kjeller.

URBAN KJELLÉN

Expert Elicitation for Risk Assessment

Risk assessment inherently involves estimation of subjective factors. While consequences of a particular hazard event may be reasonably well understood or predictable through physical models, the frequency of an event is often much harder to assess. The prevalence in the daily paper of “100-year floods”, surprise events, and unexpected system failures point out a fundamental limit of risk knowledge [1]. Even when risk frequency can be estimated by observations over time, these observations often cannot resolve fundamental uncertainty over whether that observed frequency is either stable, or could change in the future.

Expert opinion (see **Risk in Credit Granting and Lending Decisions; Credit Scoring; Operational Risk Modeling; Sampling and Inspection for Monitoring Threats to Homeland Security**) gathered in one guise or another, is often the only reasonable source of information available by which an estimate of future event frequencies can be made [2–4]. *Expert elicitation* (see **Bayesian Statistics in Quantitative Risk Assessment; Expert Judgment; Uncertainty Analysis and Dependence Modeling**) is the term for a set of methods and practices that have developed within the social sciences and statistics to gather risk information from subjective expert information in a way that (as much as possible, though never absolutely) minimizes common sources of bias and ambiguity [5].

Expert elicitation is a qualitative practice, rather than a discipline or tool. A variety of techniques exist, some of which have been evaluated under controlled tests. By and large, however, the results of such tests have revealed a number of basic caveats and biases in the ability of subjects (expert or not) to assess risk, rather than providing “the right way” to gather expert opinions.

At its heart, all attempts at expert elicitation reduce to a structured questioning strategy. Elicitation of expert opinion is most often conducted *via* face-to-face interviews with *subject matter experts*. However, a wide range of alternative elicitation strategies are also common, including group exercises, use of survey instruments, simulations and role playing, and direct observation of expert choice.

The development of a particular elicitation strategy, often in the form of a set of questions or protocol, called an *elicitation instrument*, requires a detailed understanding both of the risk subject matter and the capabilities, needs, and communication styles of the subject matter experts. Before embarking upon an effort to gather expert opinions on a particular risk, it is vital that, at a minimum, the following factors are taken into account:

- Is the elicitation instrument ethical? Is information being gathered in a way that protects the welfare of the expert subjects? In particular, while gathering information on medical risks, it may be necessary to submit the elicitation instrument for a human subjects committee review.
- What cultural, social issues may cause elicitation subjects to misunderstand, or misanswer the questions being asked?
- Is the elicitation instrument being applied to the right community of experts? How well do the questions probe issues about which the experts have knowledge and the ability to effectively assess the meaning of the risks being examined?
- Has the elicitation instrument, and the way questions are formulated, been crafted such that they obtain risk estimates that are as free from bias and ambiguity as possible?

Deciding upon the application of a specific expert elicitation strategy is highly driven by situational factors, rather than the application of a rule-set or “cook-book”. What follows, therefore is an attempt to distill from the literature, a set of models for different risk expert elicitation scenarios, along with an introduction to the various uses of risk elicitation that have been reported in the literature.

Models of Expert Elicitation of Risk

There are different kinds of risks and sources of risk. Approaches to expert elicitation of risks vary in the literature depending upon which of these different kinds of risks are being considered. Understanding which kind of risk, and the underlying type of mechanism behind it, are essential in creating an effective, suitable elicitation instrument. On the basis of the underlying mental or knowledge model being used to describe a risk, very different kinds of elicitation strategies may be employed, and the ability of

subject matter experts to provide useful *contributory expertise* would vary [6]. From the standpoint of developing an elicitation strategy, four different categories of risk origination need to be distinguished from one another: basic events, system processes, competitive games, and social negotiation.

Basic Events

The vast majority of expert elicitation case studies focus on risks that are to be characterized using a single probability distribution, or frequency estimate. From an elicitation standpoint, experts grappling with a basic event risk estimation problem are modeling risk as being generated by a single (though possibly internally complex) black-box process. Expert elicitation is used as a “second best” source in lieu of being able to observe the process long enough to obtain direct measurements for the characteristics of the unknown black-box function (central tendency, variance, etc.).

The goal of this type of expert risk elicitation is to discover a “best available estimate” for this potential distribution on the basis of the pattern of outcomes that have been observed so far, and obtain judgments from one or more experts for the characteristics of the actual black-box outcome distribution [7].

Expert elicitation of basic event modeled risks focus on gathering quantitative values. For example, an elicitation session could focus on trying to describe the probability of a jet engine turbine blade failure in terms of the most likely number that might occur over a particular period of time, as well as the least and most number of failures that are likely to be observed. From this information, a basic sense for an underlying distribution of turbine blade failure events can be constructed from the expert estimates. This distribution, however crude it may be, can then be described using a probability distribution function, or fuzzy set (*see Environmental Remediation; Dependent Insurance Risks; Early Warning Systems (EWSs) for Predicting Financial Crisis; Near-Miss Management: A Participative Approach to Improving System Reliability*), and used as part of any number of different risk analysis methods (fault tree (*see Systems Reliability; Imprecise Reliability; Fault Detection and Diagnosis*), simulation (*see Asset–Liability Management for Life Insurers; Equity-Linked Life Insurance;*

Markov Modeling in Reliability; Reliability Optimization), etc.).

Examples of such uses of expert risk elicitation include analyses from virtually every field of risk study; everything from nuclear power plant safety studies [8], to managing the risks of organ donation IT systems [9], to Search and Rescue resource allocation [10], to real estate markets [11], to climate change modeling [12].

In part, because of the prevalence of “basic event risks” in the expert elicitation literature, and in part, because this type of model for risk is the most verifiable *via* controlled experiments against observed risk data, a good deal is known about how and under what circumstances experts do, and do not, perform well [13, 14]. In broad-brush strokes, systemic errors in expert perceptions of risk can be divided among two main sources of biases: *cognitive biases* and *motivational biases* (*see Expert Judgment*).

Cognitive biases, the most rigorously studied category of bias from the standpoint of behavioral science experimentation, covers a host of different mechanisms. All of these mechanisms stem from basic cognitive and mental traits that seem to be part of the inherent limits on human risk reasoning. When constructing elicitation instruments, it is important to approach these biases with a degree of humility – for a variety of evolutionary reasons these biases seem to be a deeply embedded set mental behaviors for all people, and a questioner simply cannot ever be completely certain that the results are as unbiased as one hopes.

Availability bias reflects the observed tendency of unusual or memorable events to skew risk estimates in comparison to more mundane (but often far more prevalent) events. Availability bias is commonly seen in surveys of risk perceptions of air traffic safety following a well-publicized aviation accident [15, 16]. *Anchoring bias* is the tendency to latch onto an initial belief of the magnitude of a risk estimate, and insufficiently adjust this initial mental risk estimate as new information becomes available. Anchoring bias has been demonstrated in a number of controlled studies in which experts were presented with an initial arbitrary number prior to being asked to judge a specific risk [17]. In the real world, the classic strategy of letting the “rube” win the first couple games of roulette, demonstrates, in part, how anchoring bias can be manipulated. *Inconsistencies in probabilistic*

logic also are a common source of cognitive biases in risk elicitation [18]. These logic inconsistencies stem from a variety of memory, logic, and reasoning failures that all subjects (expert or not) seem to share. Common examples of inconsistencies in probabilistic logic include failing to remember that the sum of all probabilities assigned to the outcomes of a risk must sum to 1, or failing to assign a lower probability to the union of a set of independent events, than to any single member of that set.

In addition to biases caused by expert cognitive limitations, expert opinion may be biased by motivational effects – such as relationships, employment, and past experiences. Motivational biases can also unwittingly be introduced by the elicitation instrument *via interviewer bias*. Just as subject matter experts may have a tendency to misestimate risk frequencies on the basis of preconceptions and cognitive limits, so too may bias be introduced during interpretation based on what results the elicitation administrator thinks “make sense”, or from the administrator’s belief in a prior risk hypothesis. Likewise, there is the possibility that the elicitation instrument itself may be structured in a way that unintentionally causes bias. Even if the elicitation instrument is free of any bias held by the administrator, it may still be flawed, unclear, or structured in a way that will lead to unintended (and always very unwanted from the standpoint of the poor researcher) patterns in risk estimate results [19].

The most important mitigation strategy for all types of bias is to carefully craft, administer, and test the elicitation instrument. Whenever possible, the elicitation instrument should be edited by other researchers familiar with the kinds of bias common to expert risk elicitation/general survey analysis, and holders of subject matter contributory expertise who will *not* participate in the final elicitation. Common strategies to minimize bias include the following:

- ensuring that risk elicitation questions are nondirective, do not suggest quantitative anchor points either in the question or in any examples on how to answer the question, and do not ask responders to estimate one risk in relation to a risk estimated in a previous question;
- use of group risk elicitation exercises (*see Group Decision*) such as the Delphi method to control

against any single source of preconception, limitations on any single expert’s ability to imagine a risk scenario;

- use of structured probability distribution methods, such as providing Sherman Kent scales, the Stanford probability wheel, use of odd ratios to describe particularly small probability events, employment of interval elicitation, and other methods to help ensure risk concepts, values are clearly and consistently displayed;
- employment of logic and probabilistic reasoning aids such as analytical hierarchy process reasoning methods [20];
- frequent reminders during the administration of the elicitation instrument of definitions, what terms mean, and how risks are described;
- ensuring that the elicitation instrument meets the time constraints for the elicitation, and that risk experts will not be overly tired, bored, or annoyed by the end of the session.

System Process

Elicitation of expert opinion for risks generated by system processes requires a set of approaches somewhat different from those employed for basic event risks. Risks emerging from single events are assumed to be *independent*, allowing the subrisks contributing to an overall frequency estimate to be elicited separately and later combined together as part of a fault tree or similar quantitative structure such as a Monte Carlo simulation.

Risks modeled as being part of system processes require that the elicitation focus on the problem of gathering estimates for joint, *correlated*, risk distributions. These risks are generated by the complex interactions between individual risk generating processes—the term *system process* is used because underlying the creation of an overall risk is an underlying dependency structure that while perhaps only partially understood, is composed of correlated subprocesses.

Consider the problem of estimating the risk that a city levee system will fail over the next period of time. Suppose little or no data exists from which to estimate this risk as a single distribution obtained from observation (since we can’t flood a city on a whim), or from the experience of analog city levee system failures. Instead, such data as does exist in the form of storm prevalence, the relationship between

maintenance practices, levee upkeep and patterns of maintenance disruption by bad weather, and on a host of other subsystem activities (perhaps the relationship between storm patterns and political allocation of levee maintenance resources) required to keep the levee system functional against a subpopulation of storms that might be encountered. To estimate the risk of the levee system failing during a storm, these disparate data series can be stitched together to create an overall risk from the interwoven *system* of correlated processes contributing to overall failure frequency [21].

From an elicitation standpoint, two interesting difficulties exist. The first is that, as in the case of many engineered systems, no single expert will have the span of necessary contributory expertise to characterize the risk system as a whole. This implies that much of the challenge of the expert elicitation is to develop a common representation for the risk dependency structure. The second challenge of eliciting system process risks is that ultimately there may never be sufficient system level data to validate the overall risk model. Development of a reasonable elicitation derived estimate for the risk depends on reconciling different, possibly conflicting expert opinions.

Elicitation of a common representation for the risk process depends upon understanding what is known about the individual processes that relate to risk generation. For example, in the case of the levee example above, there is some relationship between maintenance on the levee and the likelihood that it would fail during a particular storm. If the set of, and relationships between, contributing risk processes are fairly simple, it may be possible to sketch and annotate a representation of the important dependencies between risk processes on a piece of paper over the course of a couple of expert elicitation sessions.

If the interactions between risk processes are more complex, elicitation sessions may be required with many different experts, each of whom may only be familiar with a very small set of the overall risk processes contributing to the overall risk estimate. In these cases, a more formalized process may be required to help ensure that the elicitation instruments, and results, are administered evenly across the spectrum of experts. More formalized risk modeling structures, such as Bayesian networks, may be constructed, during the elicitation

exercise to help negotiate a common risk descriptive representation [22].

The elicitation of system process risks is much less well studied than that of basic event models for risk [23]. Far less is known about the kinds of bias that occur when experts are asked to develop a system structure model. The use of a common *ontology* of concepts involved in the system process risk is a useful way to prevent confusion over what risk processes exist, and how they relate to each other. The use of software supporting representation of how these risk processes interrelate *via* the use of network graphs can also be valuable. Expert group development of a simulation model for the risk process is another way that has been explored to capture complex system interrelationships more accurately.

Elicitation of system process risks is a relatively new approach to risk estimation. It has been made useful largely by the spread of Markov chain Monte Carlo (*see Enterprise Risk Management (ERM); Global Warming; Reliability Demonstration; Bayesian Statistics in Quantitative Risk Assessment; Geographic Disease Risk*) approaches to joint distribution estimation [24]. Unlike the elicitation of basic event modeled risks, system process risk elicitation does not have a robust body of quantitative, behavioral science inspired experiments from which to characterize cognitive biases and inform “best practice” elicitation methods [25–27]. Many of the elicitation techniques applicable to basic event risk elicitation are applicable and can be used. Ensuring that as many different relevant sources of contributory expertise are included in the process to develop, review, and revise the risk process representation is as good advice as can be offered, given the state of research into this type of expert elicitation.

Competitive Game

Competitive game risks (*see Managing Infrastructure Reliability, Safety, and Security; Risk Measures and Economic Capital for (Re)insurers; Mathematics of Risk and Reliability: A Select History; Game Theoretic Methods*) arise from the interaction between two nonprobabilistic actors, acting in opposition to one another. Determining the risk of a terrorist or criminal action of a particular kind is a prevalent example of a competitive game

risk elicitation problem. Eliciting risks resulting from competitive game processes as a basic event risk is an analytic mistake. It is, unfortunately, from a policy-making standpoint, also a prevalent one. A basic understanding of *game theory* is desirable to develop an effective elicitation instrument for these types of risks.

Risk from competitive game processes do not arise from natural processes, but from hazards being *actively generated by another actor* with competing objectives. For example, in the case of screening for firearms at the airport, the distribution of attempts to slip a gun through a detector are not created by a “state-of-nature” black-box process. Rather, this type of criminal action is the result of a set of interactions between the strength of detection defenses and the cost–benefit calculation of those who want to hijack an aircraft. From the standpoint of frequency estimation, the distribution of this risk is due to uncertainties over the costs, benefits, motivations, and abilities of each side in this repeated period strategy transaction.

Expert elicitation of risks from competitive game processes should *not be performed as though these risks stem from basic event processes* [28]. In general, it can be expected that assessment of a particular risk will greatly depend on the following:

- how well experts understand the strategy, information available, and cost–benefit calculations of each side;
- the nature of the game between both sides – is it a single transaction or a multiperiod game, for example;
- the ability to elicit information from the different parties to the competitive game process.

The biggest single difficulty for this type of expert risk elicitation is finding sources of contributory expertise for both sides of the competition. Typically, criminal and terrorist networks do not respond to elicitation requests. To get this risk information, a common strategy is to establish red teams, or employ role-playing exercises that (as much as possible) force experts to immerse themselves in the decision-making context of the other side [29]. Such elicitation exercises can greatly vary in formality from counterthinking scenarios, to table-top exercises, to the formation of “red teams”, or even computer assisted war-gaming type exercises.

Whichever type of elicitation strategy is employed, typically the result is a ranked list of preferred strategies for each side to the competition, given the strategy chosen by the other side. One strategy to convert these ranked preferences into a scenario response distribution is to normalize across ranking or utility scores, to produce a rough distribution describing scenario responses.

Social Negotiation

Social negotiation situations are the most subjective risk process for expert elicitation. Social negotiation risk elicitation exercises differ from elicitation of other risks in that the actual risk quantification is not the purpose of the exercise. Instead, through the language of risk, elicitation is used as a means for negotiation and compromise over disagreements involving social justice, allocation of resources, and distribution of social costs and benefits.

Social negotiation models for risk are commonly encountered in the context of facility site determination, not in my backyard (NIMBY), and similar social-technical disputes. Very often all parties to the elicitation will enter the exercise with their own risk estimates. Whether these are based on “the best available science” or not, these estimates are an amalgamation of communal needs and fears [30]. The purpose of the elicitation is not inherently to convince one side or the other their risk beliefs are wrong, that, for example, the risk of living next to a nuclear power plant is equivalent to “eating 40 tablespoons of peanut butter” [31]. Rather, regardless of what the “actual” risk level is, the purpose of the elicitation is to see whether a resolution can be reached to a political dispute *via* the language of risk estimation [32].

As an inherently political process, elicitation of expert opinion in this context implies a broad definition of contributory expertise to include almost anyone identifiable as a stakeholder in the hazard as “an expert”. Development of an elicitation instrument may also rely more on communal processes to develop fair and open procedures. Meetings, town-hall sessions, and “open microphones” are more likely to be used as the elicitation mechanism rather than structured interviews or survey instruments.

The main success metric for elicitation of risk stemming from social negotiations is the creation of a fair process for information gathering. As formal

quantification of expert opinion is unlikely to be needed beyond development of a clear consensus for the odds of the hazard, the issue of bias is inherently secondary to whether both sides can live with the compromise. As an exercise in political decision making, the design of the elicitation instrument will be highly situational depending on customs and the nature of the risk dispute.

Uses of Expert Elicitation of Risks

Basic event, system process, and competitive game risk elicitation exercises are generally conducted so that qualitative information can be converted into quantitative results. Up through the 1980s, results gathered from expert elicitation were primarily used to populate the nodes on fault or event trees. Expert elicitation served as a supplementary technique to fill areas on a fault tree that could not be estimated using observed data. The focus of elicitation centered on central tendency or odds estimation, especially for large technical systems.

With improvements in microcomputers, gathering the shape of basic event risk distributions became more prevalent. Software tools like Analytica, Crystal-Ball, and other Monte Carlo sampling tools allowed users to quickly and cheaply analyze both the central tendency and spread of a risk distribution resulting from combinations of basic events. To populate these risk models, however, it was necessary to gather, at a minimum, the basic family of distributions described by the expert's risk understanding. Such risk elicitation methods provided data that became incorporated into a wide variety of complex risk simulations.

Bayesian methods (*see Reliability Demonstration; Bayesian Statistics in Quantitative Risk Assessment; Clinical Dose-Response Assessment*) for risk estimation have always provided a natural justification for use of expert elicitation. The use of expert opinion to provide an informative prior is at the heart of Bayesian reasoning. Mathematically, however, the joint integration required to effectively combine the expert derived prior with the data derived posterior, was often intractable [33]. With the development of practical tools implementing the Metropolis-Hasting algorithm and other Bayesian estimation methods, the use of expert risk elicitation has become more important than a technique

to supplement a lack of directly observed data [34]. As expert elicitation of risk is increasingly used by Bayesian statistical practitioners, a better understanding of how techniques to gather system process risks (in the form of Markov chain Monte Carlo) are likely to emerge over the next decade.

References

- [1] Chiles, J.R. (2002). *Inviting Disaster: Lessons from the Edge of Technology*, HarperCollins, New York.
- [2] Morgan, G. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, Cambridge.
- [3] O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E. & Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*, John Wiley & Sons, Chichester and Hoboken.
- [4] Jasanoff, S. (1993). Bridging the two cultures of risk analysis, *Risk Analysis* **13**(2), 123–129.
- [5] Ayoub, B. (2001). *Elicitation of Expert Opinion: Theory, Application, and Guidance*, CRC Press.
- [6] Cooke, R. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*, Oxford University Press, Oxford.
- [7] Meyer, M. & Booker, J. (2001). *Eliciting and Analyzing Expert Judgment: A Practical Guide*, ASA-SIAM, Philadelphia.
- [8] Vesely, W.E., Goldberg, F.F., Roberts, N.H. & Haasl, D.F. (1981). *US Nuclear Regulatory Commission, Fault Tree Handbook (NUREG-0492)*, USGPO, Washington, DC.
- [9] Santori, G., Valente, R., Cambiaso, F., Ghirelli, R., Castiglione, A. & Valente, U. (2003). Preliminary results of an expert-opinion elicitation process to prioritize an informative system funded by Italian Ministry of Health for Cadaveric Donor Management, Organ Allocation, and Transplantation Activity, *Transplantation Proceedings* **36**(3), 433–434.
- [10] Russell, A., Quigley, J. & van der Meer, R. (2006). Modeling the reliability of search and rescue operations within the UK through Bayesian Belief Networks, *First International Conference on Availability, Reliability and Security, Proceedings*, Vienna, Austria, pp. 810–816.
- [11] Qing, X. (1998). Structured subjective judgment on quantifying uncertainty of the real estate project, *Conference of American Real Estate Society*, The Netherlands, pp. 1–9.
- [12] Morgan, M.G. & Keith, D.W. (1995). Subjective judgments by climate experts, *Environmental Science and Technology* **29**, A468–A476.
- [13] Tversky, A. & Kahnemann, D. (1974). Judgement under uncertainty: heuristics and biases, *Science* **185**, 1124–1131.

- [14] Slovic, P. (2000). *The Perception of Risk*, Earthscan Press Ltd, London.
- [15] Kahneman, D. & Tversky, A. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment, *Psychological Review* **90**(4), 293–315.
- [16] Stapel, D.A. & Reicher, S.D. (1994). Social identity, availability and the perception of risk, *Social Cognition* **12**(1), 1–17.
- [17] Plous, S. (1989). Thinking the unthinkable: the effects of anchoring on likelihood estimates of nuclear war, *Journal of Applied Social Psychology* **19**(1), 67–91.
- [18] Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice, *Science* **211**, 453–458.
- [19] Booker, J. & McNamara, L. (2005). Expert knowledge in reliability characterization: a rigorous approach to eliciting, documenting, and analyzing expert knowledge, in *Engineering Design Reliability Handbook*, E. Nikolaidis, D.M. Ghiocel, S. Singhal & N. Nikolaidis, eds, CRC Press, New York.
- [20] Saaty, T.L. (2000). *Fundamentals of Decision Making and Priority Theory*, 2nd Edition, RWS Publications, Pittsburgh.
- [21] Bennett, T.R., Booker, J.M., Keller-McNulty, S. & Singpurwalla, N.D. (2003). Testing the untestable: reliability in the 21st century, *IEEE Transactions on Reliability* **52**(1), 118–124.
- [22] Fischhoff, B. (1989). Eliciting knowledge for analytical representation, *IEEE Transactions on Systems Man and Cybernetics* **19**(3), 448–461.
- [23] Bedford, T., Quigley, J. & Walls, L. (2006). Expert elicitation for reliable system design, *Statistical Science* **21**(4), 428–450.
- [24] Jensen, F. (2002). *Bayesian Networks and Decision Graphs*, Springer-Verlag, New York.
- [25] Jenkins, G.M. (1969). The systems approach, *Journal of Systems Engineering* **1**, 3–49.
- [26] Fortune, J. & Peters, G. (2001). Turning hindsight into foresight – our search for the ‘Philosopher’s Stone’ of failure, *Systemic Practice and Action Research* **14**, 6.
- [27] Hee, D.D., Pickrell, B.D., Bea, R.G., Roberts, K.H. & Williamson, B. (1999). Safety management assessment system (SMAS): a process for identifying and evaluating human and organizational factors in marine system operations with field test results, *Reliability Engineering and System Safety* **65**(2), 125–140.
- [28] Schelling, T.C. (1980). *The Strategy of Conflict*, Harvard University Press, Cambridge.
- [29] Kunreuther, H., Michel-Kerjan, E. & Porter, B. (2003). *Assessing, Managing and Financing Extreme Events: Dealing with Terrorism*, NBER Working Paper 10179, Cambridge, p. 35.
- [30] Freudenburg, W.R. (1996). Risky thinking: irrational fears about risk and society, *The Annals of the American Academy of Political and Social Science* **545**, 545–553.
- [31] Cohen, B.L. & Lee, I.S. (1991). Catalog of risks extended and updates, *Health Physics* **61**, 317–335.
- [32] Kunreuther, H. & Slovic, P. (1996). Science, values, and risk, *The Annals of the American Academy of Political and Social Science* **545**, 116–125.
- [33] Gelman, A., Carlin, J., Stern, H.S. & Rubin, D. (2003). *Bayesian Data Analysis*, 2nd Edition, CRC Press.
- [34] Anderson-Cook, C., Graves, T., Hengartner, N., Klamann, R., Koehler, A., Wilson, A.G., Anderson, G. & Lopez, G.S. (2008). *Reliability Modeling Using Both System Test and Quality Assurance Data*, Forthcoming in the *Journal of the Military Operations Research Society*.

ANDREW C.K. WIEDLEA

Expert Judgment

Judgment involves the weighing of available evidence and reaching a balanced conclusion from that evidence. We bring in experts to provide these judgments because they have developed the mental tools needed to make sound evaluations. These mental tools include knowledge of what evidence can be brought to bear on the question, the abilities to weigh the validity of various pieces of evidence and to interpret the relative importance of various facts or assertions, and to craft a view from an ensemble of information that may be inherently limited or self-conflicted. In risk analysis these judgments nearly always entail uncertainty so that the judgments are not definitive but reflect what we know and what we know we do not know.

The natural language of uncertainty is probability. It provides a precise way to encode knowledge that is inherently imprecise. It puts judgments in a form where they can be manipulated mathematically, and thereby be integrated with other pieces of information and used in models to assess risks. These judgments are essential in many analyses of risk. For example, see [1–5].

When we set about acquiring expert judgments for an analysis, there are a number of decisions that must be made about how to proceed. These include the following:

- selecting the issues to be addressed by the experts
- selecting the experts
- organizing the effort
- choosing a method for combining multiple judgments, if needed.

Posing Questions to the Experts

The first stage of developing an expert judgment process is to determine the objectives and desired products. Judgments can be made about a number of different things. Some judgments are about facts, while others are about values. Roughly speaking, a fact is something that can be verified unambiguously, while a value is a measure of the desirability of something.

For example, you might believe that it is better to expend tax dollars to fight HIV abroad than it is to improve the educational level of impoverished

students at home. This is a value judgment. You might even be able to express in quantitative terms your affinity for one use of tax dollars *versus* another. But another person may hold very different views and neither of you can be said to unequivocally correct. It is a matter of preference, not of fact. In contrast, you and another person may hold different views about the rate of HIV infection. In principle, it is possible to determine this rate (through medical testing and statistical analysis) and thus the judgments are about a fact. This article is limited to judgments about facts, while the discussions in the articles on Utility Functions and Preference Functions focus on values.

Judgments about facts can be further classified. Focusing on judgments often made in risk analysis, we find that judgments can be made about

- the occurrence of future events
- the values of parameters
- the appropriateness of competing models in their ability to reflect reality.

Not all questions, however, are of equal consequence in quantifying risk. There will normally be a few major questions that drive the uncertainty about the risk. These questions are candidates for a more structured expert judgment activity. Other issues – those that play a minor role – can often be treated less formally or through sensitivity analysis, saving the resources for the more important issues. A sensitivity analysis using initial estimates of probabilities and probability distributions is often performed after an initial risk model has been structured. The sensitivity analysis identifies those questions deserving of a more penetrating study.

However, not all issues lend themselves to quantification through expert judgment. In addition to being important contributors to uncertainty and risk, an issue that is a candidate for expert judgment analysis should satisfy the following conditions:

- It should be resolvable in that given sufficient time and/or resources, one could conceivably learn whether the event has occurred or learn the value of the quantity in question. Hence, the issue concerns a fact or set of facts.
- It should have a basis upon which judgments can be made and can be justified.

The requirement of resolvability means that the event or quantity is knowable and physically

measurable. We consider a counterexample. In a study of risk from a radioactive plume following a power plant failure, a simple Gaussian dispersion model of the form $y = ax^b$ was employed [1]. In this model, a and b are simply parameters that give a good fit to the relation between x , downwind distance, and y , the horizontal width of the plume. But not all experts subscribe to this model. More complex alternatives have been proposed with different types of parameters. Asking an expert to provide judgments about a and b violates the first principle above. One cannot verify if the judgments are correct, experts may disagree on the definition of a and b , and experts who do not embrace the simple model will find the parameters not meaningful. It is very difficult to provide a value for something you do not believe exists.

The second requirement is that there is some knowledge that can be brought to bear on the event or quantity. For many issues, there is no directly applicable data so that data from analogs, models using social, medical or physical principles, etc., may form the basis for the judgments. If the basis for judgments is incomplete or sketchy, the experts should reflect this by expressing greater uncertainty in their judgments.

Once issues have been identified, it is necessary to develop a statement that presents the issue to the experts in a manner that will not color the experts' responses. This is called *framing* the issue. Part of framing is creating an unbiased presentation that is free of preconceived notions, political overtones, and discussions of consequences that might affect the response. Framing also provides a background for the question. Sometimes there are choices about whether certain conditions should be included or withheld from the analysis and whether the experts are to integrate the uncertainty about the conditions into their responses. For example, in a study of dry deposition of radioactivity, the experts were told that the deposition surface was northern European grassland, but they were not told the length of the grass which is thought to be an important determinant of the rate of deposition [1]. Instead, the experts were asked to treat the length of grass as an unknown and to incorporate any uncertainty that they might have into their responses. The experts should be informed about those factors that are considered to be known, those that are constrained in value, those that are uncertain, and, perhaps, those that should be excluded from their analyses.

Finally, once an issue has been framed and put in the form of statement to be submitted to the experts, it should be tested. The best way to do this testing is through a dry-run, with stand-in experts who have not been participants in the framing process. Although this seems like a lot of extra work, experience has shown that getting the issue right is both critical and difficult [2]. All too often, the expert's understanding of the question differs from what was intended by the analyst who drafted the question. It is also possible that the question being asked appears to be resolvable to the person who framed the question, but not to the expert who must respond.

Selecting the Experts

The identification of experts requires that one develop some criteria by which expertise can be measured. Generally, an expert is one who "has or is alleged to have superior knowledge about data, models and rules in a specific area or field" [3]. But measuring against this definition requires one to look at indicators of knowledge rather than knowledge *per se*. The following list contains such indicators:

- research in the area as identified by publications and grants
- citations of work
- degrees, awards, or other types of recognition
- recommendations and nominations from respected bodies and persons
- positions held
- membership or appointment to review boards, commissions, etc.

In addition to the above indicators, experts may need to meet some additional requirements. The expert should be free from motivational biases caused by economic, political, or other interest in the decision. The choice of whether to use internal or external experts often hinges on the appearance of motivational biases. Potential experts who are already on a project team may be much easier to engage in an expert judgment process, but questions about the independence of their judgments from project goals may be raised. Experts should be willing to participate and they should be accountable for their judgments [6]. This means that they should be willing to have their names associated with their specific responses. At times, physical proximity or availability will be an important consideration.

How the experts are to be organized also impacts the selection. Often, when more than one expert is used, the experts will be redundant of one another, meaning that they will perform the same tasks. In such a case, one should attempt to select experts with differing backgrounds, responsibilities, fields of study, etc., so as to gain a better appreciation of the differences among beliefs. In other instances, the experts will be complementary, each bringing unique expertise to the question. Here, they act more like a team and should be selected to cover the disciplines needed.

Some analyses undergo extreme scrutiny because of the public risks involved. This is certainly the case with radioactive waste disposal or purity of the blood supply. In such instances, the process for selecting (and excluding) experts should be transparent and well documented. In addition to written criteria, it may be necessary to isolate the project staff from the selection process. This can be accomplished by appointing an independent selection committee to seek nominations and make recommendations to the staff [7].

How many experts should be selected? Experience has shown that the differences among experts can be very important in determining the total uncertainty expressed about a question. Clemen and Winkler [8] examine the impact of dependence among experts using a normal model and conclude that three to five experts are adequate. Hora [9] created synthetic groups from the responses of real experts, and found that three to six or seven experts are sufficient, with little benefit from additional experts beyond that point. When experts are organized in groups, and each group provides a single response, this advice would apply to the number of groups. The optimal number of experts within a group has not been investigated and is likely to be dependent on the complexity of issues being answered.

The Quality of Judgments

Because subjective probabilities are personal and vary from individual to individual and from time to time, there is no “true” probability that one might use as a measure of the accuracy of a single elicited probability. For example, consider the question “what is the probability the next elected president of the United States is a woman?” Individuals may hold different probabilities or degrees of belief about this

event occurring. There is, however, no physical, verifiable probability that could be known but remains uncertain. The event will resolve as occurring or not but, will not resolve to a frequency or probability.

It is possible to address the goodness of probabilities, however. There are two properties that are desirable to have in probabilities:

- probabilities should be informative
- probabilities should authentically represent uncertainty.

The first property, being informative, means that probabilities closer to 0.0 or 1.0 should be preferred to those closer to 0.5 as the more extreme probabilities provide greater certainty about the outcome of an event. In a like manner, continuous probability distributions that are narrower or tighter convey more information than those that are diffuse. The second property, the appropriate representation of uncertainty, requires consideration of a set of assessed probabilities. For those events that are given an assessed probability of p , the relative frequency of occurrence of those events should approach p .

To illustrate this idea, consider two weather forecasters who have provided precipitation forecasts as probabilities. The forecasts are given to a precision of one digit. Thus a forecast of 0.2 is taken to mean that there is a 20% chance of precipitation. Forecasts from two such forecasters are shown in Figure 1.

Ideally, each graph would have a 45° line indicating that the assessed probabilities are faithful in that they correctly represent the uncertainty about reality. Weather Forecaster B’s graph shows a nearly perfect relation while the graph for the Forecaster A shows poorer correspondence between the assessed probabilities and relative frequencies with the actual frequency of rain exceeding the forecast probability. The graph is not even monotonic at the upper end.

Graphs showing the relation between assessed probabilities and relative frequencies are called *calibration graphs* and the quality of the relationship is loosely called *calibration* which can be good or poor [10]. Calibration graphs can also be constructed for continuous assessed distributions. Following [9], let $F_i(x)$ be a set of assessed continuous probability distribution functions and let x_i be the corresponding actual values of the variables. If an expert is perfectly calibrated, the cumulative probabilities of the actual values measured on each corresponding distributions

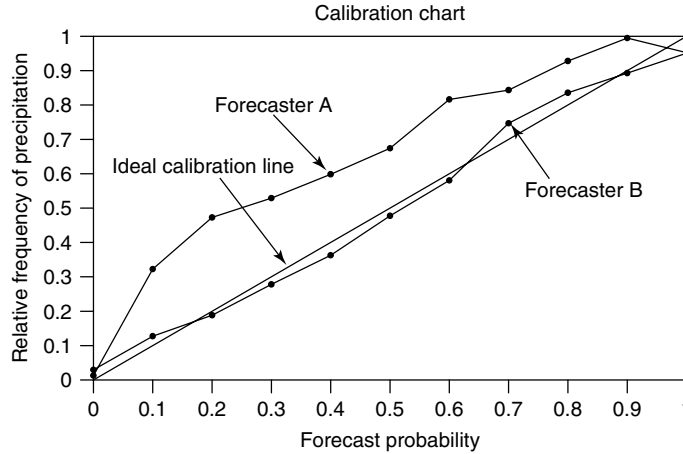


Figure 1 Calibration graph for two forecasters

function, $p_i = F_i(x)$, will be uniformly distributed on the interval $[0, 1]$. We can use the area between the 45° line of perfect calibration and the observed calibration curve as a measure of miscalibration for continuous probability assessments.

Although calibration is an important property for a set of probabilities or probability distributions to possess, it is not sufficient as the probabilities or probability distributions may not be informative. For example, in an area where it rains on 25% of the days, a forecaster who always predicts a 25% chance of rain will be perfectly calibrated but provide no information from day to day about the relative likelihood of rain. But information and calibration are somewhat at odds. Increasing the information by making probabilities closer to zero or one or by making distributions tighter may reduce the level of calibration.

One approach to measuring the goodness of probabilities is through scoring rules. Scoring rules are functions of the assessed probabilities and the true outcome of the event or value of the variable that measure the goodness of the assessed distribution and incorporate both calibration and information into the score. The term *strictly proper scoring function* refers to the property that the expected value of the function is maximized when the probabilities or probability functions to which the function is applied are identical to the probabilities or probability functions that are used to take the expectation. An example will clarify.

A simple strictly proper scoring rule for the assessed probability of an event is the Brier or quadratic rule [11]:

$$\begin{aligned} S(p) &= -(1-p)^2 \text{ if the event occurs} \\ &= -p^2 \text{ if the complement of the event} \\ &\quad \text{occurs} \end{aligned} \quad (1)$$

where p is the assessed probability. For any probability q , the mathematical expectation $E_q[S(p)] = -q(1-p)^2 - (1-q)p^2$ is maximized with respect to p by setting $p = q$. Thus, if an expert believes the probability is q , the expert will maximize the perceived expectation by responding with q . In contrast, the scoring rule $S(p) = -p$ if the event occurs and $S(p) = -(1-p)$, while intuitively pleasing, does not promote truthfulness. Instead, the expected score is maximized by providing a probability p of either 0.0 or 1.0 depending on whether q is less than or larger than 0.5. Winkler [12] provides a discussion of this Brier rule and other strictly proper scoring rules. See also [6, 10].

The concept of a strictly proper scoring rule can be extended to continuous distributions [13]. For example, the counterpart to the quadratic scoring rule for continuous densities is:

$$S[f(x), w] = 2f(w) - \int_{-\infty}^{\infty} f^2(x) dx \quad (2)$$

Expected scores can sometimes be decomposed into recognizable components. The quadratic rule for continuous densities can be decomposed in the following manner. Suppose that an expert's uncertainty

is correctly expressed through the density $g(x)$, but the expert responds with $f(x)$ either through inadvertence or intention. The expected score can be written as follows:

$$E_g \{S[f(x), w]\} = I(f) - C(f, g)$$

$$\text{where } I(f) = \int_{-\infty}^{\infty} f^2(x) dx \text{ and } C(f, g)$$

$$= 2 \int_{-\infty}^{\infty} f(x)[f(x) - g(x)] dx \quad (3)$$

$I(f)$ is the expected density associated with the assessed distribution and is a measure of information. $C(f, g)$ is a nonnegative function that increases as $g(x)$ diverges from $f(x)$. Thus $C(f, g)$ is a measure of miscalibration. Further discussion of decomposition can be found in [10, 14, 15]. Haim [16] provides a theorem that shows how a strictly proper scoring rule can be generated from a convex function. See also Savage [17].

Combining Expert Judgments

There are two situations that may require probability judgments to be combined. The first is when an expert has provided judgments about the elements of a model that result in a top event probability or distribution of a value. Examples of such decompositions include fault trees, event trees (*see Decision Trees*), and influence diagrams (*see Influence Diagrams*). With an event or probability tree, the recomposition can be accomplished by simple probability manipulations. In more complicated situations, you may need to employ simulation methods to obtain a top event probability or distribution for a quantity. This is a relatively straightforward process.

Judgments may also be combined across multiple experts. While using multiple experts to address a single question allows for a greater diversity of approaches and often provides a better representation of the inherent uncertainty, doing so creates the problem of having multiple answers when it would be convenient to have a single answer. If you decide to evaluate the risk model separately, using the judgments of each individual expert and you have multiple points in the model where different experts have given their judgments, the number of separate evaluations of the model is the product of the number of experts used at each place in the model and can

be very large. Aggregation of the judgments into a single probability or distribution avoids this problem.

There are two classes of aggregation methods, behavioral and mathematical. Behavioral approaches entail negotiation to reach a representative or consensus distribution. Mathematical methods, in contrast, are based on a rule or formula. The approaches are not entirely exclusive, however, as they may be both be used to greater or lesser degree to perform an aggregation.

You may be familiar with the “Delphi” technique, developed at the Rand Corporation in the 1960s by Norman Dalkey [18]. In the Delphi method, the interaction among the experts is tightly controlled. In fact, they do not meet face to face but remain anonymous to one another. This is done to eliminate the influence that one might have because of position or personality. The judgments are exchanged among the experts, along with the reasoning for the judgments. After viewing all the judgments and rationales, the experts are given the opportunity to modify their judgments. The process is repeated—exchanging judgments and revising them—until the judgments become static or have converged to a consensus. Often times, it will be necessary to apply some mathematical rule to complete the aggregation process.

Another behavioral approach is called the *nominal group* technique [19]. This technique, like Delphi, controls the interaction among experts. The experts meet and record their ideas or judgments in written form without discussion. A moderator then asks each expert to provide the idea or judgment and records this on a public media display such as a white board, flip chart, or computer screen. There may be several rounds of ideas/judgments which are then followed by discussion. The judgments/ideas are then ranked individually and anonymously by the experts and the moderator summarizes the rankings. This process can be followed by further discussion and reranking or voting on alternatives.

Kaplan [20] proposes a behavioral method for combining judgments through negotiation with a facilitator. The facilitator and experts meet together to discuss the problem. The facilitator’s role is to bring out information from the experts and interpret a “consensus body of evidence” that represents the aggregated wisdom of the group.

A wide range of mathematical methods for combining probability judgments have been proposed.

Perhaps the simplest and most widely used is a simple average termed the *linear opinion pool* [21]. This technique applies equally well to event probabilities and continuous probability densities or distributions. It is important to note that with continuous distributions, it is the probabilities not the values that are averaged. For example, it is tempting, given several medians, to average the medians, but this is not the approach we are referring to. An alternative to the simple average is to provide differential weights to the various experts, ensuring that the weights are nonnegative and sum to one. The values of the weights may be assigned by the staff performing the aggregation or they may result from some measure of the experts' performance. Cooke [6] suggests that evidence of the quality of probability assessments be obtained using training quizzes with questions from the subject matter area addressed by the expert. The experts are given weights based on the product of the p value for the χ^2 test of calibration and the information, as measured by the entropy in the assessments. A cutoff value is used so that poorly calibrated experts are not included in the combination.

The most elegant approach to combining judgments is provided by Morris [22, 23]. In his approach, a decision maker assigns a joint likelihood function to the various responses that a group of experts might provide and a prior distribution for the quantity or event in question. The likelihood function is conditional on the quantity or event of interest. Also see French [24] for a discussion of the axiomatic approaches to combining experts using a Bayesian approach. The decision maker can then develop the posterior distribution for the uncertain quantity or event using Bayes' theorem (*see Bayes' Theorem and Updating of Belief*).

Various mathematical methods for combining probability judgments have different desirable and undesirable properties. Genest and Zidek [25] describe the following property:

Strong set-wise function property

A rule for combining distributions has this property if the rule is a function only of the assessed probabilities and maps $[0, 1]^n \rightarrow [0, 1]$. In particular, the combination rule is not a function of the event or quantity in question.

This property, in turn, implies the following two properties:

Zero set property

If each assessor, $i = 1, \dots, n$ provides $P_i(A) = 0$, then the combined result, $P_c(A)$, should also concur with $P_c(A) = 0$.

Marginalization property

If a subset of events is considered, the marginal probabilities from the combined distribution will be the same as the combined marginal probabilities.

The strong set property also implies that the combining rule is a linear opinion pool or weighted average of the form

$$P_c(A) = \sum_{i=1}^n \alpha_i P_i(A) \quad (4)$$

where the weights, α_i , are nonnegative and sum to one.

Another property is termed the independence property is defined by:

$$P_c(A \cap B) = P_c(A)P_c(B) \quad (5)$$

whenever all experts assess A and B as independent events.

But the linear opinion pool does not have this property. Moreover, the linear opinion pool given above cannot be applied successfully to both joint probabilities and the component marginal and conditional probabilities. That is,

$$P_c(A|B)P_c(B) \neq \sum \alpha_i P_i(A|B)P_i(B) \quad (6)$$

except when one of the weights is one and all others are zero, so that one expert is a "dictator".

The strong set property was used by Dalkey [26] to provide an impossibility theorem for combining rules. Dalkey's theorem adopts seven assumptions that lead to the conclusion that while conforming to the seven assumptions, "there is no aggregation function for individual probability estimates which itself is a probability function". Boardley and Wolff [27] argue that one of these assumptions, the strong set property, is unreasonable and should not be used as an assumption.

While the linear rule does not conform to the independence property, its cousin, the geometric or logarithmic rule does. This rule is linear in the log probabilities and is given by

$$P_c(A) = k \prod_{i=1}^n P_i(A)^{\alpha_i} \text{ where } \alpha_i > 0, \sum_{i=1}^n \alpha_i = 1 \quad (7)$$

and k is a normalizing constant required because this rule does not have the strong set-wise property. The geometric rule also has the property of being externally Bayesian.

Externally Bayesian

The result of performing Bayes' theorem on individual assessments and then combining the revised probabilities is the same as combining the probabilities and then applying Bayes' theorem.

While the geometric rule is externally Bayesian, it is also dictatorial in the sense that if one expert assigns $P_i(A) = 0$, the combined result is necessarily $P_c(A) = 0$. We note that the linear opinion pool is not externally Bayesian.

It is apparent that all the desirable mathematical properties of combining rules cannot be satisfied by a single rule. The topic of selecting a combining method remains an open topic for investigation.

Expert Judgment Designs

In addition to defining issues and selecting and training the expert(s), there are a number of questions that must be addressed concerning the format for a probability elicitation. These include the following:

- the amount of interaction and exchange of information among experts;
- the type and amount of preliminary information that is to be provided to the experts;
- the time and resources that will be allocated to preparation of responses.
- What is the venue – the expert's place of work, the project's home, or elsewhere?
- Will there be training, what kind, and how will it be accomplished?
- Are the names of the experts to be associated with their judgments, and will individual judgments be preserved and made available?

The choices result in the creation of a design for elicitation that has been termed a *protocol*. Some protocols are discussed in [6, 28–30]. We briefly outline two different protocols that illustrate the range of options that have been employed in expert elicitation studies.

Morgan and Henrion [28] identify the Stanford Research Institute (SRI) assessment protocol as, historically, the most influential in shaping structured

probability elicitation. This protocol is summarized in [31]. It is designed around a single expert (subject) and single analyst engaged in a five-stage process detailed below:

- motivating – rapport with the subject is established and possible motivational biases explored;
- structuring – the structure of the uncertainty is defined
- conditioning – the subject is conditioned to think; fundamentally about his judgment and to avoid cognitive biases
- encoding – this is the actual quantification in probabilistic terms;
- verifying – checking for consistency, the responses obtained in the encoding.

The role of the analyst in the SRI protocol is primarily it to help the expert avoid psychological biases. The encoding of probabilities roughly follows a script. Stael von Holstein and Matheson [4] provide a script showing how an elicitation session might go forward.

The encoding stage for continuous variables is described in some detail in [31]. It begins with assessment of the extreme values of the variable. An interesting sidelight is that after assessing these values, the subject is asked to describe scenarios that might result in values of the variable outside of the interval and to provide a probability of being outside the interval. The process next goes to a set of intermediate values whose cumulative probabilities are assessed with the help of the probability wheel. The probability wheel provides a visual representation of a probability and its complement. Then, an interval technique is used to obtain the median and quartiles. Finally, the judgments are verified by testing for coherence and conformance with the expert's beliefs.

While the SRI protocol was designed for solitary experts, a protocol developed by Sandia Laboratories for the US Nuclear Regulatory Commission [5, 32] was designed to bring multiple experts together. The Sandia protocol consists of two meetings:

First meeting agenda

- presentation of the issues and background materials;
- discussion by the experts of the issues and feedback on the questions;

- a training session, including feedback on judgments.

The first meeting is followed by a period of individual study of approximately one month.

Second meeting agenda

- discussion by the experts of the methods, models, and data sources used;
- individual elicitation of the experts.

The second meeting is followed by documentation of rationales and opportunity for feedback from the experts. The final individual judgments are then combined using simple averaging to the final probabilities or distribution functions.

There are a number of significant differences between the SRI and Sandia protocols. First, the SRI protocol is designed for isolated experts while the Sandia protocol brings multiple experts together and allows them to exchange information and viewpoints. They are not allowed, however, to view or participate in the individual encoding sessions or comment on one another's judgments. Second, in the SRI protocol, it is assumed that the expert is fully prepared in that no additional study, data acquisition, or investigation is needed. Moreover, the SRI protocol places the analyst in the role of identifying biases and assisting the expert in counteracting these biases, while the Sandia protocol employs a structured training session to help deal with these issues. In both protocols, the encoding is essentially the same, although the probability wheel is today seldom employed by analysts. Third, the Sandia protocol places emphasis on obtaining and documenting multiple viewpoints which is consistent with the public policy issues addressed in those studies to which it had been applied.

References

- [1] Harper, F.T., Hora, S.C., Young, M.L., Miller, L.A., Lui, C.H., McKay, M.D., Helton, J.C., Goossens, L.H.J., Cooke, R.M., Pasler-Sauer, J., Kraan, B. & Jones, J.A. (1994). *Probability Accident Consequence Uncertainty Analysis*, (NUREG/CR-6244, EUR 15855 EN), USNRC and CEC DG XII, Brussels, Vols 1–3.
- [2] Hora, S.C. & Jensen, M. (2002). *Expert Judgement Elicitation*, Swedish Radiation Protection Authority, Stockholm.
- [3] Bonano, E.J., Hora, S.C., Keeney, R.L. & von Winterfeldt, D. (1989). *Elicitation and Use of Expert Judgment in Performance Assessment for High-Level Radioactive Waste Repositories*, NUREG/CR-5411, U. S. Nuclear Regulatory Commission, Washington, DC.
- [4] Stael von Holstein, C.A.S. & Matheson, J.E. (1979). *A Manual for Encoding Probability Distributions*, SRI International, Menlo Park.
- [5] Hora, S.C. & Iman, R.L. (1989). Expert opinion in risk analysis: the NUREG-1150 experience, *Nuclear Science and Engineering* **102**, 323–331.
- [6] Cooke, R.M. (1991). *Experts in Uncertainty*, Oxford University Press, Oxford.
- [7] Trauth, K.M., Hora, S.C. & Guzowski, R.V. (1994). *A Formal Expert Judgment Procedure for Performance Assessments of the Waste Isolation Pilot Plant*, SAND93-2450, Sandia National Laboratories, Albuquerque.
- [8] Clemen, R.T. & Winkler, R.L. (1985). Limits for the precision and value of information from dependent sources, *Operations Research* **33**, 427–442.
- [9] Hora, S.C. (2004). Probability judgments for continuous quantities: linear combinations and calibration, *Management Science* **50**, 597–604.
- [10] Lichtenstein, S., Fischhoff, B. & Phillips, L.D. (1982). Calibration of probabilities: the state of the art to 1980, in *Judgment Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic & A. Tversky, eds, Cambridge University Press, Cambridge.
- [11] Brier, G. (1950). Verification of weather forecasts expressed in terms of probabilities, *Monthly Weather Review* **76**, 1–3.
- [12] Winkler, R.L. (1996). Scoring rules and the evaluation of probabilities (with discussion and reply), *Test* **5**, 1–60.
- [13] Matheson, J.E. & Winkler, R.L. (1976). Scoring rules for continuous probability distributions, *Management Science* **22**, 1087–1096.
- [14] Murphy, A.H. (1972). Scalar and vector partitions of the probability score part I: two-state situation, *Journal of Applied Meteorology* **11**, 273–282.
- [15] Murphy, A.H. (1973). A new vector partition of the probability score, *Journal of Applied Meteorology* **12**, 595–596.
- [16] Haim, E. (1982). *Characterization and Construction of Proper Scoring Rules*, Doctoral dissertation, University of California, Berkeley.
- [17] Savage, L.J. (1971). The elicitation of personal probabilities and expectations, *Journal of the American Statistical Association* **66**, 783–801.
- [18] Dalkey, N.C. (1967). *Delphi*, Rand Corporation Report, Santa Monica.
- [19] Delbecq, A.L., Van de Ven, A.H. & Gustafson, D.H. (1986). *Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes*, Green Briar Press, Middleton.
- [20] Kaplan, S. (1990). 'Expert information' vs 'expert opinions': another approach to the problem of eliciting/combining/using expert knowledge in PRA, *Journal of Reliability Engineering and System Safety* **39**, 61–72.
- [21] Stone, M. (1961). The linear opinion pool, *Annals of Mathematical Statistics* **32**, 1339–1342.

- [22] Morris, P.A. (1974). Decision analysis expert use, *Management Science* **20**, 1233–1241.
- [23] Morris, P.A. (1977). Combining expert judgments: a Bayesian approach, *Management Science* **23**, 679–693.
- [24] French, S. (1985). Group consensus probability distributions: a critical survey, in *Bayesian Statistics 2*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith, eds, North-Holland, pp. 183–201.
- [25] Genest, C. & Zidek, J.V. (1986). Combining probability distributions: a critique and annotated bibliography, *Statistical Science* **1**, 114–148.
- [26] Dalkey, N. (1972). *An Impossibility Theorem for Group Probability Functions*, The Rand Corporation, Santa Monica.
- [27] Boardley, R.F. & Wolff, R.W. (1981). On the aggregation of individual probability estimates, *Management Science* **27**, 959–964.
- [28] Morgan, M.G. & Henrion, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, New York.
- [29] Merkhofer, M.W. (1987). Quantifying judgmental uncertainty: Methodology, experiences, and insights, *IEEE Transactions on Systems, Man, and Cybernetics* **17**, 741–752.
- [30] Keeney, R. & von Winterfeldt, D. (1991). Eliciting probabilities from experts in complex technical problems, *IEEE Transactions on Engineering Management* **38**, 191–201.
- [31] Spetzler, C.S. & Stael von Holstein, C.-A.S. (1975). Probability encoding in decision analysis, *Management Science* **22**, 340–358.
- [32] Ortiz, N.R., Wheeler, T.A., Breeding, R.J., Hora, S., Meyer, M.A. & Keeney, R.L. (1991). The use of expert judgment in the NUREG-1150, *Nuclear Engineering and Design* **126**, 313–331.

Related Articles

Scientific Uncertainty in Social Debates Around Risk

Subjective Probability

Uncertainty Analysis and Dependence Modeling

STEPHEN C. HORA

Extreme Event Risk

“Extreme event” is a vague concept. It has a variety of formal definitions based on specific cases and contexts. Particular extreme events include, great earthquakes, floods, droughts, nuclear meltdowns, space shuttle losses, deaths, giant sea waves, stock crashes, landslides, storms, hurricanes, tornadoes, wildfires, and consequences of any of the preceding. The event of interest might be caused by man or by nature. The events occur in time or space. They may be purely random or somehow correlated. Their size might be measured in deaths or costs. The formal study of extreme event situations has led to important research results and management methods.

The concern of this encyclopedia is quantitative risk assessment and one specific definition of extreme event risk is,

Prob{certain performance variates exceed relevant critical values as a function of explanatories}

The performance variate might be a binary variable corresponding to whether or not the event occurs, or it might be the loss and damage associated with the event. In seismic engineering a concern is the damage resulting from forces generated by an earthquake in the physical component of a structure, say a nuclear reactor. In this case the models have become quite subtle. Three dimensional motion measured in historical earthquakes is employed in simulations. The performance variate might be the force being applied to a piece of piping. An explanatory might be the duration of the shaking of the earthquake. The duration of the applied forces is crucial. A very brief application may cause no damage, but a lengthy one may cause a catastrophe.

Phrased as above one has a statistics problem whose solutions may be based on stochastic modeling, data collection, and data analysis. Particular statistical methods prove helpful. These include: Bayes rule, influence diagrams, state space formulations, threshold models, aggregation, stratification, marked point process technology, trend models, cluster models, model validation, limit theorems, asymptotic approximations, and smoothing. The Poisson process plays an important role, but so too do renewal processes and marked point processes. These last have the form $\{t_j, M_j\}$ where t_j refers to the time of the

event and the mark M_j is associated information that might perhaps involve the past. The index of the *Encyclopedia of Environmetrics* [1], references these various concepts.

The subject matter of physics, chemistry, and engineering contributes to evaluations of risk as do the ideas of systems analysis. The latter includes box and arrow diagrams, simulation, decision tools, geographic information systems, and database management tools.

Typically, one seeks risk estimates for some future time interval. In seeking solutions to such a forecasting problem one might employ leading indicators. For example, in their concern with flooding in the Amazon city of Manaus, the authorities estimate the probability of eventual flooding during the year by the height of the river at the end of March and when time has passed the height at the end of April [2].

Desired products of an extreme event risk analysis include: hazard/risk maps, formulas, graphics, time plots, and forecasts. Risk probabilities, such as the one above, have been built into government regulations by NASA and the Nuclear Regulation Commission, for example.

Difficulties of the work include small data sets, time change and other inhomogeneities, poorly defined concepts, nonstandard situations, missing data, outliers, and measurement bias.

The demand for risk analyses is growing, in part because the costs of replacing destroyed structures are growing and in part because of the steady increase in the number of people living in hazardous areas.

Some details concerning the particular cases of seismic risk assessment, wildfire occurrence and flooding of the Amazon River are provided in [2]. Reference [3] has a chapter concerning seismic risk analysis. Risk assessment for space missions and hardware are discussed in [4]. References [5] and [6] present relevant statistical models and methods with the emphasis in [6] being on the insurance case. Catastrophe models are discussed in [7]. Reference [8] provides substantive and statistical material concerning climate research. Scholarly journals containing papers devoted to risk assessment include *Extremes*, *Geneva Papers on Risk and Insurance*, *Human and Environmental Risk Assessment*, *Natural Hazards*, *Risk Analysis*, *Risk and Reliability*, *Risk Research*, *Risk and Uncertainty*,

and *Stochastic Environmental Research and Risk Assessment*.

References

- [1] El-Shaarawi, A.H. & Piegorsch, W.W. (eds) (2002). *Encyclopedia of Environmetrics*, John Wiley & Sons, New York.
- [2] Brillinger, D.R. (2003). Three environmental probabilistic risk problems, *Statistical Science* **18**, 412–421.
- [3] Bullen, K.E. & Bolt, B.A. (1985). *An Introduction to the Theory of Seismology*, 4th Edition, Cambridge University Press, Cambridge.
- [4] Cleghorn, G.E.A. (1995). *Orbital Debris: A Technical Assessment*, National Academy Press, Washington, DC.
- [5] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer, New York.
- [6] Embrechts, P., Kluppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.
- [7] National Academy of Sciences (1998). *Paying the Price*, H. Kunreuther & R.J. Roth Sr, eds, Henry Press, Washington, DC.
- [8] Storch, H. & Zwiers, F. (1999). *Statistical Analysis in Climate Research*, Cambridge University Press, Cambridge.

Related Articles

Axiomatic Measures of Risk and Risk-Value Models

Correlated Risk

.

DAVID R. BRILLINGER

Extreme Value Theory in Finance

During the last decades, the financial market has grown rapidly. As a result, this has led to the demand and, consequently, to the development of risk-management systems. The internationally recognized regulatory concept is called *Basel II accord* (see **Credit Risk Models; Solvency; Credit Scoring via Altman Z-Score; Efficacy**), which suggests international standards for measuring the adequacy of a bank's risk capital. Hence, in any discussion on risk in finance, it is of importance to keep the Basel II accord in mind; see [1] for an interesting discussion on regulation of financial risk. Also, in the Basel II accord, the different aspects of financial risk are defined. We distinguish these as follows:

- *Market risk:* (see **Options and Guarantees in Life Insurance; Reinsurance; Value at Risk (VaR) and Risk Measures**) the risk that the value of an investment will decrease due to movements in market factors.
- *Credit risk:* (see **Informational Value of Corporate Issuer Credit Ratings; Credit Migration Matrices; Mathematical Models of Credit Risk; Default Risk; Credit Value at Risk**) the risk of loss due to a debtor's nonpayment of a loan (either the principal or interest (coupon) or both).
- *Operational risk:* (see **Nonlife Insurance Markets; Compliance with Treatment Allocation**) the risk of losses resulting from inadequate or failed internal processes, people and systems, or external events.

In this article, we discuss extreme value theory (EVT) (see **Individual Risk Models; Mathematics of Risk and Reliability: A Select History**) in general, and indicate how to use it to model, measure, and assess financial risk. On balance, EVT is a practical and useful tool for modeling and quantifying risk. However, it should, as with all model-based statistics, be used with care. For interesting reading of EVT and risk management, see [2, 3].

The structure of the article is as follows. In the section titled "Extreme Value Theory", we introduce basic concepts of EVT for independent and identically distributed (i.i.d.) data, independent of financial

applications. In the section titled "Financial Risk Management", we introduce financial market risk and describe how to model and analyze it with tools from EVT. Here, we also present methods for temporal-dependent financial time series and indicate multivariate modeling, both of which are important for risk management, in assessing joint losses at subsequent times and of different instruments in a portfolio. As the role of EVT in credit risk and operational risk has not yet been clarified and is under inspection, we have refrained from a thorough presentation. We would like, however, to refer to the papers [4–7] for an extreme value approach to credit risk, and to [8–11] for interesting developments in operational risk modeling. Further references can be found in these papers.

Extreme Value Theory

EVT is the theory of modeling and measuring events that occur with very small probability. This implies its usefulness in risk modeling, as risky events per definition happen with low probability. Textbooks on EVT include [12–17]. The book [12] treats extreme value statistics from a mathematical statistician's point of view with interesting case studies, and also contains material on multivariate extreme value statistics, whereas [13] lays greater emphasis on applications. The book [14] combines theoretical treatments of maxima and sums with statistical issues, focusing on applications from insurance and finance. In [15], EVT for stationary time series is treated. The monograph [16] is an extreme value statistics book with various interesting case studies. Finally, [17] is a mathematically well-balanced book aiming mainly at multivariate EVT based on multivariate regular variation. We also mention the journal *Extremes* which specializes on all probabilistic and statistical issues in EVT and its applications.

In EVT, there are two main approaches with their own strength and weakness. The first one is based on modeling the maximum of a sample (or a few largest values of a sample, called the upper order statistics) over a time period. In [14, Chapter 3], this approach is rigorously formulated based on the Fisher – Tippet theorem (see **Reliability of Large Systems**) going back to 1928. The second approach is based on modeling excess values of a sample over a threshold, within a time period. This approach is

called *peaks over threshold* or POT method and has been suggested originally by hydrologists.

Both approaches are equivalent by the Pickands – Balkema – de Haan theorem presented in [14, Theorem 3.4.5]. Statistics based on EVT has to use the largest (or smallest) values of a sample. They can be selected in different ways, and we assume at the moment that we have i.i.d. data. The first method is based on a so-called block maxima method, where data are divided into blocks, whose maxima are assumed to follow an extreme value distribution. The second method uses the joint distribution function of upper order statistics, and the third method uses the POT method, i.e., it invokes excesses over a high threshold. In this article, we focus on the POT method for two reasons. Firstly, there is a near consensus that its performance is better than other EVT methods for estimating quantiles; see e.g. [1]. Secondly, it can be extended easily to dependent data. This is also true for the block maxima method; however, the blocks have to be chosen so that the resulting block maxima are independent, so that maximum-likelihood estimation (MLE) can be applied. Further, the block maxima method invokes considerably less data than the POT method, which makes the POT method more efficient.

The existing software for the analysis of extreme events has been reviewed in [18], where a glossary of available packages can be found. We mention EVIS, which is written in SPlus, available at www.ma.h-w.ac.uk/~mcneil with the corresponding package EVIR based on R. Also, we recommend the package EVD, also based on R, which includes

programs for multivariate extreme value analysis at www.maths.lancs.ac.uk/~stephana. Finally, the MATLAB package EVIM is available at www.bilkent.edu.tr/~faruk.

The following introduction of the POT method is based on [19]. The POT method estimates a far out tail or a very high quantile, on the basis of extreme observations of a sample and consists of three parts. Each part is based on a probabilistic principle that is explained in the following paragraphs. Figure 1 serves as an illustration.

1. Point process of exceedances

Given a high threshold u_n , we index each observation of the sample $X_1, \dots, X_n$ exceeding u_n . (In Figure 1 these are observations 2, 3, 5, 6, 10, and 12.) To obtain a limit result, we let the sample size n tend to infinity and, simultaneously, the threshold u_n increase, in the correct proportion.

For i.i.d. data, each data point has the same chance to exceed the threshold u_n , the success probability being simply $P(X_i > u_n)$ for $i = 1, \dots, n$. Hence, the number of observations exceeding this threshold

$$\#\{i : X_i > u_n, i = 1, \dots, n\} = \sum_{i=1}^n I(X_i > u_n) \quad (1)$$

follows a binomial distribution with parameters n and $P(X_i > u_n)$. Here, $I(X_i > u_n) = 1$ or 0 , according as $X_i > u_n$ or $\leq u_n$. If for some $\tau > 0$

$$nP(X_i > u_n) \rightarrow \tau, \quad n \rightarrow \infty \quad (2)$$

then by the classical theorem of Poisson, the distribution of $\#\{i : X_i > u_n, i = 1, \dots, n\}$ converges to

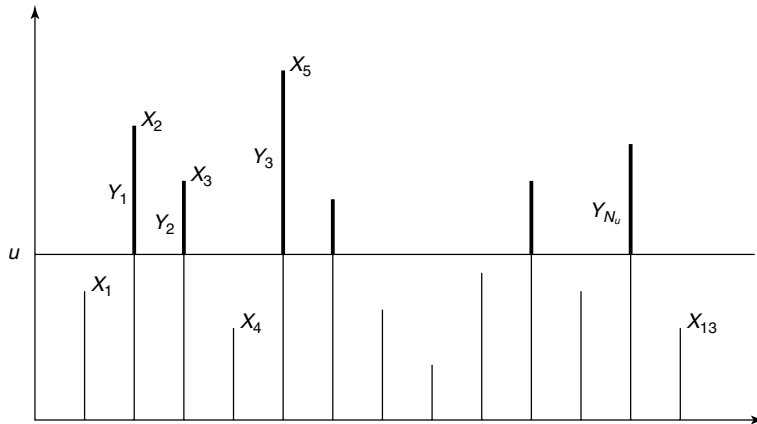


Figure 1 Data $X_1, \dots, X_{13}$ with corresponding excesses $Y_1, \dots, Y_{N_u}$

a Poisson distribution with parameter τ . If $X_i, i = 1, \dots, n$, come from an absolutely continuous distribution, equation (2) is a rather weak condition: for all known absolutely continuous distributions and every $\tau > 0$, a suitable series (u_n) can be found (see e.g. [14, Chapter 3]). Indexing all points $\{i : X_i > u_n, i = 1, \dots, n\}$ in the interval $[0, n]$, the latter becomes larger and larger, whereas the indexed points become sparser and sparser (as the threshold u_n rises with n). A more economic representation is gained by not plotting the points on the interval $[0, n]$, but rather on the interval $[0, 1]$. An observation X_i exceeding u_n is then plotted not at i , but at i/n . If, for $n \in \mathbb{N}$, we define

$$N_n((a, b]) = \#\{i/n \in (a, b] : X_i > u_n, i = 1, \dots, n\} \quad (3)$$

for all intervals $(a, b] \subset [0, 1]$, then N_n defines a point process on the interval $[0, 1]$. This process is called the *time-normalized point process of exceedances*. Choosing u_n such that equation (2) holds, the series N_n of point processes converges (as $n \rightarrow \infty$) in distribution to a Poisson process with parameter τ . For the measure theoretical background on convergence of point processes, see e.g. [14, Chapter 5].

2. The generalized Pareto distribution (see Extreme Values in Reliability)

For the exceedances of a high threshold, we are not only interested in when and how often they occur, but also in how large the excess $X - u | X > u$ is. (In Figure 1 the excesses are labeled $Y_1, \dots, Y_{N_u}$, and the number of exceedances is $N_u = 6$.) Under condition (2) it can be shown that for a measurable positive function a ,

$$\lim_{u \rightarrow \infty} P\left(\frac{X - u}{a(u)} > y \mid X > u\right) = (1 + \xi y)^{-1/\xi} \quad (4)$$

if the left-hand side converges at all. For $\xi = 0$, the right-hand side is interpreted as e^{-y} . For all $\xi \in \mathbb{R}$ the right-hand side is the tail of a distribution function, the so-called *generalized Pareto distribution*. If $\xi \geq 0$ the support of this distribution is $[0, \infty)$, for $\xi < 0$ the support is a compact interval. The case $\xi < 0$ is of no interest for our applications, and therefore, not considered.

3. Independence

Finally, it can be shown that the point process of exceedances and the excesses, that is, the sizes of the exceedances, are in the limit independent.

How can these limit theorems be used to estimate tails (see **Enterprise Risk Management (ERM)**) and quantiles?

The following paragraph illustrates the POT method for a given sample $X_1, \dots, X_n$. For a high threshold u , we define

$$N_u = \#\{i : X_i > u, i = 1, \dots, n\} \quad (5)$$

We refer to the excesses of $X_1, \dots, X_n$ as $Y_1, \dots, Y_{N_u}$, as indicated in Figure 1. The tail of F is denoted by $\bar{F} = 1 - F$. Defining $\bar{F}_u(y) = P(Y_1 > y | X > u)$ yields

$$\bar{F}_u(y) = P(X - u > y | X > u) = \frac{\bar{F}(u + y)}{\bar{F}(u)}, \quad y \geq 0 \quad (6)$$

Consequently, we get

$$\bar{F}(u + y) = \bar{F}(u) \bar{F}_u(y), \quad y \geq 0 \quad (7)$$

An observation larger than $u + y$ is obtained if an observation exceeds u , i.e., an exceedance is required, and if, furthermore, such an observation has an excess over u that is also greater than y . An estimator of the tail (for values greater than u) can be obtained by estimating both tails on the right-hand side of equation (7).

This is done by exploiting (1–4) above.

We estimate $\bar{F}(u)$ by its empirical counterpart

$$\widehat{\bar{F}}(u) = \frac{1}{n} \sum_{i=1}^n I(X_i > u) = \frac{N_u}{n} \quad (8)$$

Then, we approximate $\bar{F}_u(y)$ by the generalized Pareto distribution, where the scale function $a(u)$ has to be taken into account. The latter is integrated into the limit distribution as a parameter. This gives

$$\bar{F}_u(y) \approx \left(1 + \xi \frac{y}{\beta}\right)^{-1/\xi} \quad (9)$$

where ξ and β have to be estimated (by $\hat{\xi}$ and $\hat{\beta}$). (Note that $\beta = \beta(u)$ is a function of u .) This results,

for given u , in the following tail estimator:

$$\bar{F}(u + y) = \frac{N_u}{n} \left(1 + \hat{\xi} \frac{y}{\hat{\beta}} \right)^{-1/\hat{\xi}}, \quad y \geq 0 \quad (10)$$

By inversion, one obtains for a given $\alpha \in (0, 1)$, an estimator of the α -quantile of the form

$$\hat{x}_\alpha = u + \frac{\hat{\beta}}{\hat{\xi}} \left(\left(\frac{n}{N_u} (1 - \alpha) \right)^{-\hat{\xi}} - 1 \right) \quad (11)$$

Finding the threshold u is important, but by no means trivial. It is a classical variance-bias problem, made worse by the small sample of rare events data. The choice of the threshold is usually based on an extended explorative statistical analysis; cf. [14, Section 6.2] or [13, Chapter 4]. Once the threshold is found, parameter estimation can be performed by maximum likelihood. The seminal paper [20] shows that MLEs are asymptotically normal, provided that $\xi > -0.5$, and derives the asymptotic covariance matrix of the estimators; see [14, Section 6.5] or [13, Chapter 4]. This allows also to calculate confidence intervals. Alternatively, they can also be obtained by the profile likelihood method; see [13]. There are algorithms on adaptive threshold selection, some asymptotically optimal, in settings similar to the POT method, see e.g. [12, Section 4.7]. However, to apply such black box methods can be dangerously misleading (see the Hill horror plot in [14, Figure 4.1.13]), and we view such automatic methods as a complementary tool, which can be helpful in finding an appropriate threshold, but which should not be used stand alone.

Financial Risk Management

We recall that market risk is the risk that the value of an investment decreases due to movements in market factors. Examples include equity risk, interest rate risk, currency risk, and commodity risk. The most common risk measure in the finance industry is the *value at risk* (VaR) (see **Risk Measures and Economic Capital for (Re)insurers; Credit Migration Matrices; Value at Risk (VaR) and Risk Measures**), which is also recommended in the Basel II accord. Consider the loss L of a portfolio over a given time period Δ , then VaR is a risk statistic that measures the risk of holding the portfolio for the time

period Δ . Assume that L has distribution function F_L , then we define VaR at level $\alpha \in (0, 1)$ as

$$\begin{aligned} \text{VaR}_\alpha^\Delta(L) &= \inf\{x \in \mathbb{R} : P(L > x) \leq 1 - \alpha\} \\ &= \inf\{x \in \mathbb{R} : F_L(x) \geq \alpha\} \end{aligned} \quad (12)$$

Typical values of α are 0.95 and 0.99, while Δ usually is 1 or 10 days. For an overview on VaR in a more economic setting we refer to [21].

Although intuitively a good concept, VaR has met criticism from academia and practice because of some of its theoretical properties and its nonrobustness in statistical estimation. On the basis of an axiomatic approach, Artzner *et al.* [22] suggest so-called coherent risk measures, which excludes VaR as it is, in general, not subadditive. An alternative set of axioms has been suggested in [23], leading to convex risk measures, motivated by the economic fact that risk may not increase linearly with the size of the portfolios.

A risk measure closely related to VaR, which is coherent, is the expected shortfall (ES), also known as *conditional VaR* (CVaR), which is defined as the expected loss given that we have a loss larger than VaR. Assume that $E[|L|] < \infty$, then ES is defined as

$$ES_\alpha^\Delta(L) = E[L | L > \text{VaR}_\alpha^\Delta(L)] \quad (13)$$

Provided the loss distribution is continuous, then

$$ES_\alpha^\Delta(L) = \frac{1}{1 - \alpha} \int_\alpha^1 \text{VaR}_x^\Delta(L) \, dx \quad (14)$$

Although ES is a more informative risk measure than VaR, it suffers from even higher variability, as we use information very far out in the tail. For a discussion on the problematic description of complex risk with just one single number, we refer to [3].

As VaR and ES are, per definition, based on a high quantile, EVT is a tailor-made tool for estimating both. Instead of considering the whole distribution, estimation is based on the tail of the loss distribution in the context of the acronym *letting the tail speak for itself*. This is the only way of obtaining an estimator based on the relevant data for an extreme event. We concentrate here on the estimation procedure using the POT method, as described in the section titled “Extreme Value Theory”.

Concretely, using the generalized Pareto distribution, $\text{VaR}_\alpha^\Delta(L)$ is estimated as the α -quantile of the log-differences (called logreturns) of the portfolio loss process $(L_t)_{t \in \mathbb{N}_0}$ in a time interval of length Δ .

The logreturns are assumed to constitute a stationary time series, then the data used for estimation are given by

$$X_t^\Delta = \log(L_t) - \log(L_{t-\Delta}), \quad t = \kappa \Delta, \kappa \in \mathbb{N}_0 \quad (15)$$

Now, using equation (11) for a well-chosen threshold $u > 0$, we obtain an estimate for the quantile of the logreturn loss distribution as

$$\widehat{\text{VaR}}_\alpha(X^\Delta) = u + \frac{\hat{\beta}}{\hat{\xi}} \left(\left(\frac{n}{N_u} (1 - \alpha) \right)^{-\hat{\xi}} - 1 \right) \quad (16)$$

Similarly, ES can be estimated, provided that the logreturns have finite expectation ($\xi < 1$), and we obtain

$$\widehat{\text{ES}}_\alpha(X^\Delta) = \frac{\widehat{\text{VaR}}_\alpha(X^\Delta)}{1 - \hat{\xi}} + \frac{\hat{\beta} - \hat{\xi}u}{1 - \hat{\xi}} \quad (17)$$

In a bank, usually, a daily VaR is estimated, i.e., $\Delta = 1$ day based on daily logreturns. However, according to the Basel II accord $\Delta = 10$ days has to be estimated as well. This is statistically problematic as one needs a very long time series to get good estimates, which, on the other hand, puts stationarity to the question. To avoid this problem, and also to avoid multiple estimation, one often uses a scaling rule,

$$\begin{aligned} \widehat{\text{VaR}}_\alpha(X^\Delta) &= \Delta^\kappa \widehat{\text{VaR}}_\alpha(X^1) \quad \text{and} \\ \widehat{\text{ES}}_\alpha(X^\Delta) &= \Delta^\kappa \widehat{\text{ES}}_\alpha(X^1) \end{aligned} \quad (18)$$

Here, typically, $\kappa = 0.5$ is chosen, the so-called square root scaling rule, which is based on the central limit theorem; see [1] for a discussion on different values of κ . Applying such a scaling rule as in equation (18) can, however, be grossly misleading as realistic financial models usually do not allow for such naive scaling; cf. [24] for an example.

Time Series Approach

A stylized fact of financial logreturns is that they are seemingly uncorrelated, but the serial correlation of absolute or squared logreturns is significant. This is because of the empirical fact that a large absolute movement tends to be followed by a large absolute movement. One can view this as a varying temperature of the market introducing dependence into the data. In a naive EVT approach, one ignores this effect and, therefore, produces estimators with nonoptimal performance.

To estimate VaR and/or ES for temporally dependent data two approaches have been suggested in the literature. The first approach is called *unconditional* VaR or ES estimation and is based on a POT method, where possible dependence is taken into account. In such models, provided they are stationary, the POT method presented by (1–4) in the section titled “Extreme Value Theory”, has to be modified to account for possible clustering in the extremes. The limiting point process of (3) can, by dependence, become a marked Poisson process. Exceedances over high thresholds (see **Extreme Values in Reliability**) appear in clusters, so that a single excess as in the Poisson process, turns into a random number of excesses, whose distribution is known for certain models; see [15] or [14, Chapter 8], for the theoretical basis and [25, 26], for examples. As a consequence, the maximum value of a large sample of size n from such a dependent model behaves like the maximum value of a smaller sample size $n\theta$ for some $\theta \in (0, 1)$ of an i.i.d. sample with the same marginal distribution. The parameter θ is called *extremal index* and can, under weak regularity conditions, be interpreted as the inverse of the mean cluster size. This parameter enters into the tail and quantile estimators (equations 10 and 11), which become, for a high threshold $u > 0$,

$$\bar{F}(u + y) = \frac{N_u}{n\theta} \left(1 + \hat{\xi} \frac{y}{\hat{\beta}} \right)^{-1/\hat{\xi}}, \quad y \geq 0 \quad (19)$$

respectively, for $\alpha \in (0, 1)$,

$$\hat{x}_\alpha = u + \frac{\hat{\beta}}{\hat{\xi}} \left(\left(\frac{n\theta}{N_u} (1 - \alpha) \right)^{-\hat{\xi}} - 1 \right) \quad (20)$$

The parameter θ has to be estimated from the data and is again, of course, highly dependent of the threshold u ; see [14, Section 8.1], for details and estimation procedures. New methods have been suggested in [27, 28].

The second approach is called *conditional* or *dynamic* VaR (see **Equity-Linked Life Insurance; From Basel II to Solvency II – Risk Management in the Insurance Sector; Credit Migration Matrices**) or *ES estimation*. The most realistic models introduce stochastic volatility (see **Risk-Neutral Pricing: Importance and Relevance; Numerical Schemes for Stochastic Differential Equation Models; Volatility Smile**) and model a portfolio’s daily logreturns by

$$X_t = \mu_t + \sigma_t Z_t, \quad t = 1, \dots, n \quad (21)$$

where σ_t is the stochastic volatility, μ_t is the expected return, and Z_t is a random variable with mean zero and variance one. It starts by estimating μ_t and σ_t , usually by means of quasi-maximum likelihood (see **Statistics for Environmental Mutagenesis; Repeated Measures Analyses**), and applies the classical POT method from the section titled “Extreme Value Theory” to the residuals $\hat{Z}_t = (X_t - \hat{\mu}_t)/\hat{\sigma}_t$. The residuals $\hat{Z}_t$ are sometimes called devolatilized logreturns. We obtain for a given time period Δ ,

$$\widehat{\text{VaR}}_\alpha(X^\Delta) = \hat{\mu}_{t+\Delta} + \hat{\sigma}_{t+\Delta} \widehat{\text{VaR}}_\alpha(Z^\Delta) \quad (22)$$

and

$$\widehat{\text{ES}}_\alpha(X^\Delta) = \hat{\mu}_{t+\Delta} + \hat{\sigma}_{t+\Delta} \widehat{\text{ES}}_\alpha(Z^\Delta) \quad (23)$$

The most prominent conditional mean and variance model for μ_t and σ_t is the ARMA-(G)ARCH model which considers the mean to be an ARMA process and the volatility to be a (G)ARCH process, see [29]. It is worth mentioning that Robert F. Engle received the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel in 2003 for his development of the (G)ARCH models. For a successful implementation of estimating VaR and ES using the ARMA-(G)ARCH model, we refer to the explanatory paper [30]. It exemplifies that an AR-(G)ARCH model, together with the generalized Pareto distribution for the residuals, produces very accurate VaR and ES estimates.

Financial market data of liquid markets are tick-by-tick data, where each tick corresponds to a transaction. Such data are also termed high-frequency data and have been in the focus of research in recent years. The accessibility of high-frequency data has lead to the development of continuous-time financial models, see e.g. [31, 32] for two different models and [33] for a comparison. VaR estimation based on high-frequency data has been considered in [34], estimating VaR on different time frames using high-frequency data; see also [35] for a discussion on extreme returns of different time frames.

Multivariate Issues

Often, when modeling financial risk, one investigates the development of a portfolio’s logreturns, aggregating different risk factors. This prevents the possibility of tracking joint large losses, which may indeed lead to complete disaster. Consequently, one should model a portfolio as a multivariate random vector. Unfortunately, in financial applications, this leads easily to a

very high-dimensional problem, and modeling of the dependence structure could be very difficult, or even impossible. One way out is to use comparably few selected risk factors, or to group assets in different sectors or geographical regions.

Multivariate EVT (see **Individual Risk Models; Multivariate Reliability Models and Methods; Mathematics of Risk and Reliability: A Select History; Copulas and Other Measures of Dependence**) provides the theoretical background to model and analyze joint extreme events by concentrating on dependence in extreme observations. An indication of extreme dependence in financial data is the well-known fact that market values tend to fall together in a market crash. Multivariate EVT makes a contribution to this problem by offering a tool to model, analyze, and understand the extreme dependence structures; see Figure 2 to indicate that the commonly used normal model does not capture joint extreme events.

In multivariate EVT, there are mainly two approaches. The first one utilizes multivariate regular variation. This approach is based on the fact that marginal models for high risk data are very well modeled by Pareto-like models, whose tails decrease like $P(X > x) = x^{-\alpha} \ell(x)(1 + o(1))$ as $x \rightarrow \infty$ for some function ℓ satisfying $\lim_{t \rightarrow \infty} \ell(xt)/\ell(t) = 1$, for all $x > 0$. Given that all marginal distributions in the portfolio have such tails with the same α , extreme dependence can be modeled by the so-called spectral measure. Indeed, a multivariate, regularly varying, random vector has univariate, regularly varying, distribution tails along each direction from the origin and a measure, the spectral measure, which describes the extreme dependence. More precisely, a random vector X in $\mathbb{R}^d$ is *regularly varying with index* $\alpha \in [0, \infty)$ and spectral measure σ on the Borel sets of the unit sphere $\mathbb{S}^{d-1} := \{z \in \mathbb{R}^d : \|z\| = 1\}$, if for all $x > 0$

$$\frac{P(\|X\| > xt, X/\|X\| \in \cdot)}{P(\|X\| > t)} \xrightarrow{v} x^{-\alpha} \sigma(\cdot) \quad t \rightarrow \infty \quad (24)$$

where $\|\cdot\|$ denotes any norm in $\mathbb{R}^d$ and $\xrightarrow{v}$ denotes vague convergence; for details see [17, 36–38] and references therein. This approach has been applied, for instance, in [39, 40] to analyze high-frequency foreign exchange (FX) data.

The second approach is based on Pickands’ representation of a multivariate extreme value distribution, which aims at the asymptotic frequency for

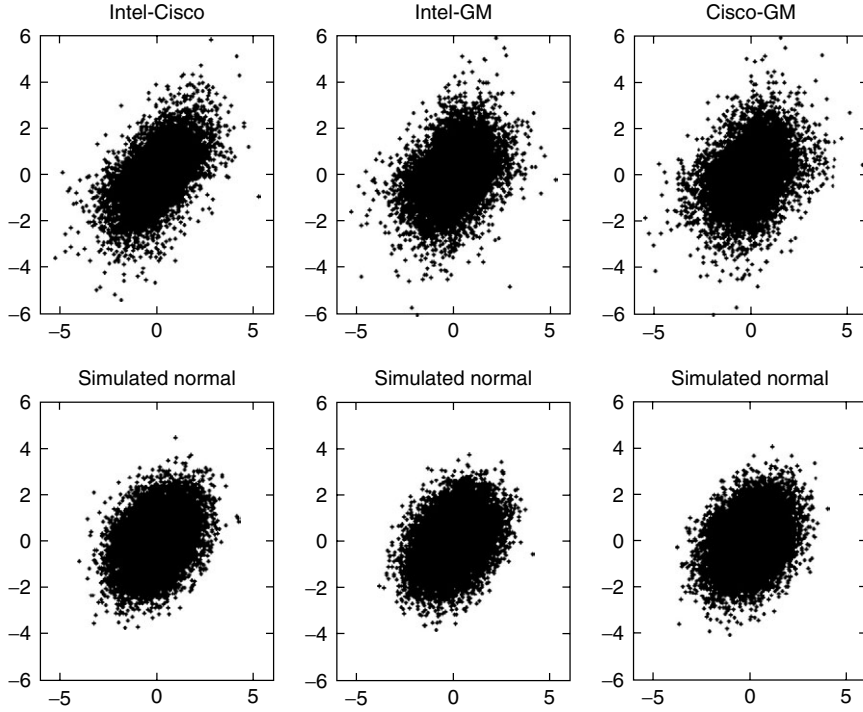


Figure 2 Bivariate stock data *versus* simulated normal random numbers with same (estimated) means and variances

joint large values in a portfolio. For two risk factors or assets X and Y with distribution functions F and G , respectively, the risk manager is interested in $P(X > x \text{ or } Y > y)$ for x, y large. This is measured by the quantity $(x \vee y = \max(x, y))$ and $x \wedge y = \min(x, y)$.

$$\begin{aligned} \Lambda(x, y) &= \lim_{n \rightarrow \infty} nP \left(F(X) > 1 - \frac{x}{n} \text{ or } G(Y) > 1 - \frac{y}{n} \right) \\ &= \int_0^{\pi/2} \left(\frac{x}{1 \vee \cot \theta} \vee \frac{y}{1 \vee \tan \theta} \right) \Phi(d\theta) \end{aligned} \quad (25)$$

and the limit exists under weak regularity conditions with finite measure Φ on $(0, \pi/2)$ satisfying

$$\int_0^{\pi/2} \frac{1}{1 \wedge \cot \theta} \Phi(d\theta) = \int_0^{\pi/2} \frac{1}{1 \wedge \tan \theta} \Phi(d\theta) = 1 \quad (26)$$

The measure Φ has to be estimated from the data, which can be done parametrically and nonparametrically. As it is more intuitive to estimate (dependence)

functions instead of measures, *Pickands' dependence function* and a so-called tail-dependence function have been introduced. They both aim at assessing joint extreme events. For Pickands' dependence function we refer to [13, Chapter 8], and [12, Section 8.2.5], where parametric and nonparametric estimation procedures have been suggested and investigated. The tail-dependence function has been proposed and estimated nonparametrically in [41]. Nonparametrical approaches are based on fundamental papers like [42]; for further references, see this paper.

Estimates as above can be improved for elliptical distributions or even for distributions with elliptical copula; cf. [43, 44]. The tail-dependence function is also invoked in [45], which includes an in depth analysis of multivariate high-frequency equity data.

Multivariate analogs to the POT method of the section "Extreme Value Theory" have been developed. We refer to [12, 13] for further details. Multivariate generalized Pareto distributions are natural models for multivariate POT data; see [46] and further references therein.

Acknowledgment

The article was partly written while the first author visited the Center for Mathematical Sciences of the Munich University of Technology. He takes pleasure in thanking the colleagues there for their hospitality. Financial support by the Chalmerska Forskningsfonden through a travel grant is gratefully acknowledged.

References

- [1] McNeil, A., Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management*, Princeton University Press, Princeton.
- [2] Embrechts, P. (ed) (2000). *Extremes and Integrated Risk Management*, UBS Warburg and Risk Books.
- [3] Rootzén, H. & Klüppelberg, C. (1999). A single number can't hedge against economic catastrophes, *Ambio* **28**, 550–555.
- [4] Campbell, R.A. & Huisman, R. (2003). Measuring credit spread risk: incorporating the tails, *Journal of Portfolio Management* **29**, 121–127.
- [5] Kuhn, G. (2004). Tails of credit default portfolios, Technical Report, Munich University of Technology, at <http://www-m4.ma.tum.de/Papers/>.
- [6] Lukas, A., Klaasen, P., Spreij, P. & Straetmans, S. (2003). Tail behavior of credit loss distributions for general latent factor models, *Applied Mathematical Finance* **10**(4), 337–357.
- [7] Phoa, W. (1999). Estimating credit spread risk using extreme value theory, *Journal of Portfolio Management* **25**, 69–73.
- [8] Böcker, K. & Klüppelberg, C. (2005). Operational VaR: a closed-form approximation, *Risk* 90–93.
- [9] Böcker, K. & Klüppelberg, C. (2006). Multivariate models for operational risk, Submitted for publication.
- [10] Chavez-Demoulin, V., Embrechts, P. & Nešlehová, J. (2005). Quantitative models for operational risk: extremes, dependence and aggregation, *Journal of Banking and Finance* **30**(10), 2635–2658.
- [11] Moscadelli, M. (2004). The modelling of operational risk: experience with the analysis of the data collected by the Basel committee, Technical Report Number 517, Banca d'Italia.
- [12] Beirlant, J., Goegebeur, Y., Segers, J. & Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*, John Wiley & Sons, Chichester.
- [13] Coles, S.G. (2001). *An Introduction to Statistical Modelling of Extreme Values*, Springer, London.
- [14] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.
- [15] Leadbetter, M.R., Lindgren, G. & Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer, New York.
- [16] Reiss, R.-D. & Thomas, M. (2001). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*, 2nd Edition, Birkhäuser, Basel.
- [17] Resnick, S.I. (1987). *Extreme Values, Regular Variation and Point Processes*, Springer, New York.
- [18] Stephenson, A. & Gilleland, E. (2006). Software for the analysis of extreme events: the current state and future directions, *Extremes* **8**(3), 87–109.
- [19] Emmer, S., Klüppelberg, C. & Trüstedt, M. (1998). VaR—ein Maß für das extreme Risiko, *Solutions* **2**, 53–63.
- [20] Smith, R.L. (1987). Estimating tails of the probability distributions, *Annals of Statistics* **15**, 1174–1207.
- [21] Jorion, P. (2001). *Value at Risk: The New Benchmark for Measuring Financial Risk*, McGraw-Hill, New York.
- [22] Artzner, P., Delbaen, F., Eber, J.M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**, 203–228.
- [23] Föllmer, H. & Schied, A. (2002). Convex measures of risk and trading constraints, *Finance and Stochastics* **6**(4), 429–447.
- [24] Drost, F.C. & Nijman, T.E. (1993). Temporal aggregation of GARCH processes, *Econometrica* **61**, 909–927.
- [25] Fasen, V., Klüppelberg, C. & Lindner, A. (2006). Extremal behavior of stochastic volatility models, in *Stochastic Finance*, A.N. Shiryaev, M.d.R. Grossinho, P.E. Oliveira & M.L. Esquivel, eds, Springer, New York, pp. 107–155.
- [26] Klüppelberg, C. (2004). Risk management with extreme value theory, in *Extreme Values in Finance, Telecommunication and the Environment*, B. Finkenstädt & H. Rootzén, eds, Chapman & Hall/CRC, Boca Raton, pp. 101–168.
- [27] Ferro, T.A. & Segers, J. (2003). Inference for clusters of extreme values, *Journal of the Royal Statistical Society, Series B* **65**, 545–556.
- [28] Laurini, F. & Tawn, J.A. (2003). New estimators for the extremal index and other cluster characteristics, *Extremes* **6**(3), 189–211.
- [29] Bollerslev, T., Engle, R.F. & Nelson, D. (1994). ARCH models, in *Handbook of Econometrics*, R. Engle & D. McFadden, eds, Elsevier, Amsterdam, Vol. IV, pp. 2959–3038.
- [30] McNeil, A. & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach, *Journal of Empirical Finance* **7**, 271–300.
- [31] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics (with discussion), *Journal of the Royal Statistical Society, Series B* **63**, 167–241.
- [32] Klüppelberg, C., Lindner, A. & Maller, R. (2004). A continuous time GARCH process driven by a Lévy process: stationarity and second order behaviour, *Journal of Applied Probability* **41**(3), 601–622.

- [33] Klüppelberg, C., Lindner, A. & Maller, R. (2006). Continuous time volatility modelling: COGARCH versus Ornstein-Uhlenbeck models, in *From Stochastic Calculus to Mathematical Finance. The Shiryaev Festschrift*, Y. Kabanov, R. Lipster & J. Stoyanov, eds, Springer, Berlin, pp. 393–419.
- [34] Beltratti, A. & Morana, C. (1999). Computing value at risk with high frequency data, *Journal of Empirical Finance* **6**(5), 431–455.
- [35] Dacorogna, M., Müller, U., Pictet, O. & de Vries, C. (2001). Extremal forex returns in extremely large data sets, *Extremes* **4**(2), 105–127.
- [36] Mikosch, T. (2005). How to model multivariate extremes if one must? *Statistica Neerlandica* **59**(3), 324–338.
- [37] Resnick, S.I. (2007). *Heavy Tail Phenomena: Probabilistic and Statistical Modeling*, Springer, New York.
- [38] Klüppelberg, C. & Resnick, S.I. (2008). The Pareto copula, aggregation of risks and the Emperor’s socks, *Journal of Applied Probability* **45**(1), To appear.
- [39] Hauksson, H., Dacorogna, M., Domenig, T., Müller, U. & Samorodnitsky, G. (2001). Multivariate extremes, aggregation and risk estimation, *Quantitative Finance* **1**(1), 79–95.
- [40] Stărică, C. (1999). Multivariate extremes for models with constant conditional correlations, *Journal of Empirical Finance* **6**, 515–553.
- [41] Hsing, T., Klüppelberg, C. & Kuhn, G. (2004). Dependence estimation and visualization in multivariate extremes with applications to financial data, *Extremes* **7**, 99–121.
- [42] Einmahl, J., de Haan, L. & Piterbarg, V.I. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution, *Annals of Statistics* **29**, 1401–1423.
- [43] Klüppelberg, C., Kuhn, G. & Peng, L. (2007). Semi-parametric models for the multivariate tail dependence function, *Under revision for Scandinavian Journal of Statistics*, To appear.
- [44] Klüppelberg, C., Kuhn, G. & Peng, L. (2007). Estimating the tail dependence of an elliptical distribution, *Bernoulli* **13**(1) 229–251.
- [45] Brodin, E. & Klüppelberg, C. (2006). *Modeling, Estimation and Visualization of Multivariate Dependence for High-Frequency Data*, Munich University of Technology, at <http://www.ma.tum.de/stat/>, Preprint.
- [46] Rootzén, H. & Tajiivdi, N. (2006). Multivariate generalized Pareto distribution, *Bernoulli* **12**(5), 917–930.

ERIK BRODIN AND CLAUDIA KLÜPPELBERG

Extreme Values in Reliability

Extreme value (EV) theory (*see Large Insurance Losses Distributions; Multistate Systems*) is basically developed from the maximum domain of attraction condition ([1, 2]; *see* [3, 4] for a recent introduction to the theory). It states that the distribution of the maximum (or the minimum) of a sample can be approximated, under very general conditions, by any one of the EV distributions, Gumbel (also called *Type I*), Fréchet (*Type II*), or Weibull (*Type III*); *cf.* the section titled “The Class of EV Distribution” (*see Copulas and Other Measures of Dependency*) for more details.

Alternatively, one can consider the sample of the exceedances. Given a high value (usually called the *threshold*), the exceedances are the positive differences (*see Structural Reliability*) between the sample values and the threshold. Then, the generalized Pareto (GP) distribution approximates the (conditional) distribution of the exceedances properly normalized; *cf.* the section titled “GP Distribution and Peaks over Threshold”.

For instance, in corrosion analysis, pits of larger depth are of primary interest as pitting corrosion can lead to the failure of metallic structures such as tanks and tubes. From the maximum domain of attraction condition, the class of EV distributions are the natural distributions to approximate to the distribution of the maximum pit depth. Thus, e.g., estimation procedures for parameters related to the distribution of the maximum such as the EV index (in the section titled “EV Index Estimation”), or related to the tail of the underlying distribution such as high quantiles (in section “Return Period and High-Quantile Estimation”), or failure probabilities (in section “Mean Excess Function, Model Fitting and Estimation”), are well studied.

The present discussion only deals with the simplest framework, univariate EV theory and samples $X_1, X_2, \dots, X_n$ supposedly drawn from independent and identically F distributed random variables. Some of the concepts discussed below have natural extensions to higher dimensions. For more on multivariate or infinite dimensional EV theory, *see* [4]. For EV theory concerning nonindependent

or nonidentically distributed random variables *see*, e.g., [4, 5].

The Class of EV Distributions

A distribution function F is said to be in the domain of attraction of some EV distribution (and consequently the maximum domain of attraction condition holds), if one can find real sequences $a_n > 0$ and b_n such that

$$\lim_{n \rightarrow \infty} \text{Prob} \left(\frac{\max(X_1, X_2, \dots, X_n) - b_n}{a_n} \leq x \right) = G(x) \quad (1)$$

for each continuity point x of G , and G is a nondegenerate distribution function. Then it is known that G must be one of the *EV distributions*, in the von Mises parameterization given by,

$$G_\gamma(x) = \exp \left(-(1 + \gamma x)^{-1/\gamma} \right), \quad 1 + \gamma x > 0 \quad (2)$$

(for $\gamma = 0$ the right-hand side should be interpreted as $\exp(-e^{-x})$) with γ real, a shape parameter called the *EV index*. Nonstandard forms are simply $G_\gamma(\sigma x + \mu)$ with $\sigma > 0$ the scale parameter and μ the location parameter.

The following distributions are often encountered as an alternative parameterization, corresponding respectively to the subclasses $\gamma = 0$, $-\gamma = 1/\alpha > 0$, and $\gamma = 1/\alpha > 0$: *double-exponential* or *Gumbel distribution*,

$$\Lambda(x) = \exp(-e^{-x}) \quad \text{for all real } x \quad (3)$$

Reverse Weibull distribution,

$$\Psi_\alpha(x) = \begin{cases} \exp(-(-x)^\alpha), & x < 0 \\ 1, & x \geq 0 \end{cases} \quad (4)$$

and *Fréchet distribution*,

$$\Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \exp(-x^{-\alpha}), & x > 0 \end{cases} \quad (5)$$

Distributions in the Gumbel domain of attraction are, e.g., normal, log-normal, gamma, exponential and, of course, the Gumbel itself. Distributions in the reverse Weibull domain of attraction are, e.g., beta and uniform. Distributions in the Fréchet domain of attraction are, e.g., Cauchy and Pareto.

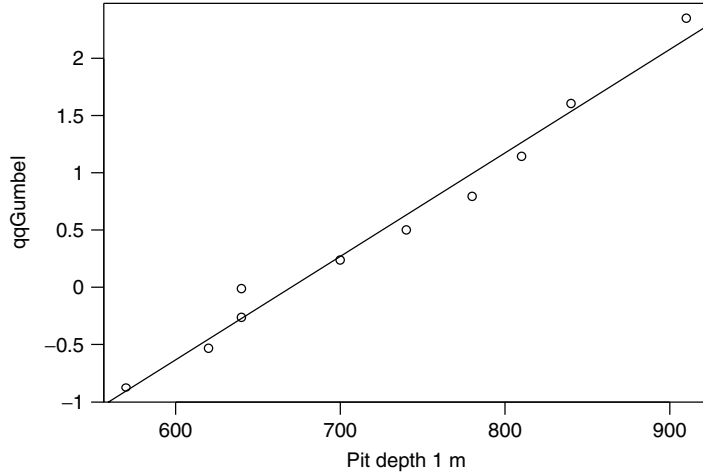


Figure 1 Gumbel Q–Q plot

Main properties distinguishing the three subclasses (equations (3)–(5)) are:

- For $\gamma = 0$ the right endpoint of the limiting distribution, i.e., $\lim_{x \rightarrow \infty} \Lambda(x)$, equals infinity, and the distribution is rather light tailed (it decays exponentially). All moments are finite.
- The reverse Weibull distribution ($\gamma < 0$) has right endpoint equal to $-1/\gamma$, it decays as a power law, and all moments exist.
- The right endpoint of the Fréchet distribution is infinity and the distribution has a rather heavy right tail, decaying as a power law. Moments of order greater than or equal to $1/\gamma$ do not exist.

Since $\min(X_1, X_2, \dots, X_n) = -\max(-X_1, -X_2, \dots, -X_n)$, EV distributions for the minimum have a one-to-one correspondence with the ones for the maximum, say $G^*(x) = 1 - G(-x)$, for all x . For instance, the *Weibull distribution* is one of the possible limiting distributions for conveniently normalized minima (i.e., when $\gamma < 0$),

$$\Psi_\alpha^*(x) = 1 - \Psi_\alpha(-x) = \begin{cases} 0, & x \leq 0 \\ 1 - \exp(-x^\alpha), & x > 0 \end{cases} \quad (6)$$

Application 1: The Gumbel Model for Maximum Pit Depth

The observations (in micrometers) correspond to the maximum pit depth developed on $N = 10$ metal

samples after 1 month exposure to tap water [6]: (570, 620, 640, 640, 700, 740, 780, 810, 840, 910). We shall fit a model to the maxima, and from it estimate the return period and high quantiles.

Gumbel Q–Q Plot and Model Fitting

The Gumbel Q–Q plot provides a quick “visual selection method” for distribution fitting. Plotting the ordered sample against the Gumbel quantiles $\Lambda^{-1}(i/(N+1)) = (-\log(-\log(i/11)))$, $i = 1, \dots, N$, the observations fit quite well in a straight line (cf. Figure 1). Generally one decides for the Gumbel, Weibull, or Fréchet distributions if the points follow approximately a straight line, convex or concave curves, respectively.

More advanced tests can also be applied. For instance, for the “common” likelihood ratio test, which in this case tests within the EV model, $H_0 : \gamma = 0$ versus $H_1 : \gamma \neq 0$, we obtained a p value of 0.69, hence inducing for not rejecting the null hypothesis. Tests for checking the EV condition can be found in [4]. Specific tests for selecting EV distributions within the maximum domain of attraction is, e.g., the Hasofer and Wang test [7]. A class of goodness-of-fit tests for the Gumbel model was investigated by Stephens [8].

Now back to the application, taking the Gumbel as the parametric model fitted to the distribution of the maximum pit depth, the maximum likelihood estimates of the location and scale parameters are,

respectively, $\mu = 696.43$ and $\sigma = 52.13 \mu\text{m}$. These estimates were obtained numerically as there are no explicit analytical expressions for the maximum likelihood estimators in the Gumbel model. Parameter estimation for EV distributions is discussed, e.g., in [3, 9]; [10], a paper discussing on available software for EV theory.

Return Period and High-Quantile Estimation

The return period (*see Bayesian Statistics in Quantitative Risk Assessment*) at a level u say, is a common characteristic to estimate in these kinds of problems. In the example, it corresponds to the number of coupons that on an average must be exposed to tap water in order to obtain a pit depth greater than u . Solely based on the sample, we would have 10 coupons (the sample size) exposed for the return period at $850 \mu\text{m}$ (the sample maximum). Then note that the sample is of no use anymore if u is larger than the sample maximum. Yet, using the fitted parametric model, extrapolation beyond the range of the available data is possible. From the fitted Gumbel distribution one obtains for the return period at u ,

$$\begin{aligned} \frac{1}{1 - F(u)} &\approx \frac{1}{1 - \Lambda_{\hat{\mu}, \hat{\sigma}}(u)} \\ &= \frac{1}{1 - \exp(-e^{-(u-696.43)/52.13})} \end{aligned} \quad (7)$$

The reverse problem of estimating the level u for a mean of C coupons exposed to corrosion, leads to a major problem in EV statistics, high-quantile estimation, high meaning that estimates may be beyond the sample maximum. Again using the fitted model information, in this case the inverse function of the fitted distribution, we get the estimator

$$\begin{aligned} \hat{u} &= \Lambda_{\hat{\mu}, \hat{\sigma}}^{\leftarrow}(1 - 1/C) \\ &= 696.43 - 52.13 \times \log(-\log(1 - 1/C)) \end{aligned} \quad (8)$$

For instance, $C = 100$ corresponds to the 0.99-quantile estimate, $\hat{u} = \Lambda_{\hat{\mu}, \hat{\sigma}}^{\leftarrow}(0.99) = 936.24 \mu\text{m}$. Note that this value is beyond 910, the sample maximum.

GP Distributions and Peaks over Threshold

The maximum domain of attraction condition is equivalent to the following: there exists a positive real function f such that

$$\lim_{t \rightarrow x^*} \frac{\text{Prob}(X > t + xf(t))}{\text{Prob}(X > t)} = (1 + \gamma x)^{-1/\gamma} \quad (9)$$

for all x for which $1 + \gamma x > 0$, and $x^* = \sup\{x : F(x) < 1\}$ denotes the right endpoint of the underlying distribution. The left-hand side of equation (9) gives the *generalized Pareto (GP) distributions*,

$$H_{\gamma}(x) = 1 - (1 + \gamma x)^{-1/\gamma} \quad (10)$$

(for $\gamma = 0$ the right-hand side should be interpreted as $1 - e^{-x}$) for $x > 0$ if $\gamma \geq 0$ and $0 < x < -1/\gamma$ otherwise, as the class of distributions to approximate to the conditional distribution of $(X - t)/f(t)$ given X is above the threshold t . The parameter γ is the same EV index as in the EV distributions.

Particular cases of GP distributions are when $\gamma = 0$, with the exponential distribution, $\gamma > 0$ with Pareto distributions, and $\gamma = -1$ with the uniform distribution. Nonstandard forms are straightforward as in the EV class. A characteristic property of the GP distribution is its stability with respect to conditioning: if X is a random variable GP distributed, then the conditional distribution of $X - t$ given $X > t$, for any t in the range of possible values of X , is also GP with the same EV index.

Estimation of the GP parameters, in a parametric setting, is well studied (e.g., [9]). The semiparametric approach connected with equation (9), *peaks over threshold (POT)*, (*see Extreme Value Theory in Finance*) is more interesting to EV analysis. In this case, statistical inferences in the tail of distributions are based on the excesses over a high threshold whose distribution is approximated by a GP distribution.

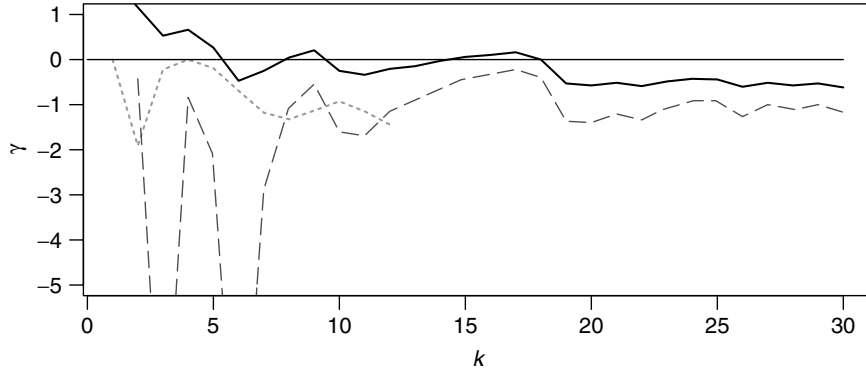
EV Index Estimation

EV index estimation is important in EV statistics since any inference concerning tails of distributions necessarily involves the estimation of γ (recall that the approximate distributions depend on this parameter).

For a sample of size n , define the order statistics (*see Uncertainty Analysis and Dependence Modeling*) as $X_{1,n} \leq X_{2,n} \leq \dots \leq X_{n,n}$. It is common in

Table 1 Fatigue data, sample size equal to 49

1.051	4.921	7.886	10.861	13.520	1.337	5.445	8.108	11.026
13.670	1.389	5.620	8.546	11.214	14.110	1.921	5.817	8.666
11.362	14.496	1.942	5.905	8.831	11.604	15.395	2.322	5.956
9.106	11.608	16.179	3.629	6.068	9.711	11.745	17.092	4.006
6.121	9.806	11.762	17.568	4.012	6.473	10.205	11.895	17.568
4.063	7.501	10.396	12.044	—	—	—	—	—

**Figure 2** EV index estimation for the fatigue data: dotted line for Pickands, dashed for moment and straight for PWM

EV theory to use only the k th highest order statistics, $X_{n-k,n} \leq \dots \leq X_{n,n}$, to estimate γ , which can be seen as an application of the POT approach with the random threshold $t = X_{n-k,n}$. Solutions of approximate maximum likelihood estimators require numerical methods and are theoretically well known. Among estimators with explicit formulas is Hill's estimator [11], $\hat{\gamma} = k^{-1} \sum_{i=0}^{k-1} \log X_{n-i,n} - \log X_{n-k,n}$, very frequently cited though it is consistent only for positive values of γ (see below for other estimators and properties).

Estimation of the EV index is vastly studied with known theoretical properties where, for instance, confidence intervals and testing may be straightforward (cf. [3, 4]).

Application 3: GP in a Fatigue Problem

Next we apply the POT approach to the fatigue data for the Kevlar/Epoxy strand lifetime (in hours) at a stress level equal to 70% [12] (cf. Table 1). With this kind of data one is usually interested in the left tail of the underlying distribution. The GP distribution is applicable as in equation (9), by taking the symmetric

sample, which transforms the lower tail into the upper tail.

Mean Excess Function, Model Fitting, and Estimation

Since GP distributions are the only distributions where the *mean excess function* (also called *mean residual life function*), $E(X - t | X > t)$, $0 \leq t \leq x^*$, is a straight line, its empirical version provides a common diagnostic tool for model fitting. If, as a function of the threshold, it is close to a straight line, it indicates a possible good GP fitting. This is the case for the present data. More advanced tests for model fitting can also be employed and given the equivalence between equations (1) and (9), tests mentioned before can be applied.

Proceeding with model fitting, in Figure 2 are estimates of the EV index, as a function of the number of upper order statistics used in the estimation (above denoted by k). We illustrate the three estimators: moment [13], Pickands [14], and probability weighted moment (PWM) [15]. A common “problem” with these estimators and with the POT approach, in general, is the effect of the choice of the threshold on the results. This can be seen in

Figure 2, the usual variance-bias trade-off. The estimators usually show large variance for small values of k (i.e., few sample information) and large bias for large values of k (i.e., where one might be escaping from the EV condition). Though some optimal threshold choice procedures are available (e.g. [16–18]), for a “quick guess” one usually looks for the estimate in a stable region of the curves; in Figure 2, for example, -1 may be the guess.

Having a satisfactory fit to the tail, extrapolation beyond the sample range is possible and reasonable from EV theory. For instance, high quantile estimation, which for the present data would mean the estimation of some fatigue limit that would be exceeded with some given low probability. Or, the reverse problem for some given low fatigue threshold to estimate the failure probability. The estimation of the left endpoint of the pertaining distribution could be similarly addressed; note that for the present data there is an indication of a negative EV index, which corresponds to a finite endpoint of the distribution under study; cf. [4] for more on these estimators.

Further Reading and References

Many references on the applications of EV theory in reliability are easily available. Just to mention a few, included in literature more specialized in probability and statistics: in [19] the Weibull distribution was first used in the analysis of breaking strengths of metals; in [20] are inferences based on GP distributions on the sizes of large inclusions in bulks of steel; the book [21] contains a number of references on applications in structural engineering, hydraulics engineering, material strength, fatigue strength, electrical strength of materials, etc. More of a theoretical nature are: Galambos [22] with motivation of applications to strength of materials and failure of equipments; Harris [23] to sequential system lifetimes; Schuller [24] to structural engineering; Taylor [25] with an extension of the Weibull model in fitting tensile strength data; Coles and Tawn [26] with an application of multivariate extremes to structural design.

References

- [1] Fisher, R.A. & Tippett, L.H.C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society* **24**, 180–190.
- [2] Gnedenko, B.V. (1943). Sur la distribution limite du terme du maximum d’une série aléatoire, *Annals of Mathematics* **44**, 423–453.
- [3] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, Berlin.
- [4] de Haan, L. & Ferreira, A. (2006). *Extreme Value Theory: An Introduction*, Springer, New York.
- [5] Leadbetter, M.R., Lindgren, G. & Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer, Berlin.
- [6] Kuhlra, C.B. (1967). *Applications of extreme value theory in the reliability analysis of non-electronic components*, Master thesis, Air Force Institute of Technology, Wright-Patterson AFB.
- [7] Hasofer, A.M. & Wang, Z. (1992). A test for extreme value domain of attraction, *Journal of the American Statistical Association* **87**, 171–177.
- [8] Stephens, M.A. (1977). Goodness of fit for the extreme value distribution, *Biometrika* **64**, 583–588.
- [9] Johnson, N.L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions*, John Wiley & Sons, New York, Vols. 1 and 2.
- [10] Stephenson, A. & Gilleland, E. (2006). Software for the analysis of extreme events: the current state and future directions, *Extremes* **8**, 87–109.
- [11] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics* **3**, 1163–1174.
- [12] Andrews, D.F. & Herzberg, A.M. (1985). *A Collection of Problems from Many Fields for the Student and Research Worker*, Springer-Verlag, New York.
- [13] Dekkers, A.L.M., Einmahl, J.H.J. & de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution, *Annals of Statistics* **17**, 1833–1855.
- [14] Pickands, J. (1975). Statistical inference using extreme order statistics, *Annals of Statistics* **3**, 119–131.
- [15] Hosking, J.R.M. & Wallis, J.R. (1987). Parameter and quantile estimation for the generalized Pareto distribution, *Technometrics* **29**, 339–349.
- [16] Beirlant, J., Vynckier, P. & Teugels, J.L. (1996). Tail index estimation, Pareto quantile plots and regression diagnostics, *Journal of the American Statistical Association* **91**, 1659–1667.
- [17] Danielsson, J., de Haan, L., Peng, L. & de Vries, C.G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation, *Journal of Multivariate Analysis* **76**, 226–248.
- [18] Drees, H. & Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation, *Stochastic Processes and Their Applications* **75**, 149–172.
- [19] Weibull, W. (1952). Statistical design of fatigue experiments, *Journal of Applied Mathematics* **19**, 109–113.
- [20] Anderson, C.W. & Coles, S.G. (2003). The largest inclusions in a piece of steel, *Extremes* **5**, 237–252.
- [21] Castillo, E., Hadi, A., Balakrishnan, N. & Sarabia, J.M. (2005). *Extreme Value and Related Models with*

- Applications in Engineering and Science*, John Wiley & Sons, Hoboken.
- [22] Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*, 2nd Edition, Krieger, Malabar, pp. 188–196.
- [23] Harris, R. (1970). An application of extreme value theory to reliability theory, *Annals of Mathematical Statistics* **41**, 1456–1465.
- [24] Schuëller, G.I. (1984). Application of extreme values in structural engineering, in *Statistical Extremes and Applications: Proceedings of the NATO Advanced Study Institute on Statistical Extremes and Applications*, J. Tiago de Oliveira, ed, Reidel Publications, pp. 221–234.
- [25] Taylor, H.M. (1994). The Poisson-Weibull flaw model for brittle fiber strength, in *Extreme Value Theory and Applications: Proceedings of Gaithersburg Conference*, J. Galambos, J. Lechner & E. Simiu, eds, Kluwer Academic Publishers, Vol. I, pp. 43–59.
- [26] Coles, S.G. & Tawn, J.A. (1994). Statistical methods for multivariate extremes – an application to structural design, *Journal of the Royal Statistical Society, Series C* **43**, 1–48.

ANA FERREIRA

Extremely Low Frequency Electric and Magnetic Fields

Exposure

Power-frequency electric and magnetic fields (EMFs) occupy the extremely low frequency (ELF) portion of a wider electromagnetic spectrum encompassing frequencies that range from above approximately 10^{20} Hz for γ rays at the high end of the spectrum to below power frequencies of 50–60 Hz at the low end. Between the two ends of the spectrum, in order of decreasing frequency, are X rays, ultraviolet radiation, visible light, infrared radiation, microwaves, and radio waves.

ELF EMFs are associated with all aspects of the production, transmission, and use of electricity. The fields are imperceptible and are ubiquitously present in modern societies. ELF EMFs are composed of two separate components, electric fields and magnetic fields. Electric fields are created by electric charges and are measured in volts per meter (V m^{-1}). Typical residential exposure levels are under 10 V m^{-1} . In the immediate vicinity of electric appliances exposure levels can reach as high as few hundreds of volts per meter, whereas exposure levels immediately under high-tension power lines can reach several thousand volts per meter, or several kilovolts per meter (kV m^{-1}). Magnetic fields are created by moving electric charges. The unit of measurement for magnetic fields is tesla (T). Another commonly used unit in engineering sciences is gauss (G); 1 T is equivalent to $10\,000 \text{ G}$. Environmental exposure levels are described in microtesla (μT), or 10^{-6} T . Typical residential exposure levels are around $0.1 \mu\text{T}$. In the immediate vicinity of electric appliances that are in use, fields could be as high as several hundreds of microtesla, and immediately under high-tension power lines can reach up to $20 \mu\text{T}$.

Epidemiology

Since the late 1970s, numerous epidemiologic studies have investigated possible health risks from residential and occupational exposure to EMFs. Most of

some 20 epidemiologic studies investigating childhood leukemia and ELF magnetic fields have found an association between higher than average magnetic field exposure levels and leukemia risk, although a robust biologic hypothesis that could explain this association is lacking. In 2000, two independently conducted pooled analyses of previously published studies showed a statistically significant, approximately twofold increase in childhood leukemia risk for average residential exposure levels above 0.3 or $0.4 \mu\text{T}$, compared to the lowest exposure category of $\leq 0.1 \mu\text{T}$ [1, 2]. Risk increases this small or smaller are notoriously hard to evaluate in epidemiology because it is usually difficult to achieve enough precision to distinguish a small risk from no risk. Such small effect estimates, compared to larger ones, are also more likely to result from inadvertent error, or bias, that can occur in epidemiologic studies [3]. One type of bias that often goes undetected and unmeasured is confounding, which occurs when another exposure that is present along with the exposure under study is actually responsible for observed effects; another is selection bias, which can arise in the process used to select or enroll study participants [4]. Given the small associations observed in studies of magnetic fields and childhood leukemia, a limited understanding of causal risk factors for childhood leukemia, and methodological difficulties such as exposure assessment and the potential for selection bias, a conclusive interpretation of these findings remains a challenge. The lack of a robust biophysical mechanism that would explain how environmental magnetic fields could cause cancer and a lack of support from laboratory investigations further detract from the epidemiologic findings.

Epidemiologic investigation of two health outcomes, breast cancer and cardiovascular disease, has been motivated by biologically based hypotheses. Although initial studies appeared to support these hypotheses, recent, larger, and more rigorous studies have found no effects and have failed to confirm earlier findings, suggesting that ELF magnetic fields do not play a role in the development of breast cancer or cardiovascular disease.

A number of additional health outcomes (including adult leukemia, adult and childhood brain cancer, motor neuron diseases, and reproductive outcomes) have also been investigated in epidemiologic studies. Most of these outcomes were investigated to a lesser extent than childhood leukemia and the results

are inconsistent. Most of these studies also were not based on specific biologic hypotheses and lack supportive laboratory evidence.

Risk Assessment

During the past decade, a number of national and international expert panels, including panels assembled by the US National Institute of Environmental Health Sciences (NIEHS), the International Agency for Research on Cancer (IARC), the World Health Organization (WHO) [5] have reviewed the available evidence on the potential relationship between exposure to ELF EMF and various adverse health outcomes [6, 7]. Evaluations by these expert panels generally agree that short-term, adverse cellular effects do not occur with exposures to magnetic fields below 100 μ T.

The NIEHS, IARC and WHO classified ELF magnetic fields as a “possible human carcinogen”, or a group 2B carcinogen. This classification was based both on epidemiologic evidence showing a consistent association between exposure to ELF magnetic fields and childhood leukemia and on laboratory studies in animals and cells, which do not support an association between exposure to ELF magnetic fields and cancer development. The NIEHS assessment (but not the IARC or WHO assessment) concluded that there was also sufficient epidemiologic evidence for an association between adult chronic lymphocytic leukemia and occupational exposure to ELF magnetic fields to warrant a 2B classification.

Several risk characterization attempts have indicated that the public health impact of residential magnetic fields on childhood leukemia is likely to be limited [5, 8]. Positing sensitivity and bias models for methodologic problems in the epidemiologic studies and accounting for uncertainties regarding the prevalence of field levels that may have effects suggest that, in light of the available data, both no public health impact and a large impact remain possibilities [3].

The combination of widespread exposures, established biological effects from acute, high-level exposures, and the possibility of leukemia in children from low-level, chronic exposures have made it both necessary and difficult to develop consistent public health

policies. In view of these uncertainties, it might be advisable to adopt general no- and low-cost exposure reduction measures [9].

References

- [1] Ahlbom, A., Day, N., Feychting, M., Roman, E., Skinner, J., Dockerty, J., Linet, M., McBride, M., Michaelis, J., Olsen, J.H., Tynes, T. & Verkasalo, P.K. (2000). A pooled analysis of magnetic fields and childhood leukaemia, *British Journal of Cancer* **83**, 692–698.
- [2] Greenland, S., Sheppard, A.R., Kaune, W.T., Poole, C. & Kelsh, M.A. (2000). A pooled analysis of magnetic fields, wire codes, and childhood leukemia, *Epidemiology* **11**, 624–634.
- [3] Greenland, S. & Kheifets, L. (2006). Leukemia attributable to residential magnetic fields: results from analyses allowing for study biases, *Risk Analysis* **26**, 471–481.
- [4] Mezei, G. & Kheifets, L. (2006). Selection bias and its implications for case-control studies: a case study of magnetic field exposure and childhood leukemia, *International Journal of Epidemiology* **35**, 397–406.
- [5] WHO – World Health Organization (2007). *Extremely Low Frequency Fields Environmental Health Criteria*, World Health Organization, Geneva, Vol. 238.
- [6] National Institute of Environmental Health Sciences (NIEHS) (1999). *NIEHS Report on Health Effects from Exposure to Power-Line Frequency Electric and Magnetic Fields*, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NIH Publication No. 99–4493.
- [7] International Agency for Research on Cancer (IARC) (2002). *Non-Ionizing Radiation, Part 1: Static and Extremely Low-Frequency (ELF) Electric and Magnetic Fields*, in *Monographs of the Evaluation of Carcinogenic Risks to Humans*, Lyon, Vol. 80.
- [8] Kheifets, L., Afifi, A. & Shimkhada, R. (2006). Public health impact of extremely low frequency electromagnetic fields, *Environmental Health Perspectives* **114**, 1532–1537.
- [9] Kheifets, L., Sahl, J., Shimkhada, R. & Repacholi, M. (2005). Developing policy in the face of scientific uncertainty: interpreting 0.3 μ T or 0.4 μ T cut points from EMF epidemiologic studies, *Risk Analysis* **25**, 927–935.

Related Articles

Environmental Hazard

Environmental Health Risk

Risk and the Media

LEEKA KHEIFETS AND GABOR MEZEI

Failure Modes and Effects Analysis

Failure mode and effects analysis (FMEA) (*see* **Human Reliability Assessment; Reliability Integrated Engineering Using Physics of Failure; Probabilistic Risk Assessment**) is a systematic safety analysis method that identifies the possible system failure modes associated with a system and evaluates the effects on the operation of the system, should the failure mode occur. The focus of the analysis is on using the analysis to improve the safety of the system design (*see* **Reliability Data**).

Engineers have always informally considered what would happen if a part of the system they were designing failed. FMEA formalizes this process, ideally considering all possible failure modes, and deriving from them the effects on the system caused by the occurrence of any failure. In order to cover all aspects of the system, the FMEA process will typically bring together a team of engineers involved in all aspects of the design of the system. Once the team has examined the consequences of all possible failure modes, they should plan actions to mitigate the risks or effects of failure, documenting the actions taken and their intended effects.

FMEA, as a formal analysis discipline, originated with the American military, and was documented in military standard 1629A [1]. The automotive industry developed the practice of FMEA further, and developed a standard for performing automotive FMEA as part of its QS-9000 quality process [2]. Although QS-9000 has now been superseded by ISO/TS16949, which merges QS-9000 with European quality standards, the requirement to perform FMEA remains a part of the process. The Society of Automotive Engineers has produced two sets of standards for performing FMEA: SAE J1739 [3] for automobile applications and ARP5580 [4] for nonautomobile applications.

Types of FMEA

The main distinction between types of FMEA is between product FMEA (*see* **No Fault Found**) and process FMEA. Product FMEA considers the design for a product and concentrates on the effect of failure

of components of the design on the overall product. Process FMEA (*see* **Reliability of Consumer Goods with “Fast Turn Around”**) focuses on the potential for mistakes in the manufacturing process, which can result in the production of faulty components or systems.

FMEA can be performed on anything from a single screw up to a full car or aircraft. It will typically be requested from component suppliers by subsystem or system suppliers and from the Tier 1 suppliers by the automobile and aircraft manufacturers. FMEA effects at the component level can be fed as failure modes into higher levels of the design, forming a hierarchy of FMEAs. The severity of a failure effect at the component level cannot always be meaningfully evaluated, as it is dependent on the use of the component. This can mean that top level system FMEA results might need to be fed down to the component manufacturers through a demand for greater reliability in a specific component of the system where it is identified as vital to the safety of the overall system.

The FMEA Process

Overview

The process of producing and acting upon an FMEA report comprises the following steps:

- identify the structure and the boundaries of the system for the purposes of this FMEA;
- convene a relevant team of experts;
- identify the possible failure modes and effects;
- decide the failure effects for each possible failure mode;
- evaluate the significance of the failure mode;
- rank the failure modes by significance;
- act to mitigate the significance of the most important failure modes;
- reevaluate the significance of those failure modes taking the new action into account.

Each of these steps is explained in greater detail in the following, using an electrical system design FMEA as an example. A more extensive example of carrying out an FMEA is given in [5].

Identify the Structure and the Boundaries of the System for the Purposes of this FMEA

Typically, a product will be broken down into a set of systems in order to make FMEA tractable. It is necessary to identify the boundaries of the system in order to make sure that all aspects of the product are covered by the FMEA activity, without unnecessary duplication. For example, in a vehicle, if every electrical system considered the fusing strategy for the vehicle, there would be a great deal of wasted effort. Typically in that case, there would be a separate analysis of the fusing strategy, and the results of that fusing analysis would be used in the individual electrical FMEAs.

The available information should be identified – for electrical system design FMEA, this would be a circuit diagram for the system plus a detailed specification for the operation of each component in the system. For a process FMEA of a mechanical component, the available information might be the specification of the component plus a process flowchart for how the component is produced. For design FMEA of a more complex system, the appropriate information might be a functional block diagram.

Convene a Team of Relevant Experts

Typically, a range of people will have to be involved in specifying and designing the system undergoing FMEA. The FMEA is best carried out by a team having all the knowledge needed to design the system in the first place. So, if the design needed the involvement of electrical engineers, computer programmers, mechanical engineers, and diagnostics experts, then the FMEA activity also needs their involvement. Given the diverse experience of the team, it is often good for the team to walk through correct operation of the system, to ensure that all members of the team have a shared, correct understanding of the system. For the process FMEA of a component, the team will need to understand the operations involved in producing the component or system (e.g., lathing and welding), as well as how it works.

Identify the Possible Failure Modes and Effects

Where there is previous experience of performing FMEA on similar systems, both the potential failure modes and the possible effects on the system will be

known to a great extent. For example, an electrical relay whose switch contacts are open when the coil has no current flowing through it and are closed when current is flowing through it, will have known failure modes. It could fail such that it is open when the current is flowing through it, or such that it is closed when no current is flowing through it. Where there are new types of components in the design, then brainstorming among the team of experts may be needed to generate all the ways in which the component could fail. For more complex systems, the lower level items might be subsystems rather than simple components such as relays, and then it is often better to think of them in functional terms rather than physical terms. For example, a voltage regulator might fail to regulate the voltage, or might regulate it to the wrong value.

Similarly, where a new type of system is being designed, then brainstorming among the experts can help identify all the possible effects of failure. It is tempting to think of the system-level effects as being produced from analyzing the effects of failure, but it is better to identify all possible system-level effects initially, since they can then be used in a consistent manner as per the following steps.

Decide the Failure Effects for Each Possible Failure Mode

When the potential failure modes are known, the team of experts can mentally simulate the behavior of the system when each failure mode has occurred, and can predict which of the identified effects will occur for that failure.

Evaluate the Criticality of the Failure Mode

The significance of the failure mode can be calculated in different ways. A common method in the automotive industry is to produce a risk priority number (RPN). RPN is calculated by multiplying three different characteristics of the failure mode: severity, detection, and occurrence, each assessed on a scale of 1–10. Severity expresses how bad the effect will be, where 1 is *no effect* and 10 is *hazardous without warning*. If all the possible effects are listed before calculating what happens in failure situations, then severity can be associated with the effect at that point, and used consistently. Detection expresses whether the present design controls will detect the

presence of potential cause of failure, and therefore the failure mode before the system is released for production, where 1 means the design controls will almost certainly detect the problem, and 10 means the design control cannot detect the problem. Occurrence expresses the likelihood that the failure mode will occur, where 1 on the scale might represent a failure rate of 1 in 1 500 000 and 10 might represent a failure rate of 1 in 2.

Rank the Failure Modes by Significance

The failure modes can be ranked by the significance of their occurrence on the system, where a high RPN would be given a high significance. In addition, all failure modes where the effect has a high severity also merit detailed attention.

Act to Mitigate the Significance of the Most Important Failure Modes

High RPNs can be brought down in several ways. The severity of the effect of a failure mode can be reduced by adding redundancy, or the occurrence can be reduced by using more reliable components, or the detection value can be reduced by adding design verification or validation actions.

Reevaluate the Significance of Those Failure Modes Taking the New Action into Account

When changes have been made to the design, the RPN for the failure mode that has been addressed can be recalculated, and a decision whether further actions are necessary can be made.

Documentation of the FMEA Process

The standard way of documenting the results of an FMEA has long been the tabular approach [1], where one row of the table gives the analysis results for one failure mode. There is a column for each of the details of the FMEA, such as failure mode name, effect description, severity, occurrence and detection values, remedial action, and recalculated values.

The results can be documented in a spreadsheet, but there are specialized programs for documenting FMEAs. Some of those tools provide useful reordering of the FMEA results, such as bringing together

all equivalent failure modes (all failure modes that cause the same effect). Other tools have a database for storing and reusing failure mode names and effect descriptions so that they can be chosen from lists rather than typed into the report.

Automation of FMEA

The production of a system design FMEA report is very repetitive, with many similar failure situations being investigated. However, it is also necessary to have enough insight into the system to detect the situations that look similar to fairly innocuous situations but have much more severe consequences – hence the need for a team of experts to derive the results.

Much of the repetitive effort can be avoided by automating the generation of effects for each failure mode [6]. This works best for domains where it is possible to compose the behavior of the whole system from a description of the components of the system, the connections between them, and their individual behavior. Electrical or hydraulic systems are good examples of domains where this can be achieved. Compositionality of simulation means that the component that has the failure mode can be replaced by a faulty version, and the faulty behavior can be generated. The simulation of the system containing the failure mode can then be compared with the correct simulation enabling generation of the effects for each failure mode. This can be used to generate RPN values automatically.

There are commercial tools available that use simulation to generate the effects caused by each failure node and assign RPNs. Where these are used, it is important that they are treated as part of a wider process, where an FMEA team still looks at the ranked failure modes and decides what action should be taken to make the design safer. One of the main benefits of FMEA is the clearer understanding that the analysis gives the FMEA team about how the product works, and without this wider context, that benefit would be lost in an automated FMEA.

Further Issues

Multiple Failures

FMEA is addressed at single-point failures, but some of the worst design flaws are caused by multiple

failures. Consideration of some of these flaws can be addressed by the FMEA team identifying common types of multiple failures – for example, a failure detection mechanism failing before the failure it is intended to detect, but, in general, it would be too tedious for an FMEA team to explore even all pairs of failures.

Automation of FMEA, where possible, can address this, and Price and Taylor describe how this has been done for the automotive electrical domain, making an exploration of many failure mode combinations both feasible and useful.

Alternatives to RPN

While RPN has proved to be a popular method of assessing failure mode significance, especially within the automotive industry, it has flaws [5]. Perhaps the worst of these flaws is that it puts equal emphasis on detection of potential problems at manufacturing time as it does for likelihood of occurrence and for severity of effect. One of the reasons for this is the automobile industry's focus on avoiding warranty costs.

Other industries put more emphasis on safe and reliable operation over an extended lifetime, and neither MIL-STD-1629A [1] nor ARP5580 [4] has the same focus on detection at the time of manufacture.

The use of RPNs originally focused on RPNs that exceeded a set level (overall RPN of 100), whereas the other ways of assessing significance have always looked at ranking the failure mode/effect pairs and concentrating on the most important ones. So MIL-STD-1629A had only four levels of severity (catastrophic, critical, marginal, and minor). It recommended either a quantitative ranking of occurrence to five levels (frequent, reasonably probable, occasional, remote, and extremely unlikely) or a quantitative approach based on known failure rates. Severity and occurrence were plotted on a criticality matrix, enabling people to concentrate on those with the highest combination of severity and likelihood of occurrence.

Whatever method is used to assign significance to failure modes, it is important that it is looked on as a ranking among the failure modes rather than as an absolute value. If the more serious problems with a design or process are always addressed when an FMEA is performed, then this can form a firm foundation for continual product or process improvement.

Contribution of FMEA to Quantitative Risk Assessment

FMEA is a qualitative risk assessment discipline. It produces a ranked list of the failures that have the most impact on the safe operation of the system being studied. However, it can be used as a basis for further quantitative risk assessment. The FMEA is often seen as giving a necessary understanding of how a system works and how it can fail. If the FMEA only considers single failures, then it does not provide sufficient information to construct a fault tree for a specific top event, but has valuable supporting information. A multiple failure FMEA [7] potentially does provide enough information to construct the fault tree (*see Systems Reliability; Imprecise Reliability*). The fault tree can then be used for quantitative risk assessment in the standard way.

Wider Application of FMEA

FMEA originated as a way of identifying the potential effects of failure on the safety of engineered systems. Besides the well-known areas of automotive and aeronautic systems, FMEA has become a required part of the development for several types of naval craft [8].

FMEA has also started to make inroads into healthcare outside the obvious area of design FMEA of medical equipment. The US Department of Veteran Affairs National Center for Patient Safety has developed a healthcare FMEA [9]. It is a process FMEA for healthcare processes, focused on ensuring patient safety, and is based on developing a process flow diagram and identifying and exploring the ways in which each process in the flow can fail. Such work might well be useful in other areas where people are carrying out critical processes.

A second innovative application of FMEA in healthcare is the use of design FMEA to assess the quality of a new hospital [10]. The structure of the design being analyzed was block diagrams at the whole hospital level. They identified safety design principles such as minimal movement for critical patients. They also identified failure modes such as shortage of nursing staff and broken elevators. The effect of failure modes for specific scenarios such as urgent moving of patients between specified departments was also identified. Further FMEA, both

process and design, was carried out at the level of designing individual rooms.

Software FMEA is perhaps the most active development area for FMEA at present. The introduction of software into a wide range of embedded systems is driving Software FMEA (SFMEA) development; however, the challenges are significant owing to the complex and varied failure modes that may be introduced within software. An ever increasing proportion of the functionality of many systems is provided by software, and even if the microprocessor hardware is considered highly reliable, reliability engineers need to understand how the software behaves in the presence of failures of the supporting hardware, such as sensors and actuators.

A related area of recent interest is the concept of performing an FMEA on the software system itself. Often the aim is to locate the generic failure modes that afflict software, such as premature termination, infinite looping, memory leaks, buffer overflow, etc. [11, 12]. Other approaches attempt to determine the functional failure modes of the system by considering the possible effects of specific input or module failures on the overall system functions. This process often involves tracing potentially faulty values through the software [13] and is extremely tedious if performed manually. The concept of analyzing software designs and code, specifically to determine the effects of hypothetical faults, requires a comprehensive and relatively abstract analysis together with sophisticated reporting capabilities. Tools to assist in performing FMEA on software are being developed, but have not reached maturity as yet.

References

- [1] US Department of Defence (1980). Procedures for performing a failure mode, effects and criticality analysis, MIL-STD-1629A (standard now withdrawn).

- [2] Automotive Industry Action Group (1995). *Potential Failure Mode and Effects Analysis (FMEA)*, QS-9000, 2nd Edition, Detroit.
- [3] SAE J1739 (2002). *Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) and Effects Analysis for Machinery (Machinery FMEA)*, SAE International.
- [4] SAE ARP5580 (2001). *Recommended Failure Modes and Effects Analysis (FMEA) Practices for Non-Automobile Applications*, SAE International.
- [5] Bowles, J.B. & Bonnell, R.D. (1993–1998). Failure mode effects and criticality analysis: what it is and how to use it, in *Topics in Reliability and Maintainability and Statistics, Annual Reliability and Maintainability Symposium*, p. 32 (revised each year) at <http://www.rams.org>.
- [6] Struss, P. & Price, C. (2003). Model-based systems in the automotive industry, *AI Magazine* **24**(4), 17–34. Special issue on Qualitative Reasoning.
- [7] Price, C. & Taylor, N. (2002). Automated multiple failure FMEA, *Reliability Engineering and System Safety Journal* **76**(1), 1–10.
- [8] Wilcox, R. (1999). *Risk – Informed Regulation of Marine Systems Using FMEA*, U.S. Coast Guard Marine Safety Center, p. 6, at <http://www.uscg.mil/hq/msc/fmea.pdf>.
- [9] DeRosier, J., Stalhandske, E., Bagian, J.P. & Nudell, T. (2002). Using health care failure mode and effect analysis™: the VA national center for patient safety's prospective risk analysis system, *Joint Commission Journal on Quality and Patient Safety* **28**(5), 248–267.
- [10] Reiling, J.G., Knutzen, B.L. & Stoecklein, M. (2003). FMEA: the cure for medical errors, *Quality Progress* **36**(8), 67–71.
- [11] Goddard, P.L. (2000). Software FMEA techniques, *Proceedings of the Annual Reliability and Maintainability Symposium*, Los Angeles, pp. 118–123.
- [12] Bowles, J.B. (2001). Failure modes and effects analysis for a small embedded control system, *Proceedings of the Annual Reliability and Maintainability Symposium*, Philadelphia, 1–6.
- [13] Ozarin, N. & Siracusa, M. (2002). A process for failure modes and effects analysis of computer software, *Proceedings of the Annual Reliability and Maintainability Symposium*, Seattle.

CHRIS PRICE

Fair Value of Insurance Liabilities

Insurance represents an influential and increasingly international industry that has no official international standards for financial reporting. Indeed, the great diversity in accounting practices makes it problematic for investors and regulators to assess and compare insurance companies worldwide and with other financial institutions. The last decade saw an ongoing debate on this issue (e.g., see [1, 2]) and significant changes taking place in the accounting world, culminating in the release of a standard on insurance contracts by the International Accounting Standards Board (IASB) (see **Solvency**). Key drivers of the changes have been the integration and globalization of capital markets and the strong demand by users of financial statements for greater transparency.

The new standards are characterized by a move toward market-consistent accounting, also called *fair value accounting*. As a consequence, accounting standards for the asset side of the balance sheet have witnessed a more readily and effective implementation of fair value principles (e.g., issue in 1993 of FAS115 [3], “Accounting for Certain Investments in Debt and Equity Securities”, in the United States), putting insurers in the awkward situation of marking only one side of the balance sheet to market, thus distorting equity and earnings of the company. Steps toward fair valuation of the liability side have been more difficult, since a true market for insurance liabilities does not exist, and insurance accounting practices have always been inspired to different principles.

From the point of view of financial economics, insurance contracts are just a particular example of dividend-paying securities and their valuation should certainly be market consistent. The fact that an insurance contract is illiquid and nonstandardized poses only a methodological problem, as valuation must be performed in an incomplete market setting, rather than in the usual (and more amenable) “frictionless” framework. This perspective, relevant for pricing of insurance liabilities, is emphasized to a limited extent here (see section titled “Fair Value Measurement” **Equity-Linked Life Insurance; Insurance Pricing/Nonlife**). We notice that financial economists

have been applying market valuation to insurance contracts for quite a long time, whereas actuaries and accountants have started doing so only recently. Early examples are represented by the seminal papers [4–9] (see [10] for additional references).

Here, we find it more useful to adopt an accountant’s perspective to describe how widely accepted accounting practices are being challenged by new valuation principles. We begin by providing a brief historical background on International Accounting Standards (IASs).

International Accounting Standards

The development of IASs began in 1973, when the accountancy bodies of different countries founded the International Accounting Standards Committee (IASC). After major restructuring, the committee was reformed and renamed IASB in 2001. The board adopted all of the standards prepared by the IASC and decided to produce new standards known as *International Financial Reporting Standards (IFRS)*. In 1997, the IASB set up a Steering Committee to work on the project *Insurance Contracts*. In December 1999, an Issues Paper [11] was released, attracting comments from financial institutions, supervisory authorities, and insurance companies worldwide. This feedback formed the basis for the release of a Draft Statement of Principles [12] in 2001, which later materialized into the issue of IFRS4 [13], “Insurance contracts”. The European Union having already endorsed the development of the standard in 2001, the European market provided one of the key test areas for its implementation. The standard is currently applied to a limited extent, as there are exclusions and exemptions that temporarily apply. The final phase of the IASB’s insurance project is expected to be launched within the next few years.

A number of standards are relevant to insurance business. This article is limited in scope and will mainly focuses on IFRS4, while providing only a few comments on IAS39 [14], “financial instruments”. The attention on IFRS4 is justified by two main reasons. First, the most problematic exercise the insurance industry faces is marking to market of liabilities. Second, even if the asset side of an insurer is less problematic, as it is likely to consist of more or less liquidly tradable instruments, the very same valuation framework described for liabilities can be applied to assets. Similar considerations apply in

principle to company pension liabilities, covered by IAS19 [15], “employee benefits”.

Principles of Fair Value Accounting

Fair value accounting relies on a single *recognition* and *measurement* approach for all forms of insurance contracts, regardless of the type of risk underwritten. In this section, we describe the following key principles: insurance risk and recognition of insurance contracts; definition of insurance assets and liabilities; fair value of insurance liabilities; measurement of assets and liability through a prospective, direct approach. The implementation of the measurement process is discussed in the section titled “Fair Value Measurement”.

Recognition of Insurance Risk

Under fair value accounting, the recognition of an insurance asset or liability is based on the substance of the economic transaction rather than on its legal form. An insurance contract is defined as a contract containing significant insurance risk, where *insurance risk* (see **Asset–Liability Management for Life Insurers; Correlated Risk; Basic Concepts of Insurance**) refers to uncertainty generated by risk sources other than financial risk factors (interest rates, security prices, commodity prices, credit rating, or credit indexes, etc.) and *significant* means the ability to generate changes in the net present value of the cashflows emerging from the contract. Annuities and pensions, whole life and term assurances are all examples of insurance contracts. However, by our definition, some of the contracts written by insurers may not qualify as insurance contracts: this is the case for some financial reinsurance agreements or for unit-linked products providing a death benefit not significantly higher than the account balance at the time of death. On the other hand, the divide between insurance contracts and financial instruments may be tiny: for example, while a contract requiring payments against credit downgrades of a particular entity would be accounted for as a derivative under IAS39 [14], a contract requiring payments in case of a debtor failing to pay when due would qualify as an insurance contract. The test for insurance risk is performed on a *contract-by-contract basis*, so that risk can be present even in those cases when there is minimal risk of significant changes in the present value

of the payments for a book of contracts as a whole. Also, a contract that qualifies as insurance contract at some point in time remains an insurance contract until all rights and obligations expire or are extinguished. This is relevant for endowment contracts, where sums at risk may vary considerably over the life term of the policy.

Definition of Insurance Assets and Liabilities

A cornerstone of fair value accounting is the recognition of income under an *asset and liability measurement* approach. This means that insurance assets and liabilities are identified through prespecified criteria, and then income and expenses are expressed in terms of variations in the measurement of those assets and liabilities. This is in stark contrast with the *deferral and matching* approach traditionally employed in insurance accounting, where revenue and expenses from contracts are recognized progressively over time as services are provided. An example of the latter approach is represented by the US standard FAS60 [16], released by the US-based Financial Accounting Standard Board. FAS60 prescribes the following measurement process for long-duration nonprofit life policies. Revenue is defined as investment income and premium earned, where premium is recognized in proportion to performance under the contract. Acquisition costs are deferred and amortized in proportion to premium revenue over the life of a block of contracts. This leads to recognition of a *deferred acquisition costs* asset. Liabilities are valued using assumptions reflecting best estimates plus a margin for adverse deviation. Those assumptions, once chosen, are “locked in” unless severe adverse experience develops in the future. “Unlocking” of assumptions is allowed only if a gross premium valuation shows that premium deficiency exists, thus triggering a “loss recognition” (see [1] for examples and additional discussion). Contrary to this approach, IFRS4 defines insurance assets and liabilities as assets and liabilities arising under an insurance contract. In particular, an insurer or policyholder should recognize an insurance liability (asset) when it has contractual obligations (rights) under an insurance contract that result in a liability (asset). As a consequence, items such as deferred acquisition costs remain excluded from recognition. The same applies to *catastrophe and equalization reserves*. The latter are provisions for

infrequent but severe catastrophic losses and for random fluctuations of claim expenses around expected values in some lines of business (e.g., hail and credit). They represent unearned premiums whose deferral provides for events expected to affect an entire business cycle, rather than a single reference contract, and are therefore excluded from recognition. In the approach just described, contractual rights and obligations arising from a book of contracts are seen as components of a single net asset or liability, rather than separate assets and liabilities. As a consequence, although the recognition of insurance assets and liabilities happens on a contract-by-contract basis, their measurement is performed at book level. The measurement must account only for the contracts in force at the reporting date (*closed book approach*), as an open book approach is inconsistent with a definition of assets and liabilities based on the existence, as a result of past events, of a resource or present obligation.

Fair Value of Insurance Assets and Liabilities

Fair value is defined as the amount for which an asset could be exchanged or a liability settled between knowledgeable, willing parties in an arm's length transaction. In particular, the fair value of a liability is the amount that the enterprise would have to pay a third party at the balance sheet date to take over the liability. The definition adopted is the *fair value in exchange*, as opposed to "fair value as an asset" (i.e., the amount at which others are willing to hold the liability as an asset) and "fair value in settlement with a creditor" (i.e., the amount that the insurer would have to pay to the creditor to extinguish the liability). In particular, the definition refers to a hypothetical transaction with a party other than the policyholder. Fair value is also an *exit value*, as opposed to an "entry value", which is the premium that the insurer would charge in current market conditions if she were to issue new contracts that created the same remaining contractual rights and obligations. If a liquid market existed for insurance liabilities, there would be no problem in observing exit values. In practice, secondary markets for the transfer of books of contracts are extremely thin, and exchange prices are often not publicly available. If such prices are available, they typically include an implicit margin for cross-selling opportunities, customer lists, and for future benefits arising from renewals that are not in the closed

book. The margin is often difficult to quantify, thus making problematic even the use of observed secondary market prices. The absence of observable exit prices makes it essential to estimate the fair value of a liability as if a hypothetical transaction were to take place. The favored approach is a *prospective* and *direct* method based on the expected present value of all future cash flows generated by a book of contracts. This is consistent with financial economic theory, as will be explained in the section titled "Fair Value Measurement". Before examining the actual implementation of the approach, it is useful to emphasize the key features of a prospective direct approach as opposed to alternative approaches widely adopted in insurance markets. We start with retrospective approaches, which focus on an accumulation of past transactions between policyholders and insurers. A life office, for example, could measure the insurance liability initially on the basis of the premiums received and defer acquisition costs as an asset or as a liability reduction. The expected profit margins would impact the measurement gradually over the life of the contract, in a way that depends on the technical basis assumed and on the accounting standard followed. It is typical of these approaches to use the policyholder account or the surrender value as a basis of measurement for contracts with an explicit surrender value or an explicit account balance (e.g., contracts described as universal life, unit-linked, variable, or indexed). On the other hand, a prospective approach focuses instead on the future cash inflows and outflows from a closed book of contracts and quantifies insurance liabilities through an estimate of their net present value. Such an estimate represents a provision for unexpired risk and can be more or less than the premium already paid by the policyholder, thus giving rise (unlike in the retrospective case) to a possible net profit on initial recognition. Moving on to direct and indirect methods, the latter discount all cash flows from the book of contracts and the assets backing the book, to derive a quantity that is then subtracted from the measurement of assets, giving the value of the book as a result. Direct methods discount, instead, the cashflows arising from the book of contracts and measure liabilities on that basis. Direct and indirect methods can produce the same results, provided a consistent set of assumptions is used, but direct methods usually provide greater transparency. See [17] for concrete examples in the context of

stochastic valuation models. An example of the indirect approach is the *embedded value method* (see **Securitization/Life; Risk Attitude**) (see [18]), which measures liabilities on a supervisory basis and then recognizes an asset, the embedded value, representing the amounts that will be released from the book of contracts as experience unfolds and liabilities are paid. In particular, the embedded value is given by the sum of the shareholders' net assets backing the book and the value of the business in force at the valuation date. The value of in-force business is the present value of future profits expected to emerge on the supervisory basis from policies already written (see [18] for details). The embedded value method is not market consistent, because it includes the present value of estimated future cash flows from investments representing the insurance liability. Specifically, it attributes an amount different from the fair value to assets held, because it discounts the cash flows from those investments at a rate reflecting not only the risks associated with the business, but also the cost of capital locked in by capital requirements. An evolution of the approach, known as *market-consistent embedded value method*, resolves these issues in order to achieve consistency with fair valuation.

Fair Value Measurement

The measurement of fair value relies on an expected present value approach that incorporates adjustments for uncertainty reflecting the risk preferences of market participants in a hypothetical secondary market transaction. The key ingredients to the recipe are the timing and amount of the cashflows generated by a book of contracts, and the margins for risk.

Amount and Timing of Cashflows

IFRS4 favors an explicit approach to adjustments for risk, i.e., an approach based on adding market-consistent margins on top of realistic estimates of liabilities. We therefore focus first on a realistic analysis of insurance cashflows, leaving considerations on risk adjustments for the next section. The fair valuation exercise encompasses all future pre-income tax cash flows arising from the contractual rights and obligations associated with a closed book of insurance contracts. These include future benefits and premium receipts under existing contracts,

claim handling expenses, policy administration and maintenance costs, acquisition costs, etc. Consistent with our definition of assets and liabilities, the following must be excluded among, others: cashflows arising from future business, payments to and from reinsurers (an appropriate insurance asset should, instead, be recognized), regulatory requirements, and cost of capital. Two main classes of assumptions are employed for the measurement of insurance cashflows: market assumptions, such as interest rates, for which transactions in the capital markets provide easier estimation; nonmarket assumptions, such as lapse and mortality rates or claim severities. Market assumptions should be consistent with current market prices and other market-implied data, while nonmarket assumptions must be consistent with industry assumptions. Capital markets are of course the main source of market assumptions. If a suitable replicating portfolio (see **Insurance Pricing/Nonlife; Solvency**) (i.e., an asset portfolio closely matching the insurer's obligations) can be constructed, then the market value of the portfolio would immediately provide a close proxy for the fair value of an insurance liability. Sources of nonmarket assumptions are data available to individual insurers (claims already reported by policyholders, historical data on insurers' experience), industry data, and recent exit prices observed in secondary markets. Two other markets that can provide useful information are primary markets and reinsurance markets. Primary markets for the issuance of direct insurance contracts provide information about the current pricing of risk, but can only partially reflect prices paid in wholesale markets. Reinsurance markets may provide an indication of prices that would prevail in secondary markets, but have considerable limitations: they are not generally true exit prices, as most reinsurance contracts do not extinguish the cedant's contractual obligations under the direct contract, and they may include a margin for intangible items.

Adjustments for Risk and Uncertainty

As fair value measurement is performed at book level, the definition of *book of contracts* (see **Alternative Risk Transfer**) can have significant impact on the degree of risk attached to a liability, because of pooling (several homogeneous risks pooled together) and diversification (several uncorrelated risks pooled together) effects. IFRS4 states that measurement

of insurance contracts should focus on books of contracts subject to substantially the same risks, and should reflect all benefits of diversification and correlation within those books. As a consequence, while the fair value of a book of contracts is likely to be lower than what would be obtained by considering individual contracts before aggregation, at the same time a limit is imposed to risk offsetting between different exposures. As an example, term assurances and annuities cannot be included in the same book, as opposed to current practice in some countries.

According to financial economics, there are three main methods, all theoretically correct, for implementing the expected present value approach to value a set of risky future cash flows. The first approach is the traditional discounted cash flow approach: it employs physical (real-world) probabilities to determine the expected payoffs, and discounts them at the risk-free rate plus a risk premium reflecting the riskiness of the cashflows as perceived from market participants. A second approach is risk-neutral valuation (see **Equity-Linked Life Insurance; Options and Guarantees in Life Insurance; Premium Calculation and Insurance Pricing; Volatility Smile**), where risk aversion is introduced in the probabilities and discounting is performed at the risk-free rate. Risk-neutral probabilities replace physical ones to reflect the weight that risk-averse investors would place on uncertain outcomes in a world where the expected return on any security is the risk-free rate. Finally, a third approach adjusts the cash flows directly to their certainty equivalent levels, and then discounts their realistic expected values at the risk-free rate. Here, risk aversion is introduced by specifying a utility function acting directly on the cashflows. Although the insurance industry seems to prefer an approach based on explicit adjustments (so-called market value margins) to realistic estimates of the liabilities, any of the previous methodologies should produce the same results if applied consistently. In particular, the probability adjustments introduced in the second approach would translate into proper risk-adjusted discount rates in the first approach or cashflow adjustments in the third approach. We note that many insurance contracts present complex payoff structures that require an integrated treatment of financial and insurance risks; in this situation, risk-neutral valuation is the method most fruitfully employed (see **Pricing of Life**

Insurance Liabilities; Equity-Linked Life Insurance; Options and Guarantees in Life Insurance).

Conclusion

The debate on implementation of fair value accounting is a lively area of research. The IASB's project agenda is publicly available and gives an idea of the number of issues still being debated. We have not touched upon several interesting topics, such as fair valuation of life contracts with discretionary participating features, which are subject of considerable attention by accountants, actuaries, regulators, and academics. We refer the reader to the IASB's website or to the volumes [19] and [20] for additional information on these and other matters.

References

- [1] Vanderhoof, I.T. & Altman, E.I. (eds) (1998). *The Fair Value of Insurance Liabilities*, Kluwer Academic Publishers.
- [2] Vanderhoof, I.T. & Altman, E.I. (eds) (1998). *The Fair Value of Insurance Business*, Kluwer Academic Publishers.
- [3] Financial Accounting Standards Board (1993). *Statement n. 115: Accounting for Certain Investments in Debt and Equity Securities*.
- [4] Brennan, M.J. & Schwartz, E.S. (1976). The pricing of equity-linked life insurance policies with an asset value guarantee, *Journal of Financial Economics* **3**, 195–213.
- [5] Boyle, P.P. & Schwartz, E.S. (1977). Equilibrium prices of guarantees under equity-linked contracts, *The Journal of Risk and Insurance* **44**(4), 639–660.
- [6] Smith, M.L. (1982). The life insurance policy as an options package, *The Journal of Risk and Insurance* **49**(4), 583–601.
- [7] Delbaen, F. (1986). Equity linked policies, *Bulletin de l'Association Royale des Actuaire Belges* **44**, 33–52.
- [8] Bacinello, A.R. & Ortu, F. (1993). Pricing equity-linked life insurance with endogenous minimum guarantees, *Insurance: Mathematics and Economics* **12**(3), 245–258.
- [9] Briys, E. & De Varenne, F. (1994). Life insurance in a contingent claim framework: pricing and regulatory implications, *The Geneva Papers on Risk and Insurance Theory* **19**, 53–72.
- [10] Jorgensen, P.L. (2004). On accounting standards and fair valuation of life insurance and pension liabilities, *Scandinavian Actuarial Journal* **5**, 372–394.
- [11] International Accounting Standards Committee (1999). *Issues Paper on Insurance Contracts*.
- [12] International Accounting Standards Board (2001). *Draft Statement of Principles*.

- [13] International Accounting Standards Board (2001). *International Financial and Regulatory Standard n. 4: Insurance Contracts*.
- [14] International Accounting Standards Board (2003). *International Accounting Standard n. 39: Financial Instruments-Recognition and Measurement*.
- [15] International Accounting Standards Board (2001). *International Accounting Standard n. 19: Employee Benefits*.
- [16] Financial Accounting Standards Board (1982). *Statement n. 60: Accounting and Reporting by Insurance Enterprises*.
- [17] Biffis, E. (2005). Affine processes for dynamic mortality and actuarial valuations, *Insurance: Mathematics and Economics* **37**(3), 443–468.
- [18] Fine, A.E.M. & Geddes, J.A. (1988). *Recognition of Life Insurance Profits: The Embedded Value Approach*, Institute of Actuaries, London.
- [19] Briys, E. & De Varenne, F. (2001). *Insurance: From Underwriting to Derivatives-Asset Liability Management in Insurance Companies*, John Wiley & Sons.
- [20] Moller, T. & Steffensen, M. (2007). *Market-valuation Methods in Life and Pension Insurance*, Cambridge University Press.

Further Reading

Abbink, M. & Saker, M. (2002). *Getting to Grips with Fair Value*, Staple Inn Actuarial Society, London.

ENRICO BIFFIS AND PIETRO MILLOSOVICH

Fate and Transport Models

Fate and transport models are tools used to describe how chemicals spread in the environment (*see **Statistics for Environmental Toxicity; Environmental Risk Assessment of Water Pollution***). They consist of a mathematical representation of the processes a chemical undergoes when it is released, and particularly

1. how a chemical is partitioned between different physical phases;
2. which transfer mechanisms drive it in the movement between different environmental media (air, water, soil, sediments, vegetation, biological material, etc.);
3. how a chemical reacts/is degraded within each medium;
4. how a chemical is transported within each medium.

The first point is normally dealt with standard thermodynamics equations of equilibrium partitioning. The second point involves the description of transfer mechanisms, such as absorption, volatilization, and deposition through settling particles (e.g., atmospheric particulate, suspended sediments in water).

The third point requires a representation of chemical reactions. In addition, sometimes it may be necessary to study different chemical species simultaneously. This implies the use of multispecies fate and transport models that describe not just one chemical, but also its reaction products.

The fourth point relates to the processes of diffusion, dispersion, and advection, which are in essence fluid-dynamics processes. These processes occur mainly within air and water, or the air and water phases in soil.

Phase Partitioning

In general, the total mass of chemical in a specific environmental medium is subdivided (partitioned) between the n different phases of the medium (usually including solid, liquid, gas, biological material) as

$$M = \sum_{i=1}^n C_i V_i = V \sum_{i=1}^n C_i \phi_i \quad (1)$$

where C_i is the concentration of chemical in phase i ($[M][L]^{-3}$) and V_i is the volume occupied by each phase in the medium ($[L]^3$). $V = \sum_{i=1}^n V_i$ is the total volume of the medium, and ϕ_i is the fraction of volume occupied by phase i [—]. Usually, concentration in solid/biological material is expressed in terms of mass of chemical per mass of material (massic concentration C_{Mi}). In this case, $C_{Mi} = C_i / \rho_i$, where ρ_i is the mass of material in the unit volume of medium ($[M][L]^{-3}$).

Usually, partitioning is considered in environmental modeling among the following phases for each medium:

1. air partitioning (assuming the fraction of volume V occupied by aerosol is negligible) between air and aerosol:

$$M = V(C_g + C_{Ms}TSP) \quad (2)$$

2. surface water partitioning (assuming the fraction of volume V occupied by suspended sediments and biological material is negligible) among water, suspended sediments, and biological material (e.g., fish, algae...):

$$M = V(C_l + C_{Ms}TSP + C_{Mb}BIO) \quad (3)$$

3. soil/aquifers – partitioning between soil air, soil water, and the solid matrix:

$$M = V(\theta C_l + (1 - \phi)C_{Ms}\rho_s + (\phi - \theta)C_g) \quad (4)$$

In the above equations, the symbols have the meaning listed hereafter:

C_g : concentration in gas phase $[M][L]^{-3}$

C_l : concentration in liquid phase $[M][L]^{-3}$

C_{Ms} : massic concentration in solid phase $[M][M]^{-1}$

C_{Mb} : massic concentration in biological material $[M][M]^{-1}$

TSP: concentration of particles in air (aerosol) or surface water (suspended sediments) $[M][L]^{-3}$

BIO: concentration of biological material $[M][L]^{-3}$

ϕ : volume fraction of soil which is not solid, or soil porosity [—]

θ : volume fraction of soil which is occupied by water, or soil moisture content [—]

ρ_s : soil solid density $[M][L]^{-3}$.

For a more thorough discussion, the reader is referred to [1].

Often, we assume that C_g , C_{Ms} , C_{Mb} , and C_l are all related among each other independent on time, so that $C_g = f(C_l)$, $C_{Ms} = f'(C_l)$ etc., where f , f' , $\dots$, are appropriate functions. This assumption is called the *equilibrium partitioning* assumption. The most simple function that can be adopted for concentrations in two phases a, b, is in the form $C_a = kC_b$ (linear equilibrium partitioning) where k is a constant called the *partitioning coefficient*. The partitioning coefficient can be dimensionless if concentration in both phases is in the same dimensions; the solid–liquid partitioning is often expressed in $[L]^3[M]^{-1}$ as a ratio of massic and volumetric concentrations. Partitioning functions can be nonlinear; particularly common are the Langmuir and Freundlich equations [2]. The solid–liquid partitioning coefficient is also called *distribution coefficient* K_d , and is normally expressed as a product of the organic carbon content of the solid, and the octanol–water equilibrium partitioning coefficient K_{ow} , a physicochemical property of substances that can be estimated experimentally. The gas–liquid partitioning coefficient is also called (nondimensional) *Henry's constant* K_{aw} and represents a physicochemical property of substances as well. K_{ow} and K_{aw} can be strongly dependent on temperature.

The equilibrium assumption implies that concentration in all phases is determined once concentration in a single phase is known.

Transfer to Other Media

Transfer to other media can occur in the form of diffusive or advective process. The first category includes in particular absorption of gases at, and volatilization from, one surface, which can be liquid or solid.

Advective processes include transport of substances through the removal of particles, or as dissolved in moving fluids.

Such processes can be described using the following general mathematical representation:

$$Q_{a,b} = A \sum_{i=1}^n C_i u_{i,(a,b)} \quad (5)$$

where $Q_{a,b}$ is the total flux of a substance from medium a to medium b, C_i is the concentration in

each of the n phases of medium a $[M][L]^{-3}$, and $u_{i,(a,b)}$ is the transfer rate in that phase from medium a to medium b $[L][T]^{-1}$. A represents the surface area of the interface between media a and b $[L]^2$.

In the case of air, the phases involved are (a) gas, (b) aerosol particles, and (c) raindrops. Snowflakes can be also included; they are normally described in an intermediate way between raindrops and aerosol particles. For simplicity, we neglect this additional phase here, with no loss of generality in the discussion. Concentration in the gas phase, in raindrops, and on aerosol particles are related to each other under equilibrium assumptions. Media a and b are identified as air, and either soil, water surfaces or canopies. The transfer rates represent the following:

- $u_{1(a,\text{surf})}$: gas absorption velocity at the air–surface interface;
- $u_{2(a,\text{surf})}$: aerosol (dry) deposition velocity;
- $u_{3(a,\text{surf})}$: wet deposition velocity.

Wet deposition is the process by which raindrops deposit to surfaces both gaseous substances by absorption, and particles (and consequently particle-attached chemicals) by scavenging. There are several ways to parameterize the above listed transfer rates; usually, gas absorption is given as a function of wind speed over the surface and the humidity of the surface itself, dry deposition is a function of turbulence and atmospheric stability, and wet deposition is a function of rainfall intensity and the duration of interstorm periods. Also, the type of land use strongly affects the transfer through dry deposition and gas absorption. The reader is referred to the literature for further details [3, 4].

In the case of surface water, the phases involved are (a) suspended solids and (b) the water phase. The transfer rates represent the following:

- $u_{1(w,\text{sed})}$: deposition velocity of particles due to gravitational and turbulent settling, toward the water body bottom sediments;
- $u_{2(w,a)}$: velocity of volatilization to the atmosphere.

Deposition velocity of particles can be estimated using Stokes' law under laminar conditions and with semiempirical formulas under turbulent conditions [5]. The velocity of volatilization is normally assumed equal to the velocity of gas absorption and

the net exchange between air and water is driven by the relative concentrations [3].

In the case of soil, the phases involved are (a) soil solids, (b) air, and (c) water in the soil pores. The transfer rates represent the following:

- $u_{1(s, \text{sed})}$: volumetric erosion rate;
- $u_{2(s, w)}$: runoff rate + rate of percolation to the water table;
- $u_{3(s, a)}$: velocity of volatilization to the atmosphere.

The erosion rate is the amount of soil solids lost per unit time and surface. The runoff/percolation rates are the amount of water leaving the soil per unit time and surface. The velocity of volatilization is evaluated as in the case of water; normally it is assumed that volatilization occurs in parallel between soil air and air, and between soil water and air. The diffusion velocities inside the soil pores are assumed to be lower than the corresponding ones in free air or water [6].

Reactions/Degradation

Reaction kinetics can be extremely complex, but up to now it has been common practice to use very simple kinetic models, such as, in most cases, linear models. In some environmental fate and transport models, other types of kinetics are also used [7]. In general, a simplified linear representation of nonlinear, or otherwise complex phenomena is acceptable when the variation of the quantities involved is limited around certain typical values. The use of linear models is reasonable insofar as the errors introduced by this simplification are small compared with the other sources of uncertainty or errors. This is often the case in environmental modeling, which involves many parameters difficult to determine, and even many processes difficult to describe in equations. According to the linear model, the flux of substance that is degraded $[M][T]^{-1}$ is

$$Q_{\text{deg}} = V \sum_{i=1}^n k_i \phi_i C_i \quad (6)$$

where k_i is the degradation rate of the substance in phase i $[T]^{-1}$, C_i its concentration $[M][L]^{-3}$, and V the bulk volume of the medium $[L]^3$. The degradation rates are either determined experimentally, or estimated from similar properties such as reaction

rate with a known standard. Further details can be found in [3]. Usually it is recognized that the degradation rates depend strongly on temperature [3]. The reaction rates of substances in the environment are a highly uncertain parameter in most fate and transport models; often an “overall” degradation rate is preferred to phase-specific ones as in equation (6), in that it is considered as a model parameter to be calibrated.

Transport within an Environmental Medium without Reactions and Generalization to Simple Multiphase Situations with Reactions

The transport of a conserved substance within a medium (i.e., a substance not subject to degradation nor to transfer to other media) is due to two types of phenomena: advection and diffusion (called *dispersion* whenever it occurs in turbulent media). Diffusion (dispersion) can be described as a process by which a chemical tends to move from regions at higher concentration, to regions at lower concentration (Fick’s law) [8]. Accordingly, the components of the diffusion (dispersion) flux vector of a substance at point (x, y, z) $[M][T]^{-1}$ is given by

$$\overline{Q}_{\text{diff}} = \begin{bmatrix} Q_{\text{diff}, x} \\ Q_{\text{diff}, y} \\ Q_{\text{diff}, z} \end{bmatrix} = \overline{\overline{D}} \times \text{grad } C \quad (7)$$

where $\overline{\overline{D}}$ is called *diffusion (dispersion) coefficients matrix* ($[L]^2[T]^{-1}$). An element (I, j) of this matrix represents the diffusion (dispersion) flux occurring in direction i for a unit concentration gradient in direction j . Often, it is assumed to be a diagonal matrix.

Advection is a process by which a chemical is transported along the flow pathways of a motion field $v(x, y, z)$, i.e., a vector of velocities along directions (x, y, z) defined for each point in space. The advection flux vector of a substance at point (x, y, z) $[M][T]^{-1}$ is given by

$$\overline{Q}_{\text{adv}} = C \overline{v} \quad (8)$$

where v is the vector of velocity of the fluid. The mass balance equation of a conservative substance for a given reference volume of an environmental medium (e.g., air, soil, water) can be written as:

$$\frac{\partial C}{\partial t} = \text{Emission} - \text{dispersion} - \text{advection} \quad (9)$$

where the dispersion and advection terms are mathematically derived as the divergence of Q_{diff} and Q_{adv} through appropriate analytical steps [9], and emission term represents the input of substance from outside and is called sometimes *source-sink term*. In this way we obtain the advection–dispersion equation (ADE):

$$\frac{\partial C}{\partial t} = \text{Emission} - v \times \text{grad } C - \text{div} (\bar{D} \times \text{grad } C) \quad (10)$$

In the above notation, $\times$ represents the matrix product of vectors, div the divergence operator and grad the gradient operator. The ADE is a very general tool that describes the mass balance of a single substance in a single medium and phase. It can be solved analytically or numerically once initial and boundary conditions are set. Boundary conditions represent, in particular, the specific conditions of mass exchange between a medium, where the ADE is solved, and its external interfaces. Analytical solutions exist for very simple cases such as parallel flow, zero-velocity (dispersion only), and simple boundary conditions [10]. Methods to solve this equation numerically include finite elements, finite differences and the method of characteristics (MOC). Some methods of solution of the ADE require to represent it in Lagrangian form, i.e., through the substantial (Lagrangian) derivative $\frac{d\bullet}{dt} = \frac{\partial\bullet}{\partial t} + v \times \text{grad}(\bullet)$:

$$\frac{dC}{dt} = \text{Emission} - \text{div}(D \times \text{grad } C) \quad (11)$$

Additional details are provided e.g., in [11].

It is easy to generalize the ADE by including reactions in the case of a single phase. In such cases, the mass balance equation is:

$$\begin{aligned} \frac{\partial C}{\partial t} = & \text{Emission} - \text{dispersion} - \text{advection} \\ & - \text{degradation} \end{aligned} \quad (12)$$

which becomes, using a linear model for degradation,

$$\begin{aligned} \frac{\partial C}{\partial t} = & \text{Emission} - v \times \text{grad } C - \text{div} (D \text{ grad } C) \\ & - kC \end{aligned} \quad (13)$$

where k is the degradation rate. In this way, the chemical that is degraded is considered as a net loss from the system, and no description is provided about the evolution of the reaction products in the environment. The above equation is referred to as the advection–dispersion–reaction equation (ADRE).

If one assumes equilibrium between the different phases of an environmental medium, it is possible to write the ADRE to include multiple phases in a medium. Assume C is the concentration in the medium a of interest, in phase X where advection and dispersion take place. For instance, this could be the dissolved phase in water, or the gas phase in the atmosphere. Assume for simplicity linear equilibrium between phases. Concentration in each of the other phases is expressed as $C_i = \zeta_i C$, where ζ_i is the partitioning coefficient between phase i and phase X . It is clear that $\zeta_X = 1$, and the overall degradation flux is

$$Q_{\text{deg}} = V \sum_{i=1}^n k_i C_i = VC \sum_{i=1}^n k_i \zeta_i \quad (14)$$

Using the same reasoning, the removal flux from medium a toward another medium b is

$$Q_{a,b} = A \sum_{i=1}^n C_i u_{i,(a,b)} = AC \sum_{i=1}^n \zeta_i u_{i,(a,b)} \quad (15)$$

Also, the total concentration of chemical in the medium is the sum of the concentration in all phases, weighted with the volume fraction of each phase:

$$C_{\text{tot}} = \sum_{i=1}^n C_i \phi_i = C \sum_{i=1}^n \zeta_i \phi_i \quad (16)$$

The ADRE is written then with reference to the total concentration, but advection and dispersion only occur in phase X .

Under these assumptions, the ADRE can be further generalized in the following form:

$$\begin{aligned} \sum_{i=1}^n \zeta_i \phi_i \frac{\partial C}{\partial t} = & \text{Emission} - v \times \phi_X \text{ grad } C \\ & - \text{div} (\phi_X D \text{ grad } C) \\ & - C \left(\sum_{i=1}^n \phi_i k_i \zeta_i + \sum_{b=1}^m \sum_{i=1}^n \zeta_i u_{i,(a,b)} \right) \end{aligned} \quad (17)$$

where C is the concentration of substance in medium a in phase X , and m is the number of other media where the substance can be transferred from medium a . the above formulation of the ADRE is rather general and allows describing a very broad range of fate and transport problems for substances in the environment. Writing an equation of the type (17) for each medium involved in the transport process

provides a system of equations which allow predicting all concentrations simultaneously. In the following, we will provide an explicit formulation of the ADRE for specific individual media.

Fate and Transport in Soil and Aquifers

In soils and aquifers, generally one is interested in the dissolved phase of chemicals in groundwater; however, for specific problems the explicit description of air flow in soils can be required, e.g., for the design of soil remediation from volatile contaminants. We refer here to the water-dissolved concentration C for reference. Concentration in soil solids is expressed as $C_{Ms} = K_d C = f_{oc} K_{ow} C$ where f_{oc} is the fraction of organic carbon in soils; accordingly, $C_s = \rho_s C_{Ms}$. Also, concentration in soil air is $C_a = K_{aw} C$. In this case, the ADRE is written as:

$$\begin{aligned} & (\theta + (\phi - \theta)K_{aw} + (1 - \phi)\rho_s K_d) \frac{\partial C}{\partial t} \\ &= \text{Emission} - v \times \theta \text{ grad } C - \text{div } (\theta D \text{ grad } C) \\ & \quad - C(\theta k_{sw} + (\phi - \theta)K_{aw}k_{sa} + (1 - \phi)\rho_s K_d k_s \\ & \quad + u_{1,(s, \text{sed})}\rho_s K_d + u_{3,(s,a)}K_{aw}) \end{aligned} \quad (18)$$

where k_s is the degradation rate of the substance adsorbed to the soil solids, k_{sa} is the one for the substance in soil air and k_{sw} for soil water. It is worth noting that the transfer from soil/aquifer to other water bodies (e.g., deeper aquifers, surface water), represented by the term $u_{2,(s,w)}$ introduced above, is incorporated in the ADRE implicitly in the advective term, and controlled by the boundary conditions that allow solving the equation.

A special case is when soil is saturated in water ($\phi = \theta$). Such condition occurs normally in deeper layers of the soil/aquifer, so that erosion can be neglected. In this case, the ADRE is simplified to

$$\begin{aligned} & (\phi + (1 - \phi)\rho_s K_d) \frac{\partial C}{\partial t} = \text{Emission} - v \times \phi \text{ grad } C \\ & \quad - \text{div } (\phi D \text{ grad } C) \\ & \quad - C(\phi k_{sw} + (1 - \phi)\rho_s K_d k_s \\ & \quad + u_{3,(s,a)}K_{aw}) \end{aligned} \quad (19)$$

and the term $(\phi + (1 - \phi)\rho_s K_d)$, called *retardation factor*, represents the ratio of the speed of the contaminant on the speed of soil water [12]. The dispersion coefficients D are estimated on the basis

of criteria for which the reader is referred to [13]. The velocity vector is evaluated in soil and groundwater according to the Darcy law [12]:

$$v = \frac{\overline{\overline{K}} \times \text{grad } h}{\phi} \quad (20)$$

where $\overline{\overline{K}}$ is the hydraulic conductivity matrix [L] [T]⁻¹ and h is the total hydraulic head [L]. In general, K depends on soil water saturation θ and is represented through theoretical relations $K(\theta)$ [14].

In complex cases, the values of v are provided by appropriate hydrodynamic models solving the water mass balance equation [12, 15–17]. At present, many computer models exist that solve the ADRE for soil and groundwater. Among them, it's worth quoting the MT3D model for saturated soil in three dimensions, produced by the United States Environmental Protection Agency (USEPA) [18]; VS2DT for variably saturated soil in two dimensions, produced by the United States Geological Survey (USGS) [19], and SUTRA for variably saturated soil in two and three dimensions, produced by the USGS [16].

Fate and Transport in Surface Water

Following the same reasoning as in the case of soil and aquifers, but considering the only two phases present in surface water (suspended solids and the water solution), the ADRE can be written as

$$\begin{aligned} & (1 + TSP K_d) \frac{\partial C}{\partial t} = \text{Emission} - v \times \text{grad } C \\ & \quad - \text{div } (D \text{ grad } C) \\ & \quad - C(k_w + TSP K_d k_{ss}) \\ & \quad - (u_{1,(w,s)} TSP K_d \\ & \quad + u_{2,(w,a)} K_{aw}) C \end{aligned} \quad (21)$$

We refer here again to the water-dissolved concentration C for reference.

In rivers, in most practical cases flow can be regarded as one dimensional; often dispersive fluxes can be neglected ($D = 0$) as they are much smaller than advective fluxes. Moreover, if emissions are reasonably constant over a period, steady state is often reached, i.e., the time derivative in the ADRE vanishes. If these conditions are met, one obtains the so-called plug flow equation [20] which is an ordinary

differential equation of the first order. This can be written as

$$v \frac{dC}{dx} = \text{Emission} - C(k_w + TSP K_d k_{ss}) - (u_{1(w,s)} TSP K_d + u_{2(w,a)} K_{aw}) C \quad (22)$$

This equation has an analytical solution if the *emission* term can be represented in a mathematically tractable form; additional discussion can be found in [20]. A special case of interest is the one of a continuous point emission in a stream with constant velocity, settling velocity, and suspended solid concentration, with no other emission along the stream. In such case, if C_0 is the concentration in the stream at the emission point, the spatial distribution of concentration along the stream is

$$C = C_0 \exp \left(-\frac{x}{v} \left(k_w + TSP K_d k_{ss} + u_{1(w,s)} TSP K_d + u_{2(w,a)} K_{aw} \right) \right) \quad (23)$$

This case is of particular interest as it represents well the fate of a chemical emitted from e.g., an industrial outflow or a waste water treatment plant, and it allows a screening level assessment of the expected pollution in the receiving water.

Dispersion coefficients in surface water can be estimated either by empirical methods [21] or using more theoretical fluid-dynamics methods based e.g., on the deformation velocity of the fluid in motion [22]. Advection velocity is normally computed using semiempirical hydraulic functions such as Manning's formula:

$$v = \frac{1}{n} R^{\frac{2}{3}} \sqrt{J} \quad (24)$$

where R is the hydraulic radius (ratio between flow cross section area and cross section wetted perimeter [L]), n is an empirical roughness coefficient $[T][L]^{1/3}$, and J is the gradient of the total hydraulic head. For complex flow conditions, such as irregular geometry of the water body, hydraulic structures, the presence of wind at the water surface, tides or water waves at the boundary of the water body, the computation of velocities need to be described using an appropriate hydraulic model [22–24].

Some models commonly used to solve the ADRE for surface water are QUAL2K [25] for one-dimensional flow, SWAT [26] for catchment scale processes, WASP [27] for two dimensional flow; a

rather general model used for coastal waters and complex two dimensional problems is COHERENS [24].

Fate and Transport in the Atmosphere

In the case of the atmosphere (*see Air Pollution Risk*), the chemical is partitioned between a gas phase and a particle phase (aerosol or particulate matter, with concentration $TSP [M][L]^{-3}$). We refer here to the gas-phase concentration C for reference.

We call $K_{pm,a}$ the linear–equilibrium–partitioning coefficient. Usually this coefficient is expressed as a function of the substance properties K_{aw} and K_{ow} (e.g., [1]). Also, raindrops (and snow flakes) act by absorbing chemicals in gas phase and scavenging aerosol. The concentration of chemical in raindrops coming from the gas phase during rainfall is usually evaluated assuming equilibrium between rainwater and the gas phase itself, and is therefore $C_{raindrop, gas} = C K_{aw}^{-1}$. The concentration in raindrops coming from the aerosol during rainfall is evaluated through an empirical “washout factor” W_{pm} (nondimensional), so that $C_{raindrop, pm} = C W_{pm} TSP K_{pm,a}$. Under these assumptions, the ADRE becomes:

$$\begin{aligned} (1 + TSP K_{pm,a}) \frac{\partial C}{\partial t} &= \text{Emission} - v \times \text{grad } C - \text{div } (D \text{ grad } C) \\ &- \left(k_a + TSP K_d k_{pm} + u_{1(a,surf)} + u_{2(a,surf)} TSP K_{pm,a} \right. \\ &\left. + \left(\frac{1}{K_{aw}} + W_{pm} TSP K_{pm,a} \right) u_{3(a,surf)} \right) C \end{aligned} \quad (25)$$

where k_a is the degradation rate in the gas phase, and k_{pm} is the one in the aerosol phase. This equation is usually solved in three dimensions with numerical methods from appropriate initial and boundary conditions. Advection velocity represents here wind speed, and is usually provided by atmospheric circulation models. Dispersion coefficients are usually estimated theoretically from air fluid–dynamic considerations [21]. Although recent developments in computing power support the use of complex three-dimensional models also for relatively simple, local scale problems, a practical solution to the ADRE for local applications is still the one provided by the so-called Gaussian models. This class of models grounds on an analytical, steady-state solution of the ADE

(i.e., neglecting the reaction and removal terms) that corresponds to constant emission, with emission rate Q [M][T]⁻¹, in a wind field of constant direction and intensity. If x is the abscissa originating from the emission point and directed along the wind, y its perpendicular in the horizontal, and z in the vertical plane, the analytical solution is:

$$C(x, y, z) = \frac{Q}{4\pi x \sqrt{D_y D_z}} \exp\left(-\frac{vy^2}{4D_y}\right) \times \left(\exp\left(-\frac{v(z+z_s)^2}{4D_z}\right) + \exp\left(-\frac{v(z-z_s)^2}{4D_z}\right) \right) \quad (26)$$

where z_s represents the elevation of the emission point, v the wind speed, D_y and D_z the dispersion coefficients in y and z directions. This base solution describes a spatial distribution of concentrations in the form of a plume, and can be generalized to include many additional effects, through empirical or semiempirical corrections. Commonly, effects included in Gaussian models are the variability of winds (through statistical treatment of the joint-frequency function of wind direction and intensity, or wind rose, for a specific site); reactions, atmospheric deposition; building downwash (modified turbulence due to aerodynamic obstacles); effects of complex terrain such as mountains. A model commonly used in the past to treat these features is ISC3 from USEPA [28]. ISC3 also deals with areal and linear emissions, while a specific Gaussian model used for linear emissions from road traffic is CALINE from the State of California EPA [29]. Also, modifications have been proposed to cope with puff-type emissions [30] and for periods of no-wind (dispersion only) [31]. Another similar model of wide application is ADMS [32].

When using Gaussian plume models, dispersion coefficients are often represented as a simple function of x and atmospheric stability conditions [33].

Models of higher complexity, but widely used at present, include CAMx [34] and AERMOD [35]. These models are suitable for studies at regional to local scale. For wider domains such as a continent, a hemisphere or the globe, models such as TM5 [36] and MM5 [37] are better suited as they are directly linkable to general circulation models (GCMs) of the atmosphere and allow simulating transport in a fully three-dimensional atmospheric domain.

Box Models, Multimedia, and Multispecies Fate and Transport

The ADE and ADRE described above allow a detailed simulation in space and time of the fate and transport of chemicals in the environment. However, solving these equations can be demanding both in terms of computational burden, and for what concerns input data acquisition. This becomes particularly problematic when simulating at the same time more environmental media, and/or more chemical species. In such circumstances, it is common practice to describe the environment in the form of a set of well-mixed (uniform concentration) compartments, often referred to as *continuous stirred tank reactors* (CSTR), or simply *boxes*, hence the terming *box models* when using this approach. Each CSTR represents one specific phase in an environmental medium. The mass balance for a chemical in a CSTR of volume V is given by the following equation:

$$V \frac{dC}{dt} = E + \sum_{i=1}^n Q_{in}^i C_{in}^i - Q_{out} C - KC \quad (27)$$

where E is a direct mass flow rate of chemical into the CSTR, K is a term that lumps together all degradation and removal processes from the CSTR, Q_{in}^i is each of the n volumetric flow rates of the phase entering the CSTR from outside, C_{in}^i its concentration of chemical, Q_{out} is the volumetric flow rate exiting the CSTR, and C is the concentration of chemical in the CSTR. The above equation is an ordinary differential one and has an analytical solution for each period for which the terms $(E + \sum_{i=1}^n Q_{in}^i C_{in}^i)$, Q_{out} and K can be regarded as constant. Let us consider now two boxes exchanging the chemical. Let us use the subscripts 1 and 2 for the respective quantities. It is then possible to write a system of two equations as:

$$\begin{aligned} V_1 \frac{dC_1}{dt} &= E_1 + Q_{21} C_2 - Q_{out,1} C_1 - K_1 C_1 \\ V_2 \frac{dC_2}{dt} &= E_2 + Q_{12} C_1 - Q_{out,2} C_2 - K_2 C_2 \end{aligned} \quad (28)$$

where Q_{21} represents the volumetric flow rate from box 2 to box 1, Q_{12} *vice versa*. The system can be generalized to m boxes and written in matrix form as:

$$\frac{d\bar{C}}{dt} = \bar{E} - \bar{K} \bar{C} \quad (29)$$

where $\bar{C}$ is the vector ($m \times 1$) of concentrations in each box, $\bar{E}$ is the corresponding vector of emissions, $\bar{K}$ is the rate matrix ($m \times m$) representing at the generic off-diagonal position (i, j) the total rate of transfer from box i to box j , and at the diagonal position (i, i) the total removal rate of chemical at box i toward all other boxes and out of the system in form of degradation.

The scheme can be applied to all the following cases, or a combination of them:

1. a single phase: different boxes correspond to different locations in space (for instance, the water phase of a lake can be subdivided into subregions to describe chemical concentrations arising from emissions at different points);
2. more phases: different boxes correspond to different phases (for instance, we can use a box for each phase to describe the exchanges along a column of air, water, and sediment);
3. a single phase and more chemical species (for instance, we can describe organic nitrogen, ammonia, nitrogen dioxide, and nitric nitrogen in a lake where denitrification occurs in the water phase). In such case, the transfer rates represent rates of reaction from/to each chemical species.

This reasoning enables both multimedia and multispecies modeling through systems of equations all similar to the above. Instead of each individual phase one might model together an environmental medium including more phases, under the equilibrium assumption. In this case, the concentration in a medium can be a total concentration, or a concentration in a specific phase, the other phases being automatically determined by the equilibrium assumptions. A thorough discussion of box models can be found in [1].

References

- [1] Mackay, D. (2001). *Multimedia Environmental Models: The Fugacity Approach*, 2nd Edition, Lewis Publishers, New York, p. 261 (1st edition 1991).
- [2] Domenico, P.A. & Schwartz, F.W. (1990). *Physical and Chemical Hydrogeology*, John Wiley & Sons.
- [3] Schwarzenbach, R.P., Gschwend, P.M. & Imboden, D.M. (1993). *Environmental Organic Chemistry*, John Wiley & Sons, New York.
- [4] Underwood, B.Y. (1984). *Chapter 2 – Dry Deposition, in Review of Specific Effects in Atmospheric Dispersion Calculations*, Euratom Report EUR 8935EN.
- [5] Vanoni, V.A. (1975). *Sedimentation Engineering*, ASCE, NY.
- [6] Thibodeaux, L.J. (1996). *Environmental Chemodynamics – Movement of Chemicals in Air, Water and Soil*, Wiley-Interscience, New York.
- [7] Schnoor, J.L. (1996). *Environmental Modeling: Fate and Transport of Pollutants in Water, Air, and Soil*, Wiley-Interscience, New York.
- [8] Weber, W.J. & Di Giano, F.A. (1996). *Process Dynamics in Environmental Systems*, Wiley-Interscience, New York.
- [9] Carslaw, H.S. & Jaeger, J.C. (1971). *Conduction of Heat in Solids*, Oxford University Press, London.
- [10] Ellsworth, T.R. & Butters, G.L. (1993). Three-dimensional analytical solutions to the advection-dispersion equation in arbitrary Cartesian coordinates, *Water Resources Research* **29**(9), 3215–3226.
- [11] Abbott, M.B. & Mimmis, A.W. (1998). *Computational Hydraulics*, 2nd Edition, Ashgate, Aldershot.
- [12] De Marsily, G. (1986). *Quantitative Hydrogeology – Groundwater Hydrology for Engineers*, Academic Press, New York.
- [13] Bear, J. & Verruijt, A. (1987). *Theory and Applications of Transport in Porous Media*, Reidel Publishing Company, Dordrecht.
- [14] Campbell, G.S. (1985). *Soil Physics with BASIC – Transport Models for Soil/Plant Systems*, Elsevier Science, Amsterdam.
- [15] USGS (2006). *MODFLOW Software and Documentation*, at <http://water.usgs.gov/nrp/gwsoftware/modflow2000/modflow2000.html> (last accessed Nov 2006).
- [16] USGS (2006). *SUTRA Software and Documentation*, at <http://water.usgs.gov/nrp/gwsoftware/sutra/sutra.html> (last accessed Nov 2006).
- [17] Gambolati, G., Putti, M. & Paniconi, C. (1999). Three dimensional model of coupled density dependent flow and miscible salt transport in *Seawater Intrusion in Coastal Aquifers: Concepts, Methods, and Practices*, J. Bear, A.H.-D. Cheng, S. Sorek, D. Ouazar & I. Herrera, eds, Kluwer Academic Publishers, Dordrecht, pp. 315–362.
- [18] USEPA (2006). *MT3D Software and Documentation*, at <http://www.epa.gov/ada/csmos/models/mt3d.html> (last accessed Nov 2006).
- [19] USGS, VS2DT software and documentation available at the web site: <http://wwwwbr.cr.usgs.gov/projects/GW-Unsat/vs2di1.2/index.html> (last accessed Oct 22, 2007).
- [20] Rinaldi, S., Soncini-Sessa, R., Stehfest, H. & Tamura, H. (1979). *Modeling and Control of River Quality*, McGraw-Hill, New York.
- [21] Smagorinsky, J. (1963). General circulation experiments with the primitive equations – I. The basic experiment, *Monthly Weather Review* **91**(3), 99–152.
- [22] USACE (2006). *HEC-RAS Software and Documentation*, at <http://www.hec.usace.army.mil/software/hecras/> (last accessed Nov 2006).

- [23] University of Alberta (2006). *River2D Software and Documentation*, at <http://www.river2d.ualberta.ca/index.htm> (last accessed Nov 2006).
- [24] Luyten P.J., Jones J.E., Proctor R., Tabor A., Tett P. & Wild-Allen K. (2006). *COHERENS – A Coupled Hydrodynamical-ecological Model for Regional and Shelf Seas: User Documentation*, MUMM Report, Management Unit of the Mathematical Models of the North Sea, p. 914. COHERENS software and documentation available at <http://www.mumm.ac.be/~patrick/mast/coherens.html> (last accessed Nov 2006).
- [25] USEPA (2006). *QUAL2K Software and Documentation*, at <http://www.epa.gov/athens/wwqts/html/qual2k.html> (last accessed Nov 2006).
- [26] USDA (2006). *SWAT Software and Documentation*, at <http://www.brc.tamus.edu/swat/> (last accessed Nov 2006).
- [27] USEPA (2006). *WASP Software and Documentation*, at <http://www.epa.gov/athens/wwqts/html/wasp.html> (last accessed Nov 2006).
- [28] USEPA (2006). *ISC3 Software and Documentation*, at http://www.epa.gov/scram001/dispersion_alt.htm (last accessed Nov 2006).
- [29] USEPA (2006). *CALINE Software and Documentation*, at http://www.epa.gov/scram001/dispersion_prefrec.htm#caline3 (last accessed Nov 2006).
- [30] Sigma Research Corporation, *CALPUFF Software and Documentation*, at http://www.epa.gov/scram001/dispersion_prefrec.htm#calpuff (last accessed Nov 2006).
- [31] DIMULA software and documentation available at the web site: <http://www.maind.it/software/windimula.htm> (last accessed Oct 22, 2007).
- [32] ADMS software and documentation available at the web site: <http://www.cerc.co.uk/software/> (last accessed Oct 22, 2007).
- [33] Zanetti, P. (1990). *Air Pollution Modeling: Theories, Computational Methods and Available Software*, Van Nostrand Reinhold, New York, pp 454.
- [34] Environ, *CAMx Software and Documentation*, at <http://www.camx.com/> (last accessed Nov 2006).
- [35] USEPA (2006). *AERMOD Software and Documentation*, at http://www.epa.gov/scram001/dispersion_prefrec.htm#aermod (last accessed Nov 2006).
- [36] Krol, M.C., Houweling, S., Bregman, B., van den Broek, M., Segers, A., van Velthoven, P., Peters, W., Dentener, F. & Bergamaschi, P. (2005). The two-way nested global chemistry-transport zoom model TM5: algorithm and applications atmos, *Chemical Physics* **5**, 417–432.
- [37] PSU/NCAR (2006). *MM5 Software and Documentation*, at <http://www.mmm.ucar.edu/mm5/> (last accessed Nov 2006).

Related Articles

Cumulative Risk Assessment for Environmental Hazards

Environmental Performance Index

Environmental Risk Regulation

ALBERTO PISTOCCHI

Fault Detection and Diagnosis

The detection and diagnosis of malfunctions in technical systems is of great practical significance [1]. Such systems include production equipment (such as chemical plants, steel mills, paper mills, and power stations), transportation vehicles (ships, airplanes, and automobiles), and household appliances (washing machines and air conditioners). In any of these systems, malfunctions of system components may lead to damage of the equipment itself, degradation of its function or product, jeopardy of its mission, and hazard to human life. While the need to detect and diagnose malfunctions is as old as the construction of such systems, advanced fault detection has been made possible only by the proliferation of the computer. Fault detection and diagnosis actually mean a scheme in which a computer monitors the technical equipment to signal any malfunction and determine the components responsible for it. The detection and diagnosis of the fault may be followed by automatic actions enabling the fault to be corrected, such that the system may operate successfully even under the particular faulty condition, or it may lead to the emergency shutdown of the system.

Diagnostic Concepts

Fault detection and diagnosis apply to both the basic technical equipment and the actuators and sensors attached to it. In the case of a chemical plant, for example, the former includes the reactors, distillation columns, heat exchangers, compressors, storage tanks, and piping. Typical faults are leaks, plugs, surface fouling, and broken moving parts. The actuators are mostly valves, together with their driving devices (electric motors and hydraulic or pneumatic drives). The sensors are devices measuring the different physical variables in the plant (such as thermocouples, pressure diaphragms, and flowmeters). Actuator and sensor fault detection is very important because these devices are quite prone to faults [1, 2].

The on-line or real-time detection and diagnosis of faults mean that the equipment is constantly monitored during its regular operation by a permanently connected computer, and any discrepancy is signaled

almost immediately. On-line monitoring is very important for early detection of any component malfunction before it can lead to more substantial equipment failure. In contrast, off-line diagnosis involves the monitoring of the system by a special, temporarily attached device, under special conditions (for example, car diagnostics at a service station).

The diagnostic activity may be broken down into several logical stages. Fault detection is the indication of something going wrong in the system. Fault isolation is the determination of the fault location (the component which malfunctions), while fault identification is the estimation of its size. On-line systems usually contain the detection and isolation stages; in off-line systems, detection may be superfluous. Fault identification is usually less important than the other two stages.

Fault detection and isolation can never be performed with absolute certainty, because of circumstances such as noise, disturbances, and model errors (see **Subjective Expected Utility**). There is always a trade-off between false alarms and missed detections, the proper balance depending on the particular application. In professionally supervised large plants, false alarms are better tolerated and missed detections may be more critical, while in consumer equipment (including cars) the situation may be the opposite [1, 2].

Approaches

A number of different approaches to fault detection and diagnosis may be used, individually or in combination.

Limit Checking

In this approach, which is the most widely used, system variables are monitored and compared to preset limits. This technique is simple and appealing, but it has several drawbacks. The monitored variables are system outputs that depend on the inputs; to make allowance for the variations of the latter, the limits often need to be chosen conservatively. Furthermore, a single-component fault may cause many variables to exceed their limits, so it may be extremely difficult to determine the source. Monitoring the trends of system variables may be more informative, but it also suffers from the same limitations as limit checking (see **Decision Modeling**) [2].

Special/Multiple Sensors

Special sensors may be applied to perform the limit-checking function (for example, as temperature or pressure limit sensors) or to monitor some fault-sensitive variable (such as vibration or sound). Such sensors are employed mostly in noncomputerized systems. *Multiple sensors* may be applied to measure the same system variable, providing physical redundancy. If two sensors disagree, at least one of them is faulty. A third sensor is needed to isolate the faulty component (and select the accepted measurement value) by “majority vote”. Multiple sensors may be expensive, and they provide no information about actuator and plant faults [2].

Frequency Analysis

This procedure, in which the Fourier transforms of system variables are determined, may supply useful information about fault conditions. The healthy plant usually has a characteristic spectrum, which will change when faults are present. Particular faults may have their own typical signature (peaks at specific frequencies) in the spectrum (*see Non- and Semi-parametric Models and Inference for Reliability Systems*) [2].

Fault-Tree Analysis

Fault trees are the graphic representations of the cause–effect relations in the system. On the top of the tree there is an undesirable or catastrophic system event (“top event”), with the possible causes underneath (“intermediate events”) down to component failures or other elementary events (“basic events”) that are the possible root causes of the top event. The logic relationships from bottom-up are represented by AND and OR (or more complex) logic gates. Fault trees can be used in system design to evaluate the potential risks associated with various component failures under different design variants (bottom-up analysis). In a fault diagnosis framework, the tree is utilized top down; once the top event is observed, the potential causes are analyzed by following the logic paths backwards; (see example below (*see Influence Diagrams*) [3]).

Parameter Estimation

This procedure utilizes a mathematical model of the monitored system. The parameters of the model

are estimated from input and output measurements in a fault-free reference situation. Repeated new estimates are then obtained on-line in the normal course of system operation. Deviations from the reference parameters signify changes in the plant, and a potential fault. The faulty component location may be isolated by computing the new physical plant parameters and comparing them with those from the model [4].

Consistency Checking

This is another way of utilizing the mathematical system model. The idea is to check if the observed plant outputs are consistent with the outputs predicted by the model. Discrepancies indicate a deviation between the model and the plant (parametric faults) or the presence of unobserved variables (additive faults). This testing concept is also called *analytical redundancy* since the model equations are used in a similar way as multiple sensors under the physical redundancy concept described above [4, 5].

As a preparation for fault monitoring by analytical redundancy methods, a mathematical model of the plant needs to be established. This may be done from “first principles” relying on the theoretical understanding of the plant’s operation, or by systems identification using experimental data from a fault-free plant.

The actual implementation of fault monitoring usually consists of two stages (Figure 1). The first is residual generation; residuals are mathematical quantities expressing the discrepancy between the actual plant behavior and the one expected on the

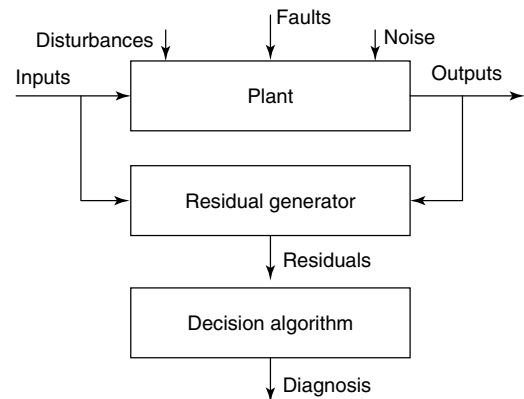


Figure 1 Model-based fault detection and diagnosis

basis of the model. Residuals are nominally zero; they become nonzero by the occurrence of faults. The second stage is residual evaluation and decision making; the residuals are subjected to threshold tests and logic analysis.

The detection of faults is complicated by the presence of disturbances and noise; these affect the plant outputs, just as faults do, and may cause the residuals to become nonzero, leading to false alarms. With model-based methods, model errors are another potential source of false alarms. Designing residual generators that are (at least somewhat) robust with respect to noise, disturbances, and model errors is a fundamental issue in technical diagnostics.

Fault isolation requires specially manipulated sets of residuals. In the most frequently used approach, residuals are arranged so that each one is sensitive to a specific subset of faults ("structured residuals"). Then, in response to a particular fault, only a fault-specific subset of residuals triggers its test, leading to binary "fault codes" (see example below).

Principal Component Analysis (PCA)

In this approach, empirical data (input and output measurements) is collected from the plant. The eigenstructure analysis of the data covariance matrix yields a statistical model of the system; the eigenvectors point at the "principal directions" of the relationships in the data while the eigenvalues indicate the data variance in the principal directions. This method is successfully used in the monitoring of large systems; by revealing linear relations among the variables the dimensionality of the model is significantly reduced. Faults may be detected by relating plant observations to the normal spread of the data; outliers indicate abnormal system situations. Residuals may also be generated from the principal component model, allowing the utilization of analytical redundancy methods in this framework (see **Copulas and Other Measures of Dependency**) [6].

Example of Fault-Tree Analysis – Simple Electrical Circuit

The schematic of a simple electrical circuit in which a light is operated by a pair of three-way switches is shown in Figure 2. (Such circuits are used e.g., in

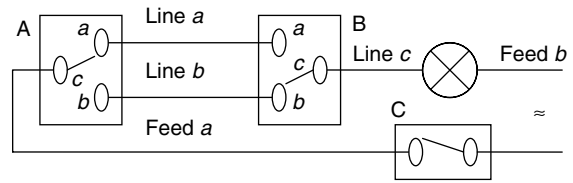


Figure 2 Simple electrical circuit: a lamp operated by a pair of three-way switches [Reproduced from [1]. © McGraw-Hill, 2006.]

long hallways.) Figure 3 shows the detailed fault tree of the circuit. The tree goes down to subcomponents (contacts of the switches) in order to illustrate the more complex logic relations on this simple system. Note that nonfailure events (operating conditions) are also among the basic events because such conditions (the position of each switch) determine whether or not a particular failure event triggers the top event [1].

Example of Consistency Checking – Application to Automobile Engines

Traditionally, a few fundamental variables, such as coolant temperature, oil pressure, and battery voltage, have been monitored in automobile engines by using limit sensors. With the introduction of on-board microcomputers, the scope and number of variables that can be considered have been extended. Active functional testing may be applied to at least one actuator, typically the exhaust-gas recirculation valve. Model-based schemes to cover the components affecting the vehicle's emission control system are gradually introduced by manufacturers.

As an example, consider the intake manifold subsystem (Figure 4). The subsystem includes two actuators, the throttle (THR) and the exhaust-gas recirculation (EGR) valve, and two sensors, measuring the manifold absolute pressure (MAP) and mass air flow (MAF). The two outputs (MAP and MAF) depend on the two inputs (THR and EGR); their relationship may be determined empirically or by physical modeling. The two model equations are

$$\text{MAP} = f_1(\text{THR}, \text{EGR}) \quad (1)$$

$$\text{MAF} = f_2(\text{THR}, \text{EGR}) \quad (2)$$

By manipulating these equations, four residuals may be generated so that each one is insensitive to

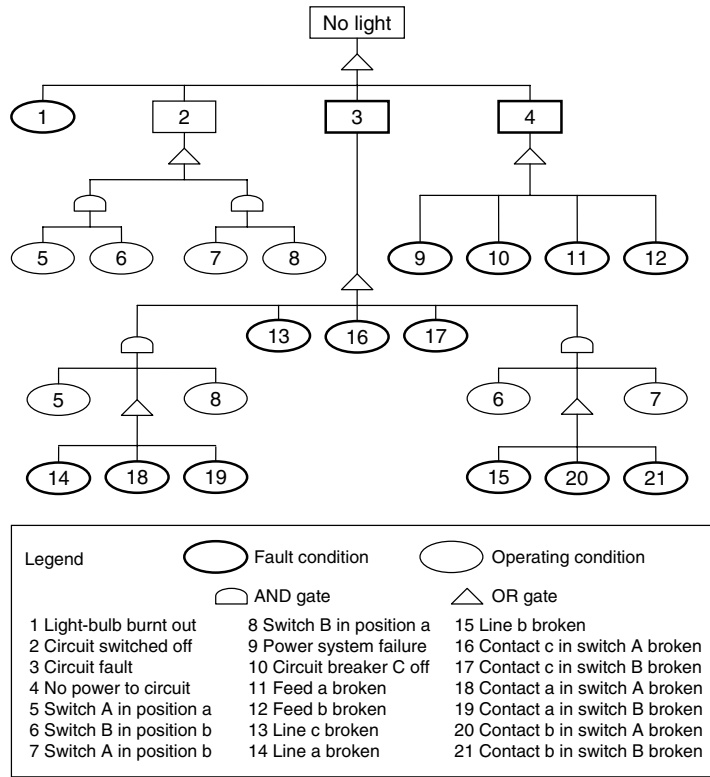


Figure 3 Detailed fault tree of the circuit shown in Figure 2 [Reproduced from [1]. © McGraw-Hill, 2006.]

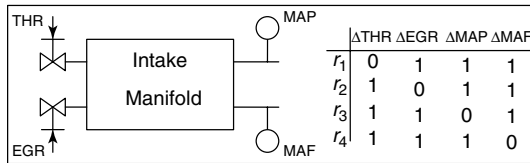


Figure 4 Car-engine manifold system and fault-to-residual structure matrix

one of the component faults, ΔTHR , ΔEGR , ΔMAP , ΔMAF , and depends on the three others:

$$r_1 = \varphi_1(\Delta\text{EGR}, \Delta\text{MAP}, \Delta\text{MAF}) \quad (3)$$

$$r_2 = \varphi_2(\Delta\text{THR}, \Delta\text{MAP}, \Delta\text{MAF}) \quad (4)$$

$$r_3 = \varphi_3(\Delta\text{THR}, \Delta\text{EGR}, \Delta\text{MAF}) \quad (5)$$

$$r_4 = \varphi_4(\Delta\text{THR}, \Delta\text{EGR}, \Delta\text{MAP}) \quad (6)$$

These four residuals form a structured set, supporting fault isolation. Their structure is also shown in Figure 4, where a “1” means that a residual

does respond to a fault while a “0” means it does not. The columns of the structure matrix are fault codes, the logic response of the residual set to each particular fault.

With an automotive diagnostic system, the critical issue is to find sufficiently general models so that a single scheme may function well across an entire automobile product line and under widely varying operating conditions [1, 7].

References

- [1] Gertler, J. (2006). Fault analysis, *Encyclopedia of Science and Technology*, 10th Edition, McGraw-Hill.
- [2] Gertler, J. (1988). Survey of model-based failure detection and isolation in complex plants, *IEEE Control Systems Magazine* 8(7), 3–11.
- [3] Clemens, P.L. (2004). *Fault Tree Analysis*, 4th Edition, Jacobs Sverdrup, <http://www.sverdrup.com/safety/fta.pdf>.
- [4] Gertler, J. (1998). *Fault Detection and Diagnosis in Engineering Systems*, Marcel Dekker.

- [5] Gertler, J. (2002). All linear methods are equal – and extendible to (some) nonlinearities, *International Journal of Robust and Nonlinear Control* **12**, 629–648.
- [6] Chiang, L.H., Braatz, R.D. & Russel, E. (2001). *Fault Detection and Diagnosis in Industrial Systems*, Springer.
- [7] Gertler, J., Costin, M., Fang, X., Kowalczyk, Z., Kunker, M. & Monajemy, R. (1995). Model-based diagnosis for automotive engines – algorithm development and testing on a production vehicle, *IEEE Transactions on Control Systems Technology* **3**, 61–69.

Related Articles

Logistic Regression

Ruin Theory

Stress Screening

JANOS J. GERTLER

Federal Statistical Systems as Data Sources for Intelligence, Investigative, or Prosecutorial Work

A quantitative data analysis, and in the current instance, the quantitative analysis of risk in the context of threats to homeland security, requires at the outset appropriate and good quality data. The ultimate analysis can only be as informative as the underlying data permit. Data analysts have understood this truism both practically and theoretically. Thus, side by side with the development of the techniques of statistical analysis, statisticians and officials in agencies, public and private, that conduct quantitative analysis have been pioneers in the development, management, and preservation of quantitative data. This data development, management, and preservation process has, in turn, led to advances in standardization of techniques, from coding systems, to questionnaire and survey design, to collection procedures.

Data development thus broadly defined has, in turn, led to the creation of data archives for secondary use and related advances in the management and use of secondary data. The historical analysis of microdata from past censuses; the aggregation of transaction data for data mining (*see Early Warning Systems (EWSs) for Predicting Financial Crisis; Risk in Credit Granting and Lending Decisions: Credit Scoring*) applications; the matching of data from multiple systems in capture/recapture applications to identify issues of coverage – all these areas of research and analysis would not be possible without the preservation of large-scale data collections and procedures for providing access to these data systems for later research.

Administrative Data and Statistical Data

National states, and in the current context, the United States, have been the chief agencies for the development and preservation of statistical data about the economic, social, and political activities within their boundaries. While conducting the routine activities

of collecting and spending taxes, identifying and registering market transactions, providing benefits to individuals and organizations, regulating disputes, or monitoring and licensing activities, governments collect and record huge amounts of information. This information is “administrative” – that is collected as part of the activities of the government itself.

Such information is not necessarily tabular or “quantitative”, but in the eighteenth century analysts initially called *statists* recognized that the resulting tabular records could be the basis of a new form of knowledge, which they called *statistics*. These early statisticians recognized that records had value apart from their administrative functions. They pioneered in the development of statistical data analysis, that is, the analysis of patterns of aggregates, and eventually they led the parallel development of governmental collection of tabular data for statistical purposes only.

Governmental data systems also record information about natural phenomena, which may be compiled into statistical data. Weather readings; water levels in lakes and rivers; records of disease in people, animals, and plants; chemical analyses of foods and nutrients, to name just a few types of data systems, may also be deployed for risk analyses and general statistical analysis. For current purposes, namely, the use of governmental data sources for intelligence, investigative, and prosecutorial work, the most controversial uses of federal data systems have arisen with respect to data on individuals, groups, or organizations suspected of illegal activity or threats to national security, and the discussion below focuses on these data.

There has been a close and often confused administrative relationship between state data collections for public informational purposes and those for surveillance (*see Managing Infrastructure Reliability, Safety, and Security; Dynamic Financial Analysis Sampling and Inspection for Monitoring Threats to Homeland Security*), tax collection, benefits administration, or even military control of people. The literature on the history of state surveillance either does not acknowledge or does not distinguish well between the practices of statistical data collection and administrative data collection [1–3] or in the recent special issue of *Contemporary Sociology*, titled “Taking a Look at Surveillance Studies” [4–8]. Higgs [9] provides an explicitly historical analysis. By the early nineteenth century, when “information came of age” [10], individuals and agencies involved

in “statistical” analysis developed their own practices, logic, and rules to guarantee the integrity of their new form of knowledge. Officials in government statistical agencies recognized that the quality of the data they collected and the credibility of the analyses derived from such data could be compromised if it was seen to be part of an administrative activity of the state. Thus, they strove to separate the statistical agencies of government from those of the administrative agencies involved in the normal surveillance and policing activities of the state. They justified such a divorce on the grounds that, without administrative independence, they could not provide unbiased, accurate, precise, and trustworthy information and statistical analysis for the proper functioning of the state [11].

Over the past century, official statistical agencies have developed practices guaranteeing respondent confidentiality (for organizational and individual respondents) both for ethical and practical reasons of assuring truthful data collection. In current official statistical practice, there is a sharp distinction between data collected for an administrative purpose and those collected for a statistical or research purpose. Administrative data (*see* **Occupational Risks; Public Health Surveillance**) are data that fundamentally serve administration purposes. The administrative agency collects identifiable information on individual respondents to secure the administration of the tax, program, or service. Statistical data are essentially anonymous, concerned with distributions, patterns, and averages, and individual identifiers function to guarantee data integrity. One does not want two responses from the same business or household. Administrative data, which are to be analyzed for statistical purposes, are stripped of their identifiers and anonymized to make the data like survey data.

The two forms of quantitative data, e.g., administrative data and statistical data, look quite similar at first blush. From the respondent’s perspective, the distinctions between these two types of data are not clear. The questionnaire one fills out to apply for a benefit program, or the tax form one completes, does not look terribly different from a census or survey form. The purpose of the administrative questionnaire is clear, e.g., the intended benefit or the tax obligation. However, the public may be wary about offering information for a purely statistical questionnaire without a clear sense of how the information will be used. Over time, therefore, safeguards have

been put in place to guarantee the data are used properly and to explain the functions of statistical data.

The system of safeguards is generally called *statistical confidentiality* (*see* **Privacy Protection in an Era of Data Mining and Record Keeping; Ethical Issues in Using Statistics, Statistical Methods, and Statistical Sources in Work Related to Homeland Security**) – that is, the system of procedures that govern who can access data, how sensitive information is protected, and how data are preserved. When an individual, organization, or business enterprise responds to statistical inquiries from the federal government, be it the population census, an economic census, or a periodic survey, the questionnaire contains a pledge that the answers to the questions will be kept confidential, will be used for statistical purposes only, and that no harm can come to an individual or business from truthful responses.

This pledge, akin to the requirements that health care providers maintain confidentiality of medical records, or the guarantees received by participants from institutional review boards in research projects, is a fundamental principle of statistical practice and has been credited for making US government statistics reliable and thus an effective base upon which to build public policy. It is also a venerable pledge, dating back in the United States to presidential census proclamations of a century ago, and the efforts of generations of government officials to clarify and refine procedures [12–14].

Deploying Statistical Data for Intelligence, Investigative, or Prosecutorial Work

It is in the context of the development of the two types of quantitative data collected and used by government on people and organizations that one must consider the issues for deploying statistical data for intelligence, investigative, or prosecutorial work. The simplest and unproblematic use of statistical data in this context is the same as that in the use of such data for any policy context, namely, to define and describe the parameters and frame the extent of any particular analysis of risk. For an analysis of the quality of the screening systems at airports, what is the size of the screener labor force? How many people are screened? How long does an average screening take? What is the likelihood of a failure in the system? Which locations are busiest, and most likely to fail under stress? For a profile of drug

trafficking, can the characteristics of known violators be compared with a profile of potential suspects in a larger population, to identify the most effective use of surveillance and police resources to deter the activity? Can the characteristics of individuals infected with HIV be used to target populations for screening and prevention? In many of these applications, the data underlying the statistical analyses are themselves sampled or complete files of existing administrative records, systematized and made publicly available for research. See, for example, the special “topical” archives at the Interuniversity Consortium for Political and Social Research sponsored by federal agencies, including the National Institute on Aging; National Institute of Justice; Substance Abuse and Mental Health Services Administration (<http://www.icpsr.umich.edu/cocoon/ICPSR/all/archives.xml>).

More problematic uses of statistical data are also possible, however. Particularly in the context of a major and immediate threat to public safety and national security, the pressures on the investigative and surveillance agencies to deter the threat can lead to (a) improper use of statistical data and short cuts in the name of the emergency or (b) full-scale efforts to break the anonymized character of statistical data and in effect deploy it for surveillance purposes. Both such uses may seem appropriate in an emergency context. Both are likely to be ineffective in responding to or deterring the original threat and will also damage the integrity of the statistical system in the long run.

The pressure to make “emergency” use of inappropriate statistical data, for a “close enough” or “down and dirty” analysis, is likely to be made when the threat is at its sharpest and the public pressure to “do something” is greatest. The data analyst in this situation may well be aware that the data deployed, for example, for an ethnic profile of potential terrorists, is highly flawed, but the effort proceeds because the need seems so great (and the risk to the researcher in doing nothing is greater than that in using the flawed data). The Department of Homeland Security’s (DHS) 2003 request for a zip-code tabulation from the 2000 census of “Arabs” is an example of such a rushed and flawed data request [15]. As many commentators have noted, the request implied a blanket racist targeting of “Arabs” – in itself an improper form of analysis, and one which

arguably undermines the trust in the statistical system overall.

Further, in this instance, it appears that neither the original agency request, nor the census officials who supplied the tabulations, discussed the capacity of these particular data to provide an indicator for homeland security. The data were derived from the sample census question on “ancestry”, and thus captured any individual who identified as “Arab” – even if that meant that the respondent was reporting that one relative from several generations back had come from the “Arab” Middle East. This flaw points to a larger issue. The variables, coding systems, and sampling rates in statistical data systems of long independence, such as the census, are designed to address issues of statistical interest. Thus the “ancestry” data in the census address the issues of immigration and cultural assimilation over generations if not centuries. They do not readily translate into tabulated data that are useful for immediate investigation and surveillance. As with all data use, the analyst needs to understand the structure, logic, strengths, and limitations of the data.

The second and more troubling potential use of statistical data for surveillance emerges when an administrative agency makes an explicit request to “breach” confidentiality and deploy microdata collected under a pledge of confidentiality for “emergency” surveillance and investigative purposes. The US Patriot Act, drafted in the weeks after the attacks on the World Trade Center and the Pentagon, contained a provision (Section 508) permitting the Attorney General to access confidential data collected in the National Center for Education Statistics for investigation and prosecution of suspected terrorists [16]. This provision remains in the law, though to date, there has been no analysis of its implementation and use. Similar legislative language permitted administrative agencies access to confidential statistical data in the Commerce Department during World War II. Information from the 1940 census, economic censuses, and lists of businesses and industries was provided to both the war planning agencies and the surveillance agencies “for use in connection with the conduct of the war” from 1942 to 1947 [17].

There is limited research on the disclosures from World War II, though it is clear that the statistical agencies were quick to support repeal of the wartime measure at the end of the war. The impact on response rates of the blurring of the line between

statistical and administrative data for surveillance has also not been systematically explored, though survey research on the more general issue of “trust” indicates damage to a statistical system that violates its pledge of confidentiality [18]. There is need for continued research on the difference between surveillance data and statistical data, and for the best ways to deploy statistical data for the analysis of security concerns without jeopardizing the integrity of the statistical data in the process.

References

- [1] Dandeker, C. (1990). *Surveillance, Power and Modernity: Bureaucracy and Discipline from 1700 to the Present Day*, Polity Press, Cambridge.
- [2] Zureik, E. (2001). Constructing Palestine through surveillance practices, *British Journal of Middle Eastern Studies* **28**(2), 205–227.
- [3] Parenti, C. (2004). *The Soft Cage: Surveillance in America from Slavery to the War on Terror*, Basic Books, New York.
- [4] Lyon, D. (2007). Sociological perspectives and surveillance studies: slow journalism and the critique of social sorting, *Contemporary Sociology* **36**(2), 107–111.
- [5] Zureik, E. (2007). Surveillance studies: from metaphors to regulation to subjectivity, *Contemporary Sociology* **36**(2), 112–116.
- [6] Torpey, J. (2007). Through thick and thin: surveillance after 9/11, *Contemporary Sociology* **36**(2), 116–119.
- [7] Cunningham, D. (2007). Surveillance and social movements: lenses on the repression mobilization nexus, *Contemporary Sociology* **36**(2), 120–125.
- [8] Marx, G.T. (2007). Desperately seeking surveillance studies: players in search field, *Contemporary Sociology* **36**(2), 125–130.
- [9] Higgs, E. (2004). *The Information State in England, the Central Collection of Information on Citizens since 1500*, Palgrave Macmillan, New York.
- [10] Headrick, D. (2000). *When Information Came of Age: Technologies of Knowledge in the Age of Reason and Revolution, 1700–1850*, Oxford University Press, New York.
- [11] Anderson, M. (2000). Building the American statistical system in the long 19th century, in *L'ère du Chiffre: Systemes Statistiques et Traditions Nationales/The Age of Numbers: Statistical Systems and National Traditions*, B. Jean-Pierre & J. Prevost, eds, Presses de l'Université du Québec, Québec, pp. 105–130.
- [12] United States Bureau of the Census (2004). *Census Confidentiality and Privacy, 1790–2002*, at <http://www.census.gov/prod/2003pubs/conmono2.pdf>.
- [13] Anderson, M. & Seltzer, W. (2007). Challenges to the confidentiality of U.S. federal statistics, 1910–1965, *Journal of Official Statistics* **23**(1), 1–34.
- [14] Division of Behavioral and Social Sciences and Education, Committee on National Statistics (2005). *Principles and Practices for a Federal Statistical Agency*, 3rd Edition, M.E. Martin, M.L. Straf & C.F. Citro, eds, National Academies Press, Washington, DC.
- [15] El Badry, S. & Swanson, D. (2007). Providing census tabulations to government security agencies in the United States: the case of Arab Americans, *Government Information Quarterly* **24**(2), 470–487.
- [16] Seltzer, W. & Anderson, M. (2003). NCES and the patriot act: an early appraisal of facts and issues, *Proceedings of the American Statistical Association*, 2002, Section on Survey Research Methods, American Statistical Association, Alexandria, pp. 3153–3156.
- [17] Seltzer, W. & Anderson, M. (2007). Census confidentiality under the second war powers act (1942–1947), Paper prepared for presentation at the session on *Confidentiality, Privacy, and Ethical Issues in Demographic Data*, Population Association of America Annual Meeting, March 29–31, 2007, New York, at <http://www.uwm.edu/~margo/govstat/integrity.htm>.
- [18] National Research Council, Panel on Data Access for Research Purposes, Division of Behavioral and Social Sciences and Education, Committee on National Statistics (2005). *Expanding Access to Research Data: Reconciling Risks and Opportunities*, National Academies Press, Washington, DC.

Related Articles

Counterterrorism

Ethical Issues in Using Statistics, Statistical Methods, and Statistical Sources in Work Related to Homeland Security

Privacy Protection in an Era of Data Mining and Record Keeping

MARGO ANDERSON AND WILLIAM SELTZER

Fraud in Insurance

In insurance, fraud is defined as the attempt to hide circumstances or distort reality in order to obtain an economic gain from the contract between the insurer and the policy holder. All insurance branches are vulnerable to fraud [1]. One of the most common types of fraud occurs when the policy holder exaggerates the losses and claims compensation in an amount higher than what would have been a just and fair payment. Even in the underwriting process, information may be withheld or manipulated in order to obtain a lower premium or a coverage that otherwise would have been denied. This is called *underwriting fraud*.

The lines of business corresponding to automobile insurance, health insurance, homeowners' insurance, and worker compensation insurance are known to be frequently affected by fraud because natural claims linked to accidents are more likely to occur than any other claims. Insurers are increasingly concerned about fraud in these areas because they represent a large percentage of the total premiums earned by their companies.

One classical example of fraudulent behavior related to automobile insurance is to make an agreement with a repair shop and claim damages that had occurred to a car before a given accident as that related to that accident. An example of fraud in health insurance is withholding information about preexisting health conditions from the insurer. In homeowners' insurance, an example of a typical fraudulent action is the claim of coverage for a broken appliance (e.g., a dishwasher) as due to an electricity surge that had never taken place. In worker's compensation, fabricated claims for medical treatments are sometimes used to misrepresent and exaggerate wage losses.

Insurance fraud can be fought by prevention and detection. Prevention aims at stopping fraud from occurring and detection includes all kinds of strategies and methods to discover fraud once it has already been perpetrated. It is difficult to analyze insurance fraud because of the small amount of systematic information available about its forms and scope. The Insurance Fraud Bureau Research Register [2] located in Massachusetts provides a vast list of resources on insurance fraud. The register was created in 1993, "to identify all available research

on insurance fraud from whatever source, to encourage company research departments, local fraud agencies and local academic institutions to get involved and use their resources to study insurance fraud and to identify all publicly available databases on insurance fraud which could be made available to researchers".

Types of Fraud

First, regarding fraud committed by the insured, in the literature we find a distinction between soft and hard fraud, that is, between claims in which the damages have been exaggerated, and those comprising damages completely invented in relation to a staged, nonexistent, or unrelated accident (this kind of fraud often requires involvement of other external agents, such as health care providers or car repair shops). Criminal fraud, hard fraud or planned fraud, are illegal activities that can be prosecuted and end with convictions. Soft fraud, buildup fraud, and opportunistic fraud are terms usually applied to cases of loss overstatement or abuse, which is not illegal and more difficult to identify. In buildup fraud costs are inflated, but the damages correspond to an accident actually related to the claim. In opportunistic fraud, the accident is also real but a portion of the claimed amount is related to damages either fabricated or related to a different accident. Most soft fraud are settled by negotiations and the insurer relies on persuasion and confrontation rather than on costly or lengthy court decisions. However, court prosecution has a positive deterrence effect on fraud [3]. Individuals become aware that fraud is a criminal activity and they also see the risks of being caught [4, 5]. Insurance companies also use other classifications of fraud, usually using some specific settings or distinct circumstances as criteria.

Secondly, besides fraud committed by the insured, as discussed above, there could be internal fraud perpetrated within the insurance companies by the agents, insurer employees, brokers, managers, or representatives. They can obstruct the investigations or simply collaborate with customers in fraudulent activities.

Fraud Deterrence

The theory underlying fraud deterrence is most often based on the basic utility model for the policyholder

[6]. The individual decision to perpetrate fraud is very intuitive: the insured would file a fraudulent claim only if the expected utility is larger than the *status quo*. The expected utility is calculated using two scenarios. First, it is assumed with a certain level of probability that the fraud will be discovered, and then the fraudster would certainly lose the opportunity to get any compensation for the claim, and moreover he/she would be fined. Secondly, if the fraud is not detected, then the insured would gain an undue compensation.

Fraud prevention can only succeed if certain mechanisms that could induce one to speak the truth are in place. One example of such a mechanism already used by the insurers is the imposition of significant penalties for fraud, while another is the introduction of lower deductibles or lower premiums based on the no-claims record of the insured. The most recommended mechanism, however, is to commit the insurer to an audit strategy as often as possible, that is, to specify in advance the way in which claims are selected for audit on a regular basis. Thus, audits do not necessarily depend on the characteristics of the claims. Another efficient practice is to inform the insured and the general public that a given company makes all possible efforts to fight fraud.

The relationship between the insurer and the insured through an insurance contract should incorporate incentives for not making fraudulent claims. Once the insurer receives the claim, he or she can decide whether it is profitable to investigate it. If such investigation seems unprofitable, the claimed amount is paid without any further checking.

Any detection system should be combined with a random audit procedure [7]. In that case, any particular claim has equal probability of being investigated. In a regulated market, credible commitment can also be ensured through external fraud investigation agencies funded by the insurers.

Some studies of the theory of insurance fraud consider auditing as a mechanism to control fraud (see [6, 8]). This approach is called *costly state verification* and was introduced by Townsend [9] in 1979. It assumes that through incurring an auditing cost the insurer can obtain valuable information about the claim. It implies that the audit process using models or detection systems and special identification units always discerns whether a claim is fraudulent or not. The challenge is to design a contract that minimizes the insurer costs, including the

total claim payment and the cost of the audit. Models commonly suggested for such purposes use the claimed amount to decide about the application of monitoring techniques. In other words, deterministic audits are compared to random audits [10], or it is assumed that the insurer would commit to some audit strategy [11].

A parallel discussion concerns *costly state falsification* [12, 13], which is based on the supposition that the insurer cannot audit the claims, i.e., determine whether the claims are truthful or not. Moreover, an optimal contract design can also be applied, based on the assumption that the claimant would exaggerate the amount of the loss but at a cost greater than zero. We may conclude that the designing of an auditing strategy is linked to the specification of an insurance contract. The analysis of the contract by Dionne and Gagné [14] revealed that insurance fraud creates a significant problem with resource allocation. For instance, in the case of automobile insurance, the straight deductible contracts shape the falsification behavior of the insured. If the deductible is higher, then the probability of reporting a small loss is lower.

Vázquez and Watt [15] have proposed other theoretical solutions, such as the optimal auditing strategies that minimize the total costs incurred from buildup or the accounting for the cost of paying undetected fraud claims. Costly state verification may be used with the optimal auditing strategy but additional assumptions have to be made about claim size, audit cost, proportion of opportunistic insured, commitment to an investigation strategy, and the extent to which the insured is able to manipulate the audit costs. Recently, a transition is observed from the formal definition of an optimal claims auditing strategy to the calibration of the model on data, and to the derivation of the optimal auditing strategy.

Fraud Detection

Fraud is not self-revealed and, thus, has to be investigated. It is necessary to find evidence of fraud and determine the factors that did not protect the system from the fraudulent actions. In order to produce evidence, particularly if the case has to be proved in the court, it is often necessary to employ intensive specialized investigation skills. If the evidence is not conclusive, the cases are usually

settled with the claimant either through compensation denial or a partial payment.

Auditing strategies serve both deterrence and detection purposes [16]. An automated detection system informs the insurer which claims to audit but it often fails because it cannot adjust itself to the continuously changing environment and the emerging opportunities for committing fraud. More information is necessary and it can be obtained when longitudinal or enriched data are analyzed. A fraud detection system, to be effective, requires unpredictability since only randomness puts the perpetrator at risk of being caught. It is difficult to calculate the return on investment in detection systems because a higher fraud detection rate may reflect an increase in fraudulent activity rather than a better detection system. The calculation of costs and benefits of such a system should also take into account the fact that an unpaid claim carries a commitment effect value to the customers.

The strategies and the detection technology that an insurer can use to detect fraud do not necessarily apply to all lines of business, or all cases. There are unique ways of doing business and how firms handle claims; the kinds of products and the characteristics of the portfolios do impact the fraud detection systems. Cultural differences, habits, and social rules also play a role in this regard.

Tools for fraud detection are equally diverse. These include human resources, including external advisors, data mining, statistical analyses, and ways of monitoring. Methods used for the detection of fraudulent or suspicious claims related to human resources include fraud awareness training, video and audio tape surveillance, manual indicator cards, and internal audits and information collected from agents or informants. Methods of data analysis require collecting both external and internal information, which are processed by means of computer software, preset variables, statistical and mathematical analytical techniques, or geographic data mapping. The cost of investigating fraud is a key factor in determining implementation of a deterrence and detection strategy, and the cost of fraud has to be taken into account when the benefits of a detection strategy are considered [17, 18].

The detection model is used to help insurance companies in making decisions and to ensure that they are better equipped to counter insurance fraud. An audit strategy with a detection model has to be

part of the claims handling process. The adjusting of claims usually puts them into two categories: express paid claims (easily paid) and target claims (that need to be investigated) [19]. Target claims are screened and eventually investigated by specialized investigation units (SIUs), which either are part of the insurance company or are contracted externally (sometimes they are governmental agencies).

Fraud Indicators

Fraud indicators, often called the *red flags*, are measurable characteristics that suggest the occurrence of fraud. Some of them, such as the date of the accident, are objective, while others are subjective. One example of the latter is a binary variable indicating that when the claim was first reported the insured seemed too familiar with the insurance terminology. Often, such expertise about the way in which claims are handled is associated with fraudsters. Other possible indicators can be obtained using text mining processes, from free-form text fields that are kept in customer databases.

Table 1 shows some examples of fraud indicators for automobile insurance that can be found in [19–22].

Some indicators easily signal a fraudulent claim, while others may be misleading or wrong. For example, witnesses are usually considered a reliable source of information on how an accident occurred, but quite often they may not actually be authentic witnesses.

Table 1 Examples of fraud indicators in automobile insurance

Date of subscription to guarantee and/or date of its modification too close to date of accident
Date/time of the accident do not match the insured habits
Harassment from policyholder to obtain quick settlement of a claim
Numerous claims submitted in the past
Production of questionable or falsified documents (copies or bill duplicates)
Prolonged recovery from injury
Shortly before the loss, the insured checked the extent of coverage with the agent
Suspicious reputation of providers
Variations in or additions to the policyholder's initial claims
Vehicle whose value does not match the income of policyholder

Steps for Data Analysis in Detecting Fraud

The claims handling process is linked to the steps for analyzing data, which, for the fraud detection model, are recommended to be as follows:

1. The construction of a random sample of claims, avoiding a selection bias.
2. The identification of red flags or other sorting indicators, avoiding the subjective ones. As time passes from the initial submission of the claim, new indicators may emerge, for instance, related to medical treatment or repair bills.
3. The clustering of claims into homogeneous categories. Sometimes a hypothesis about propensity to fraud may be helpful.
4. The assessment of fraud. Claims are classified as fraudulent or honest externally; however, the adjusters and investigators may make a mistake or have different opinions.
5. The construction of a detection model. Once the sample claims are classified, supervised models provide specific characteristics to the categories of claims, and a score is generated for this purpose [22]. However, the fact that claims are classified on the basis of similarity creates a situation of potential danger because, if a fraudulent claim is identified, similar claims may be also considered fraudulent by association. Another question is whether the claims have been classified correctly or not. Some models are equipped to take a possible misclassification into account [23, 24]. It is also necessary to evaluate the prediction performance of the fraud detection model before its implementation.
6. The monitoring of the results. Expert assessment, cluster homogeneity, and model performance are the foundations of static testing. The real-time operation of the model relates to dynamic testing. The model should be fine-tuned and the investigative proportions should be adjusted to optimize detection of fraud.

References

- [1] Derrig, R.A. (2002). Insurance fraud, *Journal of Risk and Insurance* **69**(3), 271–288.
- [2] Insurance Fraud Bureau Research Register (2002). *Insurance Fraud Research Register*, <http://www.derrig.com/ifrr/index.htm>.
- [3] Derrig, R.A. & Zicko, V. (2002). Prosecuting insurance fraud: a case study of the Massachusetts experience in the 1990's, *Risk Management and Insurance Review* **5**(2), 77–104.
- [4] Clarke, M. (1989). Insurance fraud, *The British Journal of Criminology* **29**, 1–20.
- [5] Clarke, M. (1990). The control of insurance fraud. A comparative view, *The British Journal of Criminology* **30**, 1–23.
- [6] Picard, P. (1996). Auditing claims in the insurance market with fraud: the credibility issue, *Journal of Public Economics* **63**, 27–56.
- [7] Pinquet, J., Ayuso, M. & Guillen, M. (2007). Selection bias and auditing policies for insurance claims, *Journal of Risk and Insurance* **74**(2), 425–440.
- [8] Bond, E.W. & Crocker, K.J. (1997). Hardball and the soft touch: the economics of optimal insurance contracts with costly state verification and endogenous monitoring costs, *Journal of Public Economics* **63**, 239–264.
- [9] Townsend, R.M. (1979). Optimal contracts and competitive markets with costly state verification, *Journal of Economic Theory* **20**, 265–293.
- [10] Picard, P. (2000). Economic analysis of insurance fraud, in *Handbook of Insurance*, G. Dionne, ed, Kluwer Academic Press, Boston.
- [11] Boyer, M. (1999). When is the proportion of criminal elements irrelevant? A study of insurance fraud when insurers cannot commit, in *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation*, G. Dionne & C. Laberge-Nadeau, eds, Kluwer Academic Press, Boston.
- [12] Crocker, K.J. & Morgan, J. (1998). Is honesty the best policy? Curtailing insurance fraud through to optimal incentive contracts, *Journal of Political Economy* **106**, 355–375.
- [13] Crocker, K.J. & Tennyson, S. (1999). Costly state falsification or verification? Theory and evidence from bodily injury liability claims, in *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation*, G. Dionne & C. Laberge-Nadeau, eds, Kluwer Academic Press, Boston.
- [14] Dionne, G. & Gagné, R. (2001). Deductible contracts against fraudulent claims: evidence from automobile insurance, *Review of Economics and Statistics* **83**(2), 290–301.
- [15] Vázquez, F.J. & Watt, R. (1999). A theorem on multi-period insurance contracts without commitment, *Insurance: Mathematics and Economics* **24**(3), 273–280.
- [16] Tennyson, S. & Salsas-Forn, P. (2002). Claims auditing in automobile insurance: fraud detection and deterrence objectives, *Journal of Risk and Insurance* **69**(3), 289–308.
- [17] Viaene, S., Van Gheel, D., Ayuso, M. & Guillen, M. (2004). Cost sensitive design of claim fraud screens, *Lecture Notes in Artificial Intelligence* **3275**, 78–87.
- [18] Viaene, S., Ayuso, A., Guillen, M., VanGheel, D. & Dedene, G. (2007). Strategies to detect and prevent

- fraudulent claims in the automobile insurance industry, *European Journal of Operational Research* **176**(1), 565–583.
- [19] Derrig, R.A. & Weisberg, H.I. (1998). *AIB PIP Claim Screening Experiment Final Report Understanding and Improving the Claim Investigation Process*, AIB Filing on Fraudulent Claims Payment DOI Docket R98-41, Boston.
- [20] Weisberg, H.I. & Derrig, R.A. (1991). Fraud and automobile insurance. a report on the baseline study of bodily injury claims in Massachusetts, *Journal of Insurance Regulation* **9**, 497–541.
- [21] Belhadji, E.-B., Dionne, G. & Tarkhani, F. (2000). A model for the detection of insurance fraud, *Geneva Papers on Risk and Insurance-Issues and Practice* **25**(5), 517–538.
- [22] Artís, M., Ayuso, M. & Guillen, M. (1999). Modelling different types of automobile insurance fraud behaviour in the Spanish market, *Insurance Mathematics and Economics* **24**, 67–81.
- [23] Artís, M., Ayuso, M. & Guillen, M. (2002). Detection of automobile insurance fraud with discrete choice models and misclassified claims, *Journal of Risk and Insurance* **69**(3), 325–340.
- [24] Caudill, S., Ayuso, M. & Guillen, M. (2005). Fraud detection using a multinomial logit model with missing information, *Journal of Risk and Insurance* **72**(4), 539–550.

Related Articles

Benchmark Analysis

MONTSERRAT GUILLEN AND MERCEDES
AYUSO

From Basel II to Solvency II – Risk Management in the Insurance Sector

The Evolution of the Regulatory Framework for Banks and Insurers

In the financial industry, the rules of the committee for banking supervision at the Bank of International Settlement (BIS) in Basel have drastically changed the business in all Organisation for Economic Co-operation and Development (OECD) countries. These rules are known as the *Basel rules*. The first set of Basel rules, called *Basel I*, was launched in 1988.

The justification of a fairly elaborate regulatory framework is based on the interdependence of banks and the role of the banking sector for the overall economy. Banks represent a crucial element of the financial system that in turn, is of utmost importance for the allocation of capital and risk in an economy. Banks, just like the financial system as a whole, bring together the supply of savings and the demand for investment financing and, hence, can significantly affect the real sector of the economy. Without a stable financial system, no country can have firms with healthy financial structures, high investment expenditure, or significant R&D activities. Thus, a stable financial sector is a prerequisite for sustained economic growth and further increase of welfare. The Asian crisis is seen as a textbook example of how damaging the lack of stability in the financial system and, notably, the banking system can be.

The Basel I rules have been adopted by national laws in all OECD countries. The current directive is known under Basel II. Basel II is a rigorous extension of Basel I and will become effective in most countries within this decade.

In 2002, a regulatory framework for the European insurance sector was launched by the European Commission. This framework is called *Solvency II* which does not only sound similar to Basel II – the framework carries the Basel II capital adequacy rules for the banking sector to the determination of the minimum *solvency* (see **Solvency**) capital in the insurance sector.

The justification of a regulatory framework of the insurance sector is frequently based on the

role this sector plays as a means of diversifying risk and allocating risk to market participants who can best cope with the particular risks. In a more global economy, additional risks occur that can have important effects on the magnitude and the direction of international capital flows. Without a sound risk allocation system, risk-related decisions are likely to be suboptimal and, hence, lead to a combination of higher *volatility* (see **Volatility Modeling**), following unanticipated negative events, and to less investment. At least in the life insurance sector, there is another argument supporting a regulatory framework. This has to do with the exposure of an individual who engages in such a contract. Although future income depends heavily on the performance of the insurance company, the individual neither has possibilities to control the activities of the insurer, nor can he liquidate his assets in order to invest the funds somewhere else.

Against this background, this article analyzes two aspects of risk-management regulation. First, it illustrates how the competitive positions in the global banking system have changed as an effect of Basel II and how, analogously, Solvency II will affect the insurance industry. Secondly, since insurance risks are much more uncommon than banking risks, it illustrates how the calculation of a minimum solvency capital ratio for an insurance company is performed. This will be done for two types of insurance companies: property and casualty (P&C) insurers and life insurers.

The Future of Banking – between Markets and Institutions

In the following, three aspects (points 1–3 below) are analyzed regarding the change in the competitive positions of different banking systems due to the initiation and development of the Basel accord:

Point 1: Two different types of financial systems and the development of their competitive positions over the last 15 years are described.

Point 2: The reasons for the change in the competitive positions of these two financial systems are analyzed.

Point 3: The conclusions for the insurance sector, which can be drawn from the changes in the banking sector are analyzed.

We begin with the first point. Two pure forms of banking systems can be distinguished: more financial institution oriented systems as in Germany or Japan and more investment and capital market oriented systems as in the United States or the United Kingdom. The competitive positions of the two pure forms of banking systems have changed significantly since 1988. Since then, banking has become much more market oriented and less affected by financial institutions. The league table of the biggest banks worldwide reflects this change in the global financial system: in 1990, six Japanese banks were ranked among the top six banks in terms of market capitalization and Deutsche Bank was ranked number seven. Barclays, National Westminster, and JP Morgan were the biggest Anglo-Saxon banks ranked 8–10, respectively. In 2005, the biggest bank was the Citigroup. Deutsche Bank was ranked 25 and none of the Japanese banks was ranked among the top 15 banks anymore. This is a dramatic change within only 15 years.

Of course, we can wonder what the reasons were for this dramatic change. This leads us to the second point of our considerations. A major source driving this change is regulation. The Basel accord, launched in July 1988, has enhanced the process of the capital market orientation of the global banking system very considerably. Basel I, the first Basel accord, is still effective. It requires OECD banks to hold equity capital according to the risk of asset and liability positions. Basel II, which has been very intensively discussed since 1999 and which will become effective in the OECD banking industry soon, will strengthen the process of capital market orientation even more. Before the Basel accord was designed, balance sheet risks in banks were covered by a standardized equity fraction of 8% of the assets. This figure has not been supported by valid risk data, it was rather based on intuition. The Basel accord applied risk weights to each balance sheet position, i.e., low risks imply low equity ratios and, therefore, higher returns on equity. Higher risks require a higher amount of comparatively expensive equity capital. Before the launch of Basel I, the more beneficial the bank transactions, the more riskier they were. Banks were particularly successful in environments in which financial institutions were strong. Japan and Germany are two examples for countries with relatively strong financial institutions, but relatively weak financial

markets compared to the United Kingdom and the United States.

Before the Basel accord was launched, a loan to a noninvestment-grade counterparty used to be covered by the same amount of equity capital like a loan to a first-class counterparty. Obviously, since the expected return on a low-rated counterparty is higher than on counterparties with higher ratings (*see Risk in Credit Granting and Lending Decisions: Credit Scoring*), this rule creates incentives for loans with parties that have poor credit qualities and makes loans to first-class counterparties less attractive. Under the Basel II accord, the equity coverage of a bank loan depends on the credit rating of the counterparty. One important effect is that banks tend to buy much more hedging and diversification instruments for credit risk in the capital markets. New asset classes like credit derivatives, collateralized debt obligations, or asset backed securities were created during the same time the Basel rules were developed. All of these instruments are capital market instruments. Product innovations have not only been observed for credit risks, but also for hedging and diversifying interest rate, foreign exchange, commodity price, and stock market risks.

At the same time, banks in countries with a traditional strong capital market orientation have become stronger and gained capitalization weight. The lack of a developed investment banking industry in Germany and Japan compared to very strong US and UK investment banks caused the market capitalization of German and Japanese banks to significantly lag behind their Anglo-Saxon counterparts in the last 15 years. Basel I and Basel II are shifting risk management away from financial institutions toward more transparent capital markets. This negative consequence for the German and the Japanese banks has obviously been underestimated by bank lobbyists in countries with emphasis on strong bank orientation.

These arguments lead to our third and last point of our considerations, namely, how the experience from the banking system will be transferred to the insurance sector. In February 2002, the European Commission launched a regulatory framework for the European insurance industry. This framework is referred to as *Solvency II*. Solvency II adopts the Basel II three pillars of prudential regulation to the insurance sector. Probably from 2008 onward, Solvency II will affect European insurance companies very significantly. The major goal is, like Basel II,

to achieve more transparency for the regulators and the public. This is reached by adapting the minimum solvency capital (MSC) in an insurance company, to its risks. The question now is whether such a regulatory shift will affect the insurance business to the same extent as was the case in the banking sector. Insurance risks used to be completely institutionalized. The market for capital market instruments securitizing (*see Securitization/Life*) typical insurance risks is not yet very developed. Insurance institutions diversify insurance risks and nondiversifiable insurance risks are reinsured with re-insurance companies. However, the exciting question is whether the past developments of the banking industry are going to be future trends in the insurance sector. It is not unlikely that we will observe capital market products securitizing typical insurance risks like catastrophe or weather risks in the P&C sector, and longevity or mortality risks in the life sector. Examples for derivatives on all of these risks already exist. Hence, the insurance sector has not yet gone as far as the banking sector. This is the case because the Solvency II initiative is much younger than the Basel rules. One should also not forget that Solvency II is currently a European and not yet a global initiative. This is also a reason for a slower shift of insurance risks toward capital markets than we observe for formerly typical bank risks. However, the development towards the capital market in the insurance sector has already begun. Instruments like weather derivatives, cat bonds (catastrophe bonds), or longevity bonds are not yet extremely common, but they exist and are gaining significance. We observe many indicators that predict a development away from an institutional orientation toward a market orientation of the insurance sector. Currently, the still powerful insurance companies should watch the development very carefully and should have in mind what happened earlier to the formerly powerful European banks.

Risk Management for P&C Insurers

Although the Solvency II framework is very closely related to Basel II, the implementation of a risk-management framework for insurers is more than just a simple application of a value at risk (VaR) measurement model to an insurance portfolio. Insurance companies deal with risk classes that are completely different from those of banks. This holds for both

types of insurance business, i.e., P&C as well as life insurance.

The determination of the joint probability distribution of all risks is done based on a dynamic financial analysis (DFA) approach as described by, for example, Kaufmann *et al.* [1]. The DFA goal is to identify the MSC of an insurance company. The MSC is the amount of capital required to provide a given level of safety to policy holders over a specified time horizon, given the enterprise-wide risk distribution of the insurer. Hence, the MSC is comparable to the VaR, which we know very well from bank risk management.

Figure 1 illustrates the transmission mechanism of a comparatively simple DFA model for the factors affecting the insurer's assets and liabilities. The figure shows that there are 10 probability distributions, which are needed to characterize the following 8 risk factors: interest rates, inflation, stock market returns, credit risk (i.e., the risk for a reinsurer's default), risk in the growth rate in the number of contracts, catastrophic- and noncatastrophic-risks, and the pattern of the payments over time which follows an insurance event.

The joint probability distribution of these risk factors expresses the distribution for the loss reserves of the insurance company. This cannot be determined analytically. A computation of the MSC must therefore be based on a simulation approach. DFA is such a simulation approach. Figure 2 shows an example of the joint distribution of the loss reserves. Both, the catastrophe and the noncatastrophe distribution depend on two single probability distributions, namely, loss frequency and loss severity. Severity is typically modeled as lognormal distribution. The frequency for catastrophe losses is often modeled by a Poisson distribution (*see Logistic Regression*) while noncatastrophe loss frequencies are characterized by normal distributions. Hence, the joint loss reserves distribution in Figure 2 is made up by four nonstandard probability distributions. The joint distribution can therefore not be expressed in analytical terms.

Risk Management for Life Insurers

While the preceding section dealt with P&C insurers, this section addresses the risk management of life insurance companies. The situation in determining the MSC is completely different from P&C insurers. While the complexity of risk factors was most

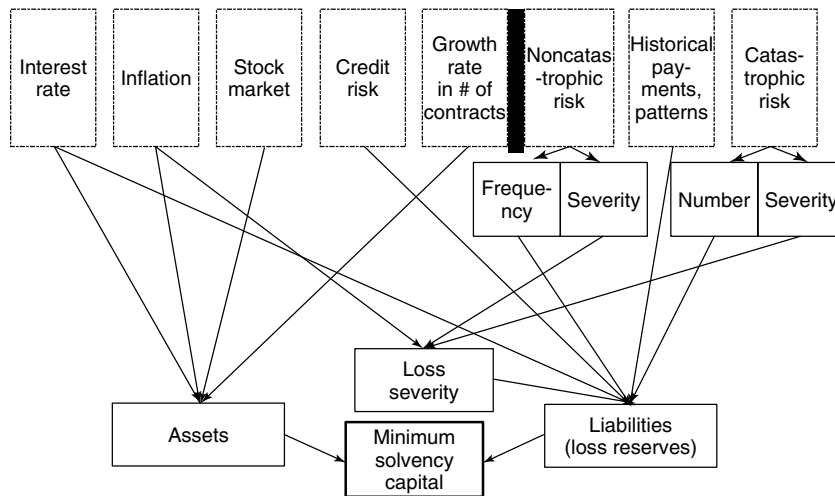


Figure 1 Transmission mechanism of the DFA analysis for a P&C insurer

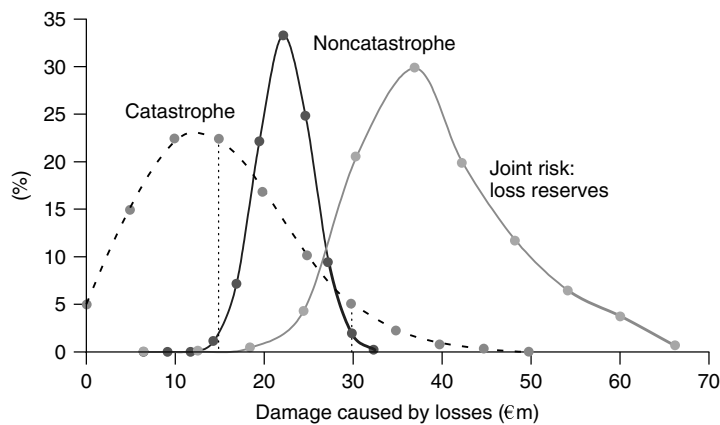


Figure 2 Noncatastrophe losses and catastrophe losses generate the distribution of the loss reserves

important in the P&C sector, there are two dominant risk factors in the life insurance business. These are interest rate risks and biometric risks (e.g., longevity).

What makes the risk management of life insurance contracts complicated are the number and the complexity of the embedded options in such contracts. Figure 3 summarizes the most important risks in the life insurance business. It shows that a life insurance contract can be considered as a combination of a risk-free bond, an interest rate bonus option, and a surrender option.

One of the best known options in life insurance contracts is the bonus option, guaranteeing

a minimum return to the policy holder (*see Bonus–Malus Systems*). This option depends on the interest rate situation on the capital market. It can therefore be hedged by fixed income derivatives, which are traded at the financial markets typically with great liquidity. The risk management of the bonus option is therefore comparable to interest rate risk management in the banking sector.

The surrender option indicated in Figure 3 is more difficult to handle. It provides the policy holder's contract with an early liquidation feature. The early liquidation likelihood is due to changes in market or regulatory conditions. It might therefore be driven

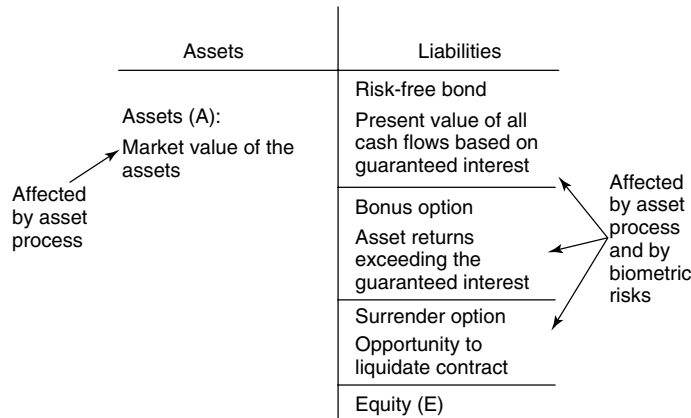


Figure 3 Risks affecting the assets and liabilities of life insurance companies

by changes in the tax regime associated with life insurance contracts. Moreover, the early liquidation likelihood does also depend on the interest rate scenario in the capital market: the higher the interest rates are, the higher will be the likelihood of an early exercise. Förterer [2] estimates the value of the early liquidation option to be roughly 5% of the entire liability value.

Finally, for the valuation and the risk management of a life insurance contract, it is not sufficient to apply the VaR (*see Value at Risk (VaR) and Risk Measures*) tools developed for bank risk management. There is a recent literature dealing with the evaluation of life insurance contracts. Two excellent examples for this literature are Grosen and Jorgensen [3] or Neilson and Sandmann [4].

Conclusions

Basel II has enhanced the capital market orientation of the banking sector very significantly. Since 2002, Solvency II is a regulatory framework of the European Commission for the regulation of the European insurance sector. Since Solvency II takes Basel II as its archetype, this article deals with the question whether a much stronger capital market orientation can also be expected for the insurance sector. The conclusion is that it is very likely that more and more insurance risks will be transferred to the capital market, although insurance risks deviate from bank risks in many respects. This article also shows the most important risk factors in the P&C insurance sector

and points to the DFA as the most frequently used tool to deal with these risks.

References

- [1] Kaufmann, R., Gadmer, A. & Klett, R. (2001). Introduction to dynamic financial analysis, *ASTIN Bulletin* **31**(1), 213–249.
- [2] Förterer, D. (2000). *Ertrags- und risikosteuerung von lebensversicherern*, Dissertation, Universität St. Gallen.
- [3] Grosen, A. & Jørgensen, P. (2000). Fair valuation of life insurance liabilities: the impact of interest guarantees, surrender options, and bonus policies, *Insurance: Mathematics and Economics* **26**, 37–57.
- [4] Nielsen, J.A. & Sandmann, K. (2002). The fair premium of an equity-linked life and pension insurance, in *Advances in Finance and Stochastics: Essays in Honor of Dieter Sondermann*, P. Schönbucher & K. Sandmann, eds, Springer-Verlag, Heidelberg, pp. 219–255.

Further Reading

Nakada, P., Koyluoglu, H.U. & Collignon, O. (1999). P&C RAROC: a catalyst for improved capital management in the property and casualty insurance industry, *The Journal of Risk Finance*, **Fall** 1–18.

Related Articles

Alternative Risk Transfer

Copulas and Other Measures of Dependency

Value at Risk (VaR) and Risk Measures

MARKUS RUDOLF AND MICHAEL FRENKEL

Game Theoretic Methods

After the September 11, 2001, terrorist attacks on the World Trade Center and the Pentagon, and the subsequent anthrax attacks in the United States, there has been an increased interest in security. However, security problems pose some significant challenges to conventional methods for risk analysis [1, 2]. One critical challenge is the intentional nature of terrorist attacks [3].

Protecting against intentional attacks is fundamentally different from protecting against accidents or acts of nature. In particular, an adversary may adopt a different offensive strategy in response to observed security measures, in order to circumvent or disable those measures. For example, Ravid [4] argues that since the adversary can change targets in response to defensive investments, “investment in defensive measures, unlike investment in safety measures, saves a lower number of lives (or other sort of damages) than the apparent direct contribution of those measures.” Game theory [5] (*see Managing Infrastructure Reliability, Safety, and Security; Optimal Risk-Sharing and Deductibles in Insurance; Risk Measures and Economic Capital for (Re)insurers*) provides a way of taking this into account. In particular, game theory identifies the optimal defense against an optimal attack. While it may not be possible to actually predict terrorist attacks, protecting against an optimal attack can arguably be described as conservative. This article discusses approaches for applying game theory to the problem of defending against intentional attacks.

Applications of Game Theory to Security

Game theory has a long history of being applied to security, beginning with military applications [6–10]. It has also been extensively used in political science [11–13]; e.g., in the context of arms control. Recently, game theory has also been applied to computer security [14, 15]. Many applications of game theory to security have been by economists; see, for example [16–27]. Much of this work has been designed to provide “policy insights” [25] – e.g., on the relative merits of public *versus* private funding of defensive investments [23], or deterrence *versus* defense [19, 20, 22, 27]. Since the events of September 11, however, there has been increased

interest in using game theory not only to explore the effects of different policies, but to generate detailed guidance in support of operational-level decisions (*see Experience Feedback*) – e.g., which assets to protect, or how much to charge for terrorism insurance.

For example, Enders and Sandler [18] and Hausken [21] study substitution effects in security. Enders and Sandler observe that “installation of screening devices in US airports in January 1973 made skyjackings more difficult, thus encouraging terrorists to substitute into other kinds of hostage missions.” Similarly, they note, “If . . . the government were to secure its embassies or military bases, then attacks against such facilities would become more costly on a per-unit basis. If, moreover, the government were not at the same time to increase the security for embassy and military personnel when outside their facilities, then attacks directed at these individuals (e.g., assassinations) would become relatively cheaper” (and hence, presumably, more frequent).

Clearly, security improvements that appear to be justified without taking into account the fact that attacks may be deflected to other targets may turn out to be wasteful (at least from a public perspective) if they merely deflect attacks to other targets of comparable value. Therefore, the Brookings Institution [28] has recommended that “policy makers should focus primarily on those targets at which an attack would involve a large number of casualties, would entail significant economic costs, or would critically damage sites of high national significance.”

The Brookings recommendation constitutes a reasonable “zero-order” suggestion about how to prioritize targets for investment. However, it is important to go beyond the zero-order heuristic of protecting only the most valuable or potentially damaging targets, to account for the success probabilities of different attacks, since terrorists appear to take the probability of success into account in their choice of targets. For example, Woo [29] has observed that “al-Qaeda is . . . sensitive to target hardening,” and that “Osama bin Laden has expected very high levels of reliability for martyrdom operations.” Thus, even if a successful attack against a given target would be highly damaging, that target may not merit as much defensive investment as a target that is less valuable but more vulnerable (and hence, more attractive to attackers).

Early models that take the success probabilities of potential attacks into account include [29–31].

Both the Brookings recommendations [28] and the results in [30] still represent “weakest-link” models (*see* **Extreme Values in Reliability**), in which defensive investment is allocated only to the target(s) that would cause the most damage if attacked. However, such weakest-link models can be unrealistic in practice. For example, Arce and Sandler [16] note that weakest-link solutions “are not commonly observed among the global and transnational collective action problems confronting humankind.” In particular, real-world decision makers typically “hedge” by defending additional targets, to cover contingencies such as misestimating which targets are most attractive to attackers.

The models in [29] and [31] achieve the more realistic result of defensive hedging at optimality, but do so with a restrictive (and possibly questionable) assumption. In particular, these models assume that the attacker can observe “the marginal effectiveness of defense... at [each] target” [31], but not which defenses have been implemented. More recent work [32, 33] achieves the result of defensive hedging at equilibrium in a different manner – assuming (perhaps conservatively) that attackers can observe defensive investments perfectly, but that defenders are uncertain about the attractiveness of possible targets to the attackers. Interestingly, hedging does not always occur even in this model. In particular, the results suggest that it is often optimal for the defender to put no investment at all into some targets – especially when the defender is highly resource constrained, and the various potential targets differ greatly in their values (both of which are likely to be the case in practice).

Recently, researchers have begun to focus on allocation of defensive resources in the face of threats from both terrorism and natural disasters [34–36]. In addition, some researchers have begun formally taking into account the effect of deterrence or retaliation on terrorist recruitment; see for example [24, 37].

Security as a Game between Defenders

The work discussed above views security primarily as a game between attacker and defender. However, defensive strategies adopted by one agent can also affect the incentives faced by other defenders. Some types of defensive actions (such as installation of visible burglar alarms or car alarms) may increase

risk to other potential victims (a negative externality), leading to overinvestment in security. Conversely, other types of defensive actions – such as vaccination, fire protection, or use of antivirus protection software – decrease the risk to other potential victims (a positive externality). This type of situation can result in underinvestment in security, and “free riding” by defenders [21, 25].

To better account for security investments with positive externalities, Kunreuther and Heal [38] propose a model of interdependent security (IDS) in which agents are vulnerable to “infection” from other agents; Heal and Kunreuther [39] apply this model to the case of airline security. More specifically, this body of work assumes that the consequences of even a single successful attack would be catastrophic – leading, for example, to business failure. Under this assumption, Kunreuther and Heal show that failure of one agent to invest in security can make it unprofitable for other agents to invest, even when they would otherwise do so. Moreover, this game can have multiple equilibrium solutions (e.g., one in which all defenders invest in security, and another in which none invest), so coordinating mechanisms can help ensure that the socially optimal level of investment is reached.

Recent work [40] has extended these results to attacks occurring over time (rather than a “snapshot” model). In this case, differences in discount rates among agents can lead some agents with low discount rates not to invest in security, if other agents (e.g., those with high discount rates) choose not to invest. Thus, heterogeneous time preferences can complicate the task of achieving security.

Unlike the above work, Keohane and Zeckhauser [22] consider both positive and negative externalities among defenders. In particular, they note that reduction of population in a major city (e.g., because of relocation decisions by members of the public) may make all targets in that city less attractive to terrorists, while government investment to make a location safer will tend to attract increased population, making that location more attractive to potential attackers.

Combining Risk Analysis and Game Theory

Conventional risk analysis does not explicitly model the adaptive response of potential attackers, and hence may overstate the effectiveness and cost-effectiveness of defensive investments. However,

much game-theoretic security work focuses on non-probabilistic games. Even those models that explicitly consider the success probabilities of potential attacks – e.g., [29–33, 36, 38–40] – generally treat individual assets in isolation, and fail to consider their possible role as components of a larger system. Therefore, combining risk analysis with game theory could be fruitful in studying intentional threats to complex systems such as critical infrastructure (*see Managing Infrastructure Reliability, Safety, and Security*).

Hausken [41] has integrated probabilistic risk analysis and game theory (although not in the security context), by interpreting system reliability as a public good, and considering the incentives of players responsible for maintaining particular components of a larger system. In particular, by viewing reliability as a game between defenders responsible for different portions of the system, he identifies relationships between the series or parallel structure of the system being analyzed and classic games such as the coordination game, battle of the sexes, chicken, and prisoner’s dilemma.

Rowe [42] argues that the implications of “the human variable” in terrorism risk have yet to be adequately appreciated, and presents a framework for evaluating possible defenses in light of terrorists’ ability to “learn from experience and alter their tactics”. This approach has been used in practice to help prioritize defensive investments among multiple potential targets and threat types. Similarly, Paté-Cornell and Guikema [43] discuss the need for “periodic updating of the model and its input” to account for the dynamic nature of terrorism, but do not formally use game theory.

Banks and Anderson [44] apply similar ideas to the threat of intentionally introduced smallpox. Their approach embeds risk analysis in a game-theoretic formulation of the defender’s decision, accounting for both the adaptive nature of terrorism, and also uncertainty about the costs and benefits of particular defenses. They conclude that this approach “captures facets of the problem that are not amenable to either game theory or risk analysis on their own”.

Bier *et al.* [30] use game theory to study security of simple series and parallel systems, as a building block to more complex systems. Results suggest that defending series systems against intentional attack is extremely difficult, if the attacker knows about the system’s defenses. In particular, the attacker’s ability

to respond to the defender’s investments deprives the defender of the ability to allocate defensive investments according to their cost-effectiveness; instead, defensive investments in series systems must equalize the strength of all defended components, consistent with results in [8]. Recent work [45] begins to extend these types of models to more complicated system structures (including both parallel and series subsystems).

Conclusions

As noted above, protecting complex systems against intentional attacks is likely to require a combination of game theory and risk analysis. Risk analysis by itself does not account for the attacker’s responses to security measures, while game theory often deals with individual assets in isolation, rather than with complex networked systems (such as computer systems, electrical transmission systems, or transportation systems). Approaches that combine risk models and game-theoretic methods can take advantage of the strengths of both approaches.

Acknowledgment

This material is based upon work supported in part by the US Army Research Laboratory and the US Army Research Office under grant number DAAD19-01-1-0502, by the US National Science Foundation under grant number DMI-0 228 204, by the Midwest Regional University Transportation Center under project number 04–05, and by the United States Department of Homeland Security through the Center for Risk and Economic Analysis of Terrorism Events (CREATE) under grant number N00014-05-0630. Any opinions, findings, and conclusions or recommendations expressed herein are those of the author and do not necessarily reflect the views of the sponsors. I would also like to acknowledge the numerous colleagues and students who have contributed to the body of work discussed here.

Revised and adapted from chapters in: *Modern Statistical and Mathematical Methods in Reliability*, World Scientific Publishing Company, ISBN: 981-256-356-3; and *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, Springer, ISBN: 038-732-904-8.

References

- [1] Bedford, T. & Cooke, R. (2001). *Probabilistic Risk Analysis: Foundations and Methods*, Cambridge University Press, Cambridge.

- [2] Bier, V.M. (1997). An overview of probabilistic risk analysis for complex engineered systems, in *Fundamentals of Risk Analysis and Risk Management*, V. Molak, ed, Lewis Publishers, Boca Raton, pp. 67–85.
- [3] Brannigan, V. & Smidts, C. (1998). Performance based fire safety regulation under intentional uncertainty, *Fire and Materials* **23**, 341–347.
- [4] Ravid, I. (2002). *Theater Ballistic Missiles and Asymmetric War*, The Military Conflict Institute, [http://www.militaryconflict.org/TBM%20and%20Asymmetric%20War%20L2%20\(1\)](http://www.militaryconflict.org/TBM%20and%20Asymmetric%20War%20L2%20(1)).
- [5] Fudenberg, D. & Tirole, J. (1991). *Game Theory*, MIT Press, Cambridge.
- [6] Berkovitz, L.D. & Dresher, M. (1959). A game-theory analysis of tactical air war, *Operations Research* **7**, 599–620.
- [7] Berkovitz, L.D. & Dresher, M. (1960). Allocation of two types of aircraft in tactical air war: a game-theoretic analysis, *Operations Research* **8**, 694–706.
- [8] Dresher, M. (1961). *Games of Strategy: Theory and Applications*, Prentice Hall, Englewood Cliffs.
- [9] Haywood, O.G. (1954). Military decision and game theory, *Journal of the Operations Research Society of America* **2**, 365–385.
- [10] Leibowitz, M.L. & Lieberman, G.J. (1960). Optimal composition and deployment of a heterogeneous local air-defense system, *Operations Research* **8**, 324–337.
- [11] Brams, S.J. (1975). *Game Theory and Politics*, Free Press, New York.
- [12] Brams, S.J. (1985). *Superpower Games: Applying Game Theory to Superpower Conflict*, Yale University Press, New Haven.
- [13] Brams, S.J. & Kilgour, D.M. (1988). *Game Theory and National Security*, Basil Blackwell, Oxford.
- [14] Anderson, R. (2001). Why information security is hard: an economic perspective, Presented at *The 18th Symposium on Operating Systems Principles*, October 21–24, Lake Louise; also *The 17th Annual Computer Security Applications Conference*, December 10–14, New Orleans; <http://www.ftp.cl.cam.ac.uk/ftp/users/rja14/ec on.pdf>.
- [15] Schneier, B. (2000). *Secrets and Lies: Digital Security in a Networked World*, John Wiley & Sons, New York.
- [16] Arce, M.D.G. & Sandler, T. (2001). Transnational public goods: strategies and institutions, *European Journal of Political Economy* **17**, 493–516.
- [17] Brauer, J. & Roux, A. (2000). Peace as an international public good: an application to Southern Africa, *Defence and Peace Economics* **11**, 643–659.
- [18] Enders, W. & Sandler, T. (2004). What do we know about the substitution effect in transnational terrorism? in *Researching Terrorism: Trends, Achievements, Failures*, A. Silke & G. Ilardi, eds, Frank Cass, London, pp. 119–137.
- [19] Frey, B.S. (2004). *Dealing With Terrorism – Stick or Carrot?* Edward Elgar, Cheltenham.
- [20] Frey, B.S. & Luechinger, S. (2003). How to fight terrorism: alternatives to deterrence, *Defence and Peace Economics* **14**, 237–249.
- [21] Hausken, K. (2006). Income, interdependence, and substitution effects affecting incentives for security investment, *Journal of Accounting and Public Policy* **25**, 629–665.
- [22] Keohane, N.O. & Zeckhauser, R.J. (2003). The ecology of terror defense, *Journal of Risk and Uncertainty* **26**, 201–229.
- [23] Lakdawalla, D. & Zanjani, G. (2005). Insurance, self-protection, and the economics of terrorism, *Journal of Public Economics* **89**, 1891–1905.
- [24] Rosendorff, B. & Sandler, T. (2004). Too much of a good thing? The proactive response dilemma, *Journal of Conflict Resolution* **48**, 657–671.
- [25] Sandler, T. & Arce, M.D.G. (2003). Terrorism and game theory, *Simulation and Gaming* **34**, 319–337.
- [26] Sandler, T. & Enders, W. (2004). An economic perspective on transnational terrorism, *European Journal of Political Economy* **20**, 301–316.
- [27] Siqueira, K. & Sandler, T. (2006). Terrorists versus the government: strategic interaction, support, and sponsorship, *Journal of Conflict Resolution* **50**, 1–21.
- [28] O’Hanlon, M., Orszag, P., Daalder, I., Destler, M., Gunter, D., Litan, R. & Steinberg, J. (2002). *Protecting the American Homeland*, Brookings Institution, Washington, DC.
- [29] Woo, G. (2003). Insuring against Al-Qaeda, *Insurance Project Workshop*, National Bureau of Economic Research, <http://www.nber.org/~confer/2003/insurance03/woo.pdf>.
- [30] Bier, V.M., Nagaraj, A. & Abhichandani, V. (2005). Protection of simple series and parallel systems with components of different values, *Reliability Engineering and System Safety* **87**, 313–323.
- [31] Major, J. (2002). Advanced techniques for modeling terrorism risk, *Journal of Risk Finance* **4**, 15–24.
- [32] Bier, V.M., Oliveros, S. & Samuelson, L. (2007). Choosing what to protect: strategic defensive allocation against an unknown attacker, *Journal of Public Economic Theory* **9**, 563–587.
- [33] Bier, V.M. (2007). Choosing what to protect, *Risk Analysis* **27**, 607–620.
- [34] Powell, R. (2007). Defending against terrorist attacks with limited resources, *American Political Science Review* **101**, 527–541.
- [35] Woo, G. (2006). Joint mitigation of earthquake and terrorism risk, *Proceedings of the 8th U.S. National Conference on Earthquake Engineering* San Francisco, California.
- [36] Zhuang, J. & Bier, V.M. (2007). Balancing terrorism and natural disasters: defensive strategy with endogenous attacker effort, *Operations Research* **55**, 976–991.
- [37] Das, S.P. & Lahiri, S. (2006). A strategic analysis of terrorist activity and counter-terrorism policies, *Topics in Theoretical Economics* **6**, <http://www.bepress.com/cgi/viewcontent.cgi?article=1295&context=bejte>.

-
- [38] Kunreuther, H. & Heal, G. (2003). Interdependent security, *Journal of Risk and Uncertainty* **26**, 231–249.
 - [39] Heal, G. & Kunreuther, H. (2005). IDS models of airline security, *Journal of Conflict Resolution* **49**, 201–217.
 - [40] Zhuang, J., Bier, V.M. & Gupta, A. (2007). Subsidies in interdependent security with heterogeneous discount rates, *The Engineering Economist* **52**, 1–19.
 - [41] Hausken, K. (2002). Probabilistic risk analysis and game theory, *Risk Analysis* **22**, 17–27.
 - [42] Rowe, W.D. (2002). Vulnerability to terrorism: addressing the human variables, in *Risk-Based Decisionmaking in Water Resources X*, Y.Y. Haimes, D.A. Moser & E.Z. Stakhiv, eds, American Society of Civil Engineers, Reston, pp. 155–159.
 - [43] Pate-Cornell, E. & Guikema, S. (2002). Probabilistic modeling of terrorist threats: a systems analysis approach to setting priorities among countermeasures, *Military Operations Research* **7**, 5–20.
 - [44] Banks, D.L. & Anderson, S. (2006). Combining game theory and risk analysis in counterterrorism: a smallpox example, in *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, A.G. Wilson, G.D. Wilson & D.H. Olwell, eds, Springer, New York, pp. 9–22.
 - [45] Azaiez, N. & Bier, V.M. (2007). Optimal resource allocation for security in reliability systems, *European Journal of Operational Research* **181**, 773–786.

VICKI M. BIER

Gene–Environment Interaction

Broad Concept

Whether nature or nurture is the predominant determinant of human traits has been a source of societal and scientific discussions for much of the last century. More recently, at least in the fields of medicine, epidemiology, and toxicology, discussion has moved away from the questions of which influence predominates and, instead, has focused on developing a fuller understanding of how genetic makeup and history of environmental exposures work together to influence an individual's traits, particularly susceptibility to diseases. This newer focus is the essence of a broad concept of gene–environment interaction: the extensive interplay of genetics and environment on risk of disease.

Closely related to this broad view of gene–environment interaction is the idea of a multifactorial or complex disease. Few diseases can be considered to have a single genetic or environmental origin; most involve influences of both types. In this context, “environment” is usually broadly construed and may include virtually anything other than “genes”. Environment could encompass things such as exposure to chemical toxicants, infectious organisms, dietary or nutritional factors, aspects of lifestyle such as smoking or exercise habits, hormonal milieu, and social dynamics associated with, for example, employment or marriage (*see* **Environmental Hazard**). Environmental factors could act anytime from conception onward throughout life. Complex diseases like cancer, asthma, and schizophrenia are widely viewed as arising from genetic predisposition, potentially involving many genes, acting together with any number of environmental factors throughout a lifetime. The complexity of sorting out the mutual influences of genes and environment on disease risk is clearly monumental.

Despite the evident challenge, the study of gene–environment interaction potentially offers substantial benefits to science and society. Consider the example of cancer. Environmental exposures appear to contribute to most cancer, but evidence is accumulating that polymorphisms in various genes, particularly metabolism and DNA-repair genes, may modulate

the effect of these exposures and, thereby, influence susceptibility [1]. Eventually, a fuller understanding of the joint effects of genetic and environmental risk factors could lead to new public health interventions that reduce cancer incidence and new clinical approaches that treat it or prevent recurrence. By identifying susceptibility genes and the metabolic pathways through which they act, one might uncover previously unrecognized carcinogens or novel genetic targets for therapeutics. Studies directed at genetically susceptible subpopulations might better delineate low levels of risk attributable to common exposures whose role was not previously recognized. A widespread exposure, even one that induces only a small increase in risk, can contribute substantially to a population's cancer burden and may, after careful risk assessment, warrant public health intervention. On the other hand, a rare genetic variant that, in the presence of certain exposures, increases an individual's risk of disease substantially would have little population-wide import but an intense impact on the few individual carriers – a situation where individually tailored counseling or clinical intervention would be invaluable.

Heterogeneity of Effect

Although the broad concept of gene–environment interaction provides the outlook and long-term goals that motivate much current biomedical research, the actual work is more restricted in scope, typically limited in any individual report to a few genes and a few exposures. Consequently, most of the statistical or methodological work that underpins current studies addresses a much narrower concept of gene–environment interaction, one that can be illustrated by considering a single genetic polymorphism and a single exposure factor. In this simple context, a defining question for this narrow concept of gene–environment interaction is whether the influence of the polymorphism and exposure together on some measure of disease risk is different from that expected based on combining their individual contributions. An equivalent question is whether the influence of one factor (e.g., exposure) differs across levels of the other factor (genotype). Thus, this narrower concept of gene–environment interaction involves heterogeneity in the statistically assessed “effect”, whether risk-enhancing or risk-reducing, of one factor across levels of the other factor.

A widely cited example of unambiguous gene–environment interaction is the disease phenylketonuria (PKU), which can lead to mental retardation [2]. A typical diet contains a certain level of the amino acid phenylalanine. Individuals who carry two defective copies of the gene whose protein product converts phenylalanine to tyrosine develop symptoms of PKU when they consume a typical diet. When these individuals consume a diet with highly restricted levels of phenylalanine, they have no symptoms. Similarly, individuals with at least one functional copy of the gene metabolize phenylalanine properly and show no symptoms. Thus, PKU requires both the genetic variant and the exposure; absence of either eliminates the adverse outcome. In this example, the outcome when genotype and exposure occur together clearly exceeds the outcome predicted on the basis of either factor alone.

While the PKU example is classic, deciding whether gene–environment interaction exists is rarely so clear-cut. In fact, the defining question itself remains ambiguous until one specifies precisely what the phrase “expected based on each individually” means. Disease outcomes might be measured on a binary scale (present or absent), an ordered categorical scale (none, mild, moderate, or severe), or a continuous scale (survival time or blood levels of some key biomarkers). Similarly, an environmental factor might be assessed on any of these scales, though genetic factors are typically assessed as binary (variant allele present or not) or categorical (zero, one, or two copies of a variant allele). Certainly, in different contexts, different statistical models would be used and implicitly set different baselines for measuring interaction. Even within a single context, alternate statistical models could be employed. A widely used class of statistical models is the generalized linear model where the outcome Y , conditional on predictors, has a distribution in the exponential family, which admits binary or continuous outcomes. In a regression setting, with a single predictor G measuring genotype and a single predictor E measuring exposure, one might model the expected value (mean) of Y conditional on G and E , $\mu(Y|G, E)$, as depending on G and E through

$$h(\mu(Y|G, E)) = \beta_0 + \beta_G G + \beta_E E \quad (1)$$

Here β_0 , β_G , and β_E are regression coefficients and $h(\cdot)$ is a strictly monotone function known as a *link*. $h(\cdot)$ could represent, among others, the identity

function, logarithmic function, or logit function. This regression model specifies that, in an outcome scale determined by $h(\cdot)$, the expected change in mean outcome corresponding to a one-unit change in E is β_E regardless of the value of G . Similarly, of course, the expected change in mean outcome corresponding to a unit change in G is β_G regardless of the value of E . Model (1) defines an expected outcome based on G and E , where the effects of G and E are additive in the scale defined by the selected link $h(\cdot)$. Thus, equation (1) provides an additive model for scale $h(\cdot)$ and represents a no-interaction null model for that scale. If, for example, G is binary and E is continuous, equation (1) describes $h(\mu(Y|G, E))$ as two parallel lines.

Among many possible ways to model departure from additivity, a simple, convenient, and often used regression model is the following:

$$h(\mu(Y|G, E)) = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} G \cdot E \quad (2)$$

The product or interaction term, $\beta_{GE} G \cdot E$, encodes a departure from the additive expected joint influence of G and E defined by equation (1). Thus, equation (2) embodies the idea that the outcome of G and E together differs from that expected from summing their individual contributions. In equation (2), again in the scale specified by $h(\cdot)$, the change in mean outcome corresponding to a unit change in E is $\beta_E + \beta_{GE} G$, a quantity that depends explicitly on G . A similar expression, depending on E , applies for the change in mean outcome for a unit change in G . Whenever $\beta_{GE} \neq 0$, gene–environment interaction is said to exist because the change in mean outcome associated with exposure differs across genotypes (alternatively, the change associated with genotype differs across exposure).

In epidemiology where the outcome is often binary (1 = disease present, 0 = disease absent), Y has a Bernoulli distribution with a conditional mean $\mu(Y|G, E)$. In fact, $\mu(Y|G, E)$ in this setting is the conditional probability of disease given the values of G and E , denoted by $\Pr(Y = 1|G = g, E = e)$ or by P_{ge} . Under equation (1), algebraic manipulations yield

$$h(P_{ge}) - h(P_{g0}) - h(P_{0e}) + h(P_{00}) = 0 \quad (3)$$

as a condition that must hold for interaction to be absent in the scale specified by link $h(\cdot)$. Taking g and e as positive and working with equation (2), one can

show that the left-hand side of equation (3) is positive whenever the coefficient β_{GE} is positive (sometimes called a *superadditive* or *synergistic interaction* because the expected outcome when G and E occur together is larger than expected under the additive model). Alternatively, the left-hand side of equation (3) is negative whenever β_{GE} is negative (sometimes called a *subadditive* or *antagonistic interaction* because the expected outcome when G and E occur together is smaller than expected under additivity).

An important characteristic of the narrow concept of gene–environment interaction enunciated above is that it is fundamentally a statistical concept of interaction, one that is defined in terms of coefficients in regression models with no necessary implication of causality. The modeling process provides a quantitative description of the joint influence of G and E . For data analysis, the choice of model is somewhat arbitrary, though limited to an extent by the type of outcome, owing to the imperative to use a model that fits the observed data well, and by any constraints imposed on what models might be estimable under the sampling design of the study. To the extent that the model itself is arbitrary, whether or not gene–environment interaction exists is also arbitrary. Suppose that data are available from a cohort study and a referent or “zero” level is designated for both G and E . If the link is the identity function, $h(P) = P$, the condition specified in equation (3) would involve differences of risk differences being zero, namely,

$$(P_{ge} - P_{g0}) - (P_{0e} - P_{00}) = 0 \quad (4)$$

If instead the link is the natural logarithm function, equation (3) would be equivalent to a certain ratio of relative risks being unity, namely,

$$\frac{RR_{ge}}{(RR_{g0} \cdot RR_{0e})} = 1$$

where

$$RR_{GE} = \frac{P_{GE}}{P_{00}} \quad (5)$$

This derived condition indicates why an additive model for the logarithm of risk is often described as a *multiplicative model* for relative risk. Similarly, for commonly used logistic regression where the link is the logit or log-odds function, equation (3) would correspond to a certain ratio of odds ratios being unity, namely,

$$\frac{OR_{ge}}{(OR_{g0} \cdot OR_{0e})} = 1$$

where

$$OR_{GE} = \frac{P_{GE}(1 - P_{00})}{(1 - P_{GE})P_{00}} \quad (6)$$

These three no-interaction conditions are distinct except in certain degenerate cases (e.g., $P_{0e} = P_{00}$) or in certain limits (e.g., OR s approximate RR s for rare diseases). In general, degenerate cases aside, if there is no interaction under a given link (e.g., logarithmic), there must be interaction under an alternate link (e.g., identity) and *vice versa*. Thus, the presence or absence of gene–environment interaction generally depends on the statistical model used to assess effects (see reference [3] for a numerical example). By the same token, a superadditive interaction under one link can correspond to a subadditive interaction under another.

On the other hand, some states of nature may correspond to the presence of interaction ($\beta_{GE} \neq 0$) regardless of the link function considered. Suppose that both G and E are binary (so that $g, e \in \{0, 1\}$ where 1 = “present”, 0 = “absent”) and further suppose that $P_{11} > P_{01} = P_{10} = P_{00}$ (this configuration, sometimes called *pure interaction*, corresponds to the synergistic interaction of the PKU scenario described earlier). Here, the no-interaction conditions for all three specific link functions fail to hold; and, in fact, equation (3) cannot be satisfied for any strictly monotone link function. Interactions, like the one in this example, that cannot be transformed by changes in outcome scale are sometimes called *qualitative interactions*. With two binary predictors, qualitative interactions would typically involve risk-enhancing effects of exposure for one genotype but null (or even risk-reducing) effects of exposure for the other genotype. Interactions that are not qualitative are called *quantitative*.

Causal Interactions

The concept of interaction as heterogeneity of effect on a specified outcome scale is a satisfactory one for many purposes. If a model aptly describes the data at hand, inferences based on that model can be scientifically useful. On the other hand, knowing that a statistical model on a certain scale requires an interaction term to adequately describe data may not tell the scientist anything fundamental about the underlying causal mechanism. In this sense, interaction *versus* no interaction is essentially a descriptive distinction;

if genotype and exposure together provide a larger effect than either alone, whether or not that larger effect is designated as interaction or simply as additive depends on the scale chosen.

The term “interaction” is widely used in scientific discussion and can mean a variety of things – most of which have causal implications distinct from the statistical meaning involving heterogeneity of effect discussed so far. Suppose that an epidemiologic case–control study of a disease has, through logistic regression analysis, revealed a synergistic gene–environment interaction. The underlying biological mechanism that explains how the gene and the exposure work together to cause disease remains to be dissected and understood in terms of signal transduction or enzymatic reactions or other cellular and physiologic processes. Clearly, such an effort goes well beyond any epidemiologic data at hand and requires integration of multiple lines of research. Elucidating the underlying causal mechanism would certainly be described as explaining the interaction of the gene and the exposure.

Relying almost exclusively on observational studies, rather than randomized experiments, to develop an understanding of disease risks, epidemiologists are naturally interested in inferring causal relationships from observational data. Consequently, they would like to use such data to infer truly causal interactions (not simply heterogeneity in statistical “effect”) without waiting for additional extensive laboratory work to have dissected mechanistic details. The idea of a causal interaction in epidemiology is that of “the interdependent operation of two or more causes to produce disease” [4]. Arguments based on counterfactuals or on sufficient causes both point to risk differences (i.e., use of the identity link as in equation (4)) as appropriate for assessing causal interactions when both the genetic and environmental factors are risk-enhancing and binary [5, 6] (when one or both factors are risk-reducing other links may be appropriate for assessing causal interactions [7]). Since the PKU example mentioned earlier departs from additivity in risk, it would be deemed a causal interaction, a conclusion confirmed by the known biological mechanism. Toxicologists use a closely related concept of causal interaction defined as departures from simple independent action [8], a concept of probabilistic independence, where the causes have completely separate modes of action. This latter concept has been adapted to studying causal

interactions in epidemiologic cohort studies through models like equations (1) and (2) using link $h(P) = \ln(1 - P)$ [7]. Some writers restrict application of the terms “synergy” or “superadditivity” and “antagonism” or “subadditivity”, mentioned earlier, to causal interactions.

The use of causal concepts and the associated specific links to determine whether observational data are consistent with a causal gene–environment interaction adds an interpretation beyond that offered by merely observing a statistical interaction with an arbitrary link. A causal inference of synergy would suggest that the gene and exposure do not act independently through separate pathways. A causal inference of antagonism might suggest competition for receptors. Thus, the ability to infer a causal interaction holds a clear advantage because it can offer clues regarding mechanisms of action. However, inference about causal interactions does have limitations. For example, in the counterfactual context, departures from risk additivity as in equation (4) imply the presence of causal interaction, whereas risk additivity does not necessarily imply its absence [5]. Unmeasured intervening variables pose problems for interpretation in that one set of observed data can be consistent with a several causal scenarios [9]. Likewise, when individuals vary in inherent susceptibility to the factors under study, causal interactions may be masked or may spuriously appear [7]. Cautious interpretation of any interaction declared causal by these approaches is warranted. In the epidemiological literature, causal interactions as discussed here are usually referred to as *biological interactions*. This terminology is justifiable in that the causal processes of health and disease are mediated through biology. Nevertheless, the terminology seems less than apt because the no-interaction or independence models employed are not derived from explicitly biological mechanisms such as kinetics or signaling; instead, they are predicated on philosophical concepts about what it means for two causes to act independently versus to interact.

Statistical Inference and Design

To this point, the presentation has centered on various definitions of gene–environment interaction; what remains is to consider how to test whether an interaction is present, how to estimate measures of

interaction, and how to design studies that are able to accomplish those tasks efficiently.

As discussed earlier, generalized linear models are often appropriate for analyzing data to address questions of gene–environment interaction. In these models, testing and estimation are usually accomplished *via* likelihood-based methods [10]. Equation (2) may be fitted by maximum likelihood techniques to provide estimates of model parameters and their standard errors. Wald tests and confidence intervals for β_{GE} , the interaction parameter, provide one route for inference about gene–environment interaction; alternatively, likelihood ratio tests and profile likelihood confidence intervals are available. Equation (2) could be elaborated to include multiple genetic or environmental predictors as well as their interactions. Higher-order interaction terms involving products of three or more variables might be considered although such higher-order terms may be difficult to interpret.

Detection of gene–environment interactions using generalized linear models requires larger sample sizes, often substantially larger, than those needed for detection of genetic or environmental effects alone – particularly when one or both factors are rare. This circumstance has prompted several proposals for increasing a study’s power to detect such interactions. One approach is to focus tests of interaction from a larger to a smaller set of parameters, a set anticipated to carry most of the effect. For example, with G and E categorical, an analogue of Tukey’s single degree of freedom test for interaction provides such a reduction in dimension [11]. A second general approach is to devise strategies that enrich a study sample for rare exposures or genotypes compared to their population prevalences. Proposals along these lines include counter matching [12], enrichment through relatives of cases [13], and various two-stage sampling plans that allow oversampling with defined probability [14–16]. A third approach involves exploiting a particular independence assumption. With a binary outcome, gene–environment interaction can be tested and a measure of interaction estimated using cases only (no control subjects) under the assumption that the genotype and the exposure occur independently in the population at large and the disease is rare [17]. This design, by exploiting the key assumption, increases the power for detecting gene–environment interaction compared to a case–control study with

the same number of cases and any number of controls; but it sacrifices the ability to test or estimate the main effects of genotype or of exposure. The assumption can also be exploited with case–control data (*see Case–Control Studies*) to increase the power for detecting gene–environment interaction as well as for the main effects [18, 19]. Inferences are subject to bias whenever the key assumption fails to hold [20], so these methods should be used circumspectly.

Confounding is an omnipresent source of bias in observational studies, and studies involving gene–environment interaction are no exception. Of particular concern are possible effects of hidden genetic population structure. Genetic epidemiologists can largely overcome this source of bias through the use of case-parents design – recruiting families with one affected offspring and two biological parents. Consider the situation where the genetic factor is a diallelic single nucleotide polymorphism (SNP). With this design, cases are stratified according to the genotypes of both parents. Assuming that gametes assort and are transmitted at random (Mendelian transmission) and that survival is nondifferential from conception to the time of study, the expected proportion of each allele at the SNP locus in each stratum is known under the null hypothesis. Deviations from those expected proportions provide evidence that the inherited SNP is associated with the disease within families. This same idea can be used to study gene–environment interaction while avoiding potential bias from population structure [21, 22]. Unbiased estimation of the interaction parameter requires independence between transmission of the SNP and exposure in offspring, conditional on parents’ genotypes, a weaker assumption than independence of G and E in the source population. A drawback to the case-parents design is that, while genetic main effects and gene–environment interaction effects can be evaluated, the main effects of the environmental factor cannot be without auxiliary information.

The design of an observational study dictates what parameters can be validly estimated and whether certain forms of interaction can be evaluated. With a *cohort study* (*see Cohort Studies*) and a binary disease outcome, absolute risks (P_{GE}) are estimable and therefore relative risks or odds ratios can be estimated too. Thus, interaction can be examined in any scale; in particular, causal interactions can be examined *via* risk differences in equation (4). For a

case–control design, odds ratios are estimable and, because the case–control design is applied to rare diseases, odds ratios are close approximations to relative risks; but absolute risks are not estimable without auxiliary information [16]. Thus, examination of causal interaction using risk differences as in equation (4) appears at first impossible. Dividing equation (4) through by P_{00} , however, yields the equivalent expression:

$$(RR_{ge} - RR_{g0}) - (RR_{0e} - 1) = 0 \quad (7)$$

Consequently, to the extent that *ORs* approximate *RRs*, causal interactions can be evaluated with case–control data even though absolute risks cannot themselves be estimated. On the other hand, the case-only design and the case-parents design do not permit the evaluation of risk differences or interaction on any scale other than multiplicative (logarithmic link), unless external data are available for estimation of main effects. Others have compared and criticized study designs and methods of data analysis for gene–environment interaction [23, 24].

The development of methods for uncovering how genes and environment work together to cause disease will remain important as long as the paradigm of gene–environment interaction continues to inform biomedical research.

References

- [1] Rothman, N., Wacholder, S., Caporaso N.E., Garcia-Closas, M., Buetow, K., Fraumeni Jr, J.F. (2001). The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens, *Biochimica et Biophysica Acta* **1471**, C1–C10.
- [2] Ottman, R. (1990). An epidemiologic approach to gene-environment interaction, *Genetic Epidemiology* **7**, 177–185.
- [3] Hlatky, M.A. & Whittemore, A.S. (1991). The importance of models in the assessment of synergy, *Journal of Clinical Epidemiology* **44**, 1287–1288.
- [4] Rothman, K., Greenland, S. & Walker, A. (1980). Concepts of interaction, *American Journal of Epidemiology* **112**, 467–470.
- [5] Rothman, K.J. & Greenland, S. (1997). *Modern Epidemiology*, 2nd Edition, Lippincott-Raven Publishers, Philadelphia, pp. 329–342.
- [6] Greenland, S. & Poole, C. (1988). Invariants and noninvariants in the concept of interdependent effects, *Scandinavian Journal of Work Environment and Health* **14**, 125–129.
- [7] Weinberg, C.R. (1986). Applicability of the simple independent action model to epidemiologic studies involving two factors and a dichotomous outcome, *American Journal of Epidemiology* **123**, 162–173.
- [8] Finney, D.J. (1971). *Probit Analysis*, 3rd Edition, Cambridge University Press, Cambridge, pp. 230–268.
- [9] Thompson, W.D. (1991). Effect modification and the limits of biological inference from epidemiologic data, *Journal of Clinical Epidemiology* **44**, 221–232.
- [10] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition, Chapman & Hall, New York, pp. 1–511.
- [11] Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U. & Wacholder, S. (2006). Powerful multilocus tests for genetic association in the presence of gene-gene and gene-environment interactions, *American Journal of Human Genetics* **79**, 1002–1016.
- [12] Andrieu, N., Goldstein, A.M., Thomas, D.C. & Langholz, B. (2001). Counter-matching in studies of gene-environment interaction: efficiency and feasibility, *American Journal of Epidemiology* **153**, 265–274.
- [13] Weinberg, C.R., Shore, D., Umbach, D.M. & Sandler, D.P. (2007). Using risk-based sampling to enrich cohorts for endpoints, genes, and exposures, *American Journal of Epidemiology* **166**, 447–455.
- [14] Breslow, N.E. & Cain, K.C. (1988). Logistic regression for two-stage case–control data, *Biometrika* **75**, 11–20.
- [15] Weinberg, C.R. & Wacholder, S. (1990). The design and analysis of case–control studies with biased sampling, *Biometrics* **46**, 963–975.
- [16] Wacholder, S. & Weinberg, C.R. (1994). Flexible maximum likelihood methods for assessing joint effects in case–control studies with complex sampling, *Biometrics* **50**, 350–357.
- [17] Piegorsch, W.W., Weinberg, C.R. & Taylor, J.A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies, *Statistics in Medicine* **13**, 153–162.
- [18] Umbach, D.M. & Weinberg, C.R. (1997). Designing and analyzing case–control studies to exploit independence of genotype and exposure, *Statistics in Medicine* **16**, 1731–1743.
- [19] Chatterjee, N. & Carroll, R.J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case–control studies, *Biometrika* **92**, 399–418.
- [20] Albert, P.S., Ratnasinghe, D., Tangrea, J. & Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interactions, *American Journal of Epidemiology* **154**, 687–693.
- [21] Schaid, D.J. (1999). Case-parents design for gene-environment interaction, *Genetic Epidemiology* **16**, 261–273.
- [22] Umbach, D.M. & Weinberg, C.R. (2000). The use of case-parent triads to study joint effects of genotype and exposure, *American Journal of Human Genetics* **66**, 251–261.

-
- [23] Weinberg, C.R. & Umbach, D.M. (2000). Choosing a retrospective design to assess joint genetic and environmental contributions to risk, *American Journal of Epidemiology* **152**, 197–203.
- [24] Clayton, D. & McKeigue, P.M. (2001). Epidemiological methods for studying genes and environmental factors in complex diseases, *Lancet* **358**, 1356–1360.

Related Articles

Linkage Analysis

DAVID M. UMBACH

Geographic Disease Risk

The availability of geographically indexed health, population, and exposure data, and advances in computing, geographic information systems, and statistical methodology, have enabled the realistic investigation of spatial variation in disease risk (*see Spatial Risk Assessment*). Each of the population, exposure, and health data may have associated exact spatial and temporal information (point data), or may be available as aggregated summaries (count data). We may distinguish between four types of studies:

1. disease mapping in which the spatial distribution of risk is summarized;
2. spatial regression in which risk summaries are modeled as a function of spatially indexed covariates;
3. cluster detection (surveillance) in which early detection of “hot spots” is attempted (*see Hotspot Geoinformatics*);
4. clustering in which the “clumpiness” of cases relative to noncases is examined.

In this article, we concentrate on the first two endeavors, since often these consider risk as a function of environmental pollutants. Much greater detail on each of these topics can be found in Elliott *et al.* [1].

Disease Mapping

Disease mapping has a long history in epidemiology [2] as part of the classic triad of person/place/time. A number of statistical reviews are available [3–6] and there are numerous examples of cancer atlases [7, 8].

Usually disease mapping is carried out with aggregate count data. We describe a model for such data. Suppose the geographical region of interest is partitioned into n areas, and Y_{ij} and N_{ij} represent the number of disease cases and the population at risk in area i , $i = 1, \dots, n$, and age-gender stratum j , $j = 1, \dots, J$. The age distribution is likely to vary by area, so a map of raw risk will reflect this distribution. Often one is interested in unexplained variability in risk and hence one controls for age and gender since these are known risk factors. For a rare disease, and

letting p_{ij} be the risk in area i and age-gender stratum j , we have

$$Y_{ij} \sim \text{Poisson}(N_{ij}p_{ij}) \quad (1)$$

For nonlarge areas in particular, the data will not be abundant enough to reliably estimate the risk in each area for each age and gender category, i.e. each p_{ij} . Hence it is usual to assume proportionality of risk, i.e., $p_{ij} = \theta_i \times q_j$, where q_j are a set of reference risks (perhaps taken from a larger region [9]), and θ_i is the *relative risk* (ratio of risks) (*see Absolute Risk Reduction*) associated with area i ; this model assumes that the effect of living in area i is constant across stratum (an assumption that should be checked in applications). Then

$$Y_i = \sum_{j=1}^J Y_{ij} \sim \text{Poisson}(E_i \theta_i) \quad (2)$$

where $E_i = \sum_{j=1}^J N_{ij}q_j$, are the *expected numbers*. The use of expected numbers is known as *indirect standardization*. The standardized mortality/morbidity ratio (SMR) is the estimate $\hat{\theta}_i = Y_i/E_i$. The variance of the estimator is $\text{var}(\text{SMR}_i) = \text{SMR}_i/E_i$, which will be large if E_i is small, which occurs when the populations are small.

To illustrate, we calculate SMRs using lung cancer mortality data over the years 1998–2002 in the state of Minnesota. These data are more fully described in [10]. The observed counts range between 14 and 3012 across counties, with median 71. Expected numbers are available, with adjustment for age and gender, and range between 16.8 and 2818. Figure 1 provides an SMR map. The SMRs range between 0.58 and 1.38, which is a relatively narrow range because the expected numbers are quite large here. The most easterly county, Cook county, has an SMR of 1.22, but an expected number of 20.5, suggesting that the high SMR may reflect sampling variability.

More reliable estimates can be obtained using random-effects models [11, 12] that use the data from all areas to provide more reliable estimates in each of the constituent areas. A popular log-linear model assumes

$$\log \theta_i = \mu + x_i \beta + U_i + V_i \quad (3)$$

where U_i and V_i are random effects with and without spatial structure, x_i are covariates associated with

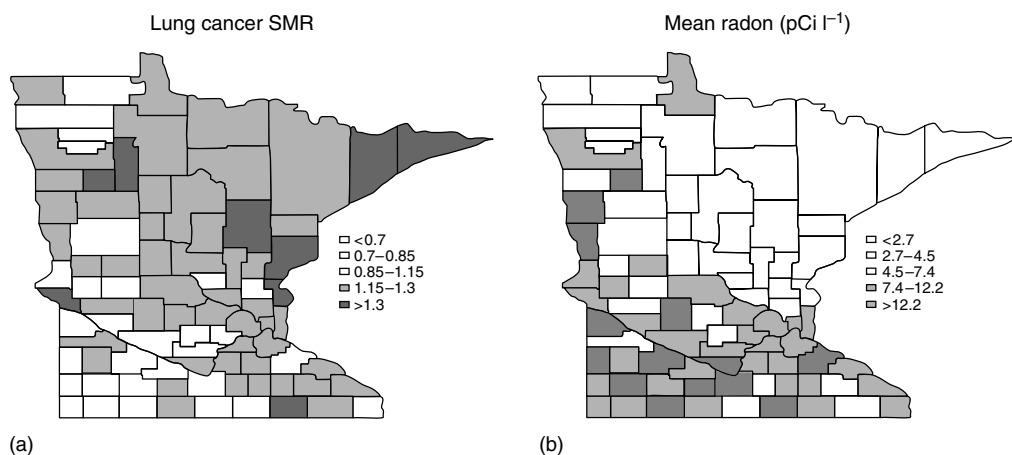


Figure 1 (a) Lung cancer standardized mortality ratios (SMRs) and (b) average radon in 87 counties of Minnesota

area i , and β is the vector of log relative risks associated with these covariates. The V_i terms give global smoothing to the overall mean, the U_i allow local smoothing so that “close-by” areas are likely to be similar, and $x_i\beta$ models similarity between areas with similar covariate values. Details of specific spatial models, including neighborhood schemes, and computational strategies are available elsewhere [6, 12]. Once reliable estimates of the relative risk are available, these may be used for public health purposes (for example, to decide on which areas to promote screening campaigns), or they may be compared with maps of environmental exposures in order to investigate the association between risk and exposure.

Spatial Regression

The formal comparison with risk and exposure is carried out using so-called ecological correlation studies, and numerous such studies have been reported [13]. Exposures may be directly measured in air, water, or soil, or be indirect surrogates such as latitude as a surrogate for exposure to sunlight, distance from a point source of risk such as an incinerator [14], or a foundry [15], or a line source such as a road.

With aggregate count data, model (3) may also be used in the context of spatial regression, though now β is of primary interest and one must be aware of ecological bias that occurs when individual-level inference is attempted from aggregated data and arises

due to within-area variability in exposures and confounders [16–18]. We illustrate by returning to the Minnesota example but now assume that the aim is to estimate the association between risk of lung cancer mortality and residential radon. Radon (*see Radon*) is a naturally occurring radioactive gas present in rocks and soil. Extensive epidemiological studies on highly exposed groups, such as underground miners, consistently indicate a substantially increased risk of lung cancer at high concentrations of radon [19]. However, extrapolation to lower doses, typical of those found in residential homes, is more controversial. Many studies that address residential radon have been ecological in design, but their usefulness is debated, in part owing to the problems of interpretation in the presence of ecological bias [20]. As a result, conclusions from such studies are conflicting, with ecological studies often displaying a negative association between radon and lung cancer. We utilize radon measurements in individual homes in Minnesota collected by the environmental protection agency (EPA) and Minnesota Department of Health in 1987. The sample sizes associated with the radon measurements, by county, are very small, ranging between 1 and 122 (median 5), with two counties having 0 individual radon measurements. Figure 1(b) displays average radon by county. Comparison with the SMRs in Figure 1(a) seems to indicate a negative association (there is a northeast to southwest decreasing trend in the SMRs, and the opposite in radon). Figure 2 shows the SMRs versus average radon, and confirms a clear negative association, indicating a protective

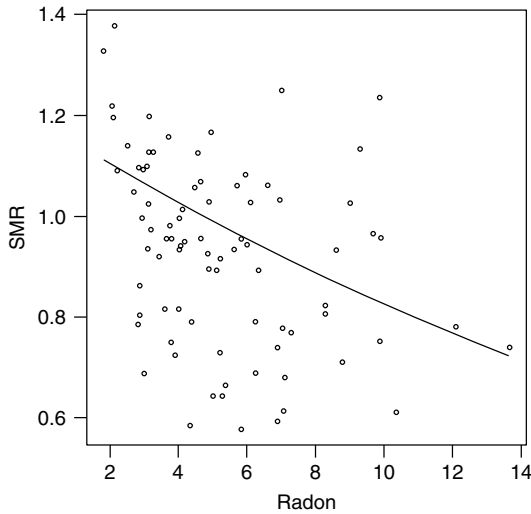


Figure 2 SMRs versus mean radon levels (pCi l^{-1}), along with fitted log-linear model

effect. We assume the model

$$E[Y_i] = E_i \exp(\mu + \beta x_i), \text{var}(Y_i) = \kappa \times E[Y_i] \quad (4)$$

where x_i is the average radon in area i , and κ allows for overdispersion. We fit this model using quasi likelihood [21], and obtain $\hat{\beta} = -0.036$ (with a standard error of 0.0090) to give a highly significant area-level relative risk of $\exp(-0.036) = 0.96$, indicating a 4% drop in risk for every 1 unit increase in radon. For these data, $\hat{\kappa} = 2.8$, so there is a large degree of excess-Poisson variability (note that this model does not account for residual dependence in the counts). The fitted curve in Figure 2 corresponds to the fit of this model. For a more thorough analysis of the radon example that addresses ecological bias, see [10]. Here we briefly mention some problems with the above analysis. Smoking is a confounder that we have not controlled for, and there is within-area variability in radon exposure that can lead to pure specification bias that arises under the aggregation of a nonlinear individual-level model. The only way to overcome ecological bias is to supplement the ecological information with individual-level data, and this is the subject of much current work [18, 22–24]. In conclusion, associations estimated at the aggregate level should be viewed with caution owing to the potential for ecological bias, but ecological data can offer clues to etiology, and add to the totality of evidence of a given relationship (see **Ecological Risk Assessment**).

In the above, we have considered spatial regression with count data. With point data a similar approach is available with a Bernoulli likelihood and a logistic, rather than log, link function. There are no problems of ecological bias with individual-level data, but the spatial aspect should still be considered. This may be less of a problem if individual-level information on confounders is gathered, as this will reduce residual dependence, since it is unmeasured variables that contribute to this dependence (along with errors in the data that have spatial structure, and other sources of model misspecification [25]). It is usual for aggregate data to provide (hypothetically) complete geographical coverage, but with point data one must carefully examine the data collection scheme since it may have spatial structure; in particular cluster sampling may have been carried out. In these cases, spatial dependence reflects not only inherent variability but also that arising from the sampling scheme.

We finally note the potential for “confounding by location”. Often risk and exposure surfaces have local and global spatial structure, so when one progresses from a model with no modeling of spatial dependence to one with spatial effects, one may see large changes in the regression coefficients of interest. This occurs because the initial coefficient may reflect not only the exposure of interest but also other variables associated with risk that are correlated with the exposure surface. This is a complex issue [26, 27] since overadjustment for spatial location is also possible. The ideal scenario is when one has good data on confounders, since in this case any adjustment for spatial dependence should not change estimated associations too drastically.

Conclusions

A major weakness of geographical analyses investigating the association between environmental pollutants and risk is the inaccuracy of the exposure assessment. Modeling of concentration surface has become more common (for a state-of-the-art treatment, see Le and Zidek [28]), but care is required in the use of such methods since inaccuracies in the modeled surface (which are often based on sparse monitor data) may introduce greater bias than that present in simple methods that, for example, use the nearest monitor for exposure assessment [29].

Acknowledgment

This article was written with support from grant R01 CA095994 from the National Institute of Health.

References

- [1] Elliott, P., Wakefield, J.C., Best, N.G. & Briggs, D.J. (2000). *Spatial Epidemiology: Methods and Applications*, Oxford University Press, Oxford.
- [2] Walter, S.D. (2000). Disease mapping: a historical perspective, in *Spatial Epidemiology: Methods and Applications*, P. Elliott, J.C. Wakefield, N.G. Best & D. Briggs, eds, Oxford University Press, pp. 223–239.
- [3] Smans, M. & Esteve, J. (1992). Practical approaches to disease mapping, in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English & R. Stern, eds, Oxford University Press, Oxford, pp. 141–150.
- [4] Clayton, D.G. & Bernardinelli, L. (1992). Bayesian methods for mapping disease risk, in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, P. Elliott, J. Cuzick, D. English & R. Stern, eds, Oxford University Press, Oxford, pp. 205–220.
- [5] Mollié, A. (1996). Bayesian mapping of disease, in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds, Chapman & Hall, New York, pp. 359–379.
- [6] Wakefield, J.C., Best, N.G. & Waller, L.A. (2000). Bayesian approaches to disease mapping, in *Spatial Epidemiology: Methods and Applications*, P. Elliott, J.C. Wakefield, N.G. Best & D. Briggs, eds, Oxford University Press, Oxford, pp. 104–127.
- [7] Kemp, I., Boyle, P., Smans, M. & Muir, C. (1985). *Atlas of Cancer in Scotland, 1975–1980: Incidence and Epidemiologic Perspective*, IARC Scientific Publication 72, International Agency for Research on Cancer, Lyon.
- [8] Devesa, S.S., Grauman, D.J., Blot, W.J., Hoover, R.N. & Fraumeni, J.F. (1999). *Atlas of Cancer Mortality in the United States 1950–94*, NIH Publications No. 99–4564, National Institutes of Health.
- [9] Wakefield, J.C. (2006). Disease mapping and spatial regression with count data, *Biostatistics* **8**, 158–183.
- [10] Salway, R. & Wakefield, J. (2008). A hybrid model for reducing ecological bias, *Biostatistics* **9**, 1–17.
- [11] Clayton, D.G. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics* **43**, 671–682.
- [12] Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics, *Annals of the Institute of Statistics and Mathematics* **43**, 1–59.
- [13] Boffetta, P. & Nyberg, F. (2003). Contribution of environmental factors to cancer risk, *British Medical Journal* **68**, 71–94.
- [14] Elliott, P., Shaddick, G., Kleinschmidt, I., Jolley, D., Walls, P., Beresford, J. & Grundy, C. (1996). Cancer incidence near municipal solid waste incinerators in Great Britain, *British Journal of Cancer* **73**, 702–707.
- [15] Lawson, A. & Williams, F. (1994). Armadale: a case-study in environmental epidemiology, *Journal of the Royal Statistical Society, Series A* **157**, 285–298.
- [16] Greenland, S. (1992). Divergent biases in ecologic and individual level studies, *Statistics in Medicine* **11**, 1209–1223.
- [17] Richardson, S. & Montfort, C. (2000). Ecological correlation studies, in *Spatial Epidemiology: Methods and Applications*, P. Elliott, J.C. Wakefield, N.G. Best & D. Briggs, eds, Oxford University Press, Oxford, pp. 205–220.
- [18] Wakefield, J.C. (2004). Ecological inference for 2×2 tables (with discussion), *Journal of the Royal Statistical Society, Series A* **167**, 385–445.
- [19] National Academy of Sciences (1999). *Health Effects of Exposure to Radon: BEIR VI*, National Academy press, Washington, DC.
- [20] Stidley, C. & Samet, J. (1994). Assessment of ecologic regression in the study of lung cancer and radon, *American Journal of Epidemiology* **139**, 312–322.
- [21] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition, Chapman & Hall, London.
- [22] Jackson, C.H., Best, N.G. & Richardson, S. (2006). Improving ecological inference using individual-level data, *Statistics in Medicine* **25**, 2136–2159.
- [23] Glynn, A., Wakefield, J., Handcock, M. & Richardson, T. (2007). Alleviating linear ecological bias and optimal design with subsample data, *Journal of the Royal Statistical Society, Series A* (online).
- [24] Haneuse, S. & Wakefield, J. (2006). Hierarchical models for combining ecological and case-control data, *Biometrics* **63**, 128–136.
- [25] Wakefield, J. & Elliott, P. (1999). Issues in the statistical analysis of small area health data, *Statistics in Medicine* **18**, 2377–2399.
- [26] Clayton, D., Bernardinelli, L. & Montomoli, C. (1993). Spatial correlation in ecological analysis, *International Journal of Epidemiology* **22**, 1193–1202.
- [27] Wakefield, J.C. (2003). Sensitivity analyses for ecological regression, *Biometrics* **59**, 9–17.
- [28] Le, N.D. & Zidek, J.V. (2006). *Statistical Analysis of Environmental Space-Time Processes*, Springer, New York.
- [29] Wakefield, J. & Shaddick, G. (2006). Health-exposure modelling and the ecological fallacy, *Biostatistics* **7**, 438–455.

Related Articles

Environmental Health Risk

Environmental Monitoring

Statistics for Environmental Toxicity

JON WAKEFIELD

Global Warming

The global average annual temperature of the Earth's surface has increased by as much as 0.7°C since the late nineteenth century, with significant warming in the last 50 years. The scientific consensus, best represented in this matter by the Intergovernmental Panel on Climate Change (IPCC), is that a significant portion of this increase, and most of the warming in the last 50 years, is very likely (i.e., with probability greater than 0.9 according to the IPCC definition of the phrase) attributable to human activities – most notably the burning of fossil fuels – leading to increased concentrations of greenhouse gases (GHGs) in the atmosphere. Further, the consensus is that a larger future warming is in store in the absence of significant curbing of GHG emissions, with 2°C additional warming being generally accepted as a lower bound estimate of the expected global warming at the end of the twenty-first century under so-called business as usual emission scenarios.

Of course, in the absence of our atmosphere, and its natural greenhouse effect, the temperature of Earth would soar during the day and fall dramatically at night. The entire amount of solar energy incoming on the surface of our planet would be immediately reflected back. The presence of GHGs in our atmosphere (CO_2 , methane, water vapor, and ozone being the main constituents), able to “trap” part of the radiation that the Earth's surface reflects back, keeps our planet inhabitable. However, human activities are dumping enormous quantities of GHGs in the atmosphere every day, mainly through the burning of fossil fuels. There is no doubt that the increase in the CO_2 concentrations in the atmosphere comes from human activities, since the origin of its molecules can be traced exactly, owing to a difference in the isotopes of naturally produced rather than man-produced gases. The higher concentrations of these gases make the planet increasingly warmer. There are positive feedbacks superimposed on this straightforward mechanism (e.g., the melting of ice caps, because of the warmer atmospheric temperature, uncovers water, which absorbs more incoming radiation than the reflective surface of ice, further warming the sea surface temperatures of northern regions, and melting more ice). There are countereffects produced in part by the same industrial activities responsible for the production of GHGs, i.e., aerosols (small

particles suspended in the atmosphere, which actually consist in large measure of natural dusts, produced by volcanic eruptions and dust storms), which act as screens, keeping incoming radiation from penetrating the atmosphere. Aerosols, however, are short-lived agents, and not as well mixed as GHGs, making their cooling effect more transitory, and geographically less homogeneous (and thus more difficult to predict). Complicating the detection of warming and human influences is the natural variability of the system, owing to low-frequency signals that impact temperatures even in their global average, like the El Niño Southern Oscillation and the North Atlantic Oscillation. The solar radiation also goes through natural phases, creating warmer and cooler multiyear phases, which have however been quantified as insufficient to explain the warming observed in the last decades. A source of uncertainty in the exact quantification of future global warming is the effect of clouds. Warmer climates mean larger amounts of water vapor in the air, so it is to be expected that changes in the formation of clouds are in store as well, with GHGs increasing. The formation of clouds is, however, an extremely complex process, which is impossible to represent explicitly in the current generation of climate models because of limited spatial resolution, and may have contrasting effects on the climate system: high clouds may screen incoming radiation with a cooling effect and low clouds may trap a larger fraction of the reflected radiation, keeping the surface of the Earth warmer. Even in the presence of all these complicating factors, analysis of the observed records and comparison with idealized experiments has quantified evidence that the warming observed so far is inconsistent with estimates of natural variability, and natural change in forcings owing to solar or other low-frequency modes of variability. The detection of an “unnatural” warming and its attribution to human causes is being performed for increasingly more variables, and phenomena that are increasingly finer in scale. For example, it is a strong result for global average temperature [1], but it has also been proved for regional temperature changes [2], and very recently for sea surface temperatures in the cyclogenesis regions of the Atlantic and the Pacific [3].

Global average increases in temperature are just an aggregation of heterogeneous regional climatic changes that are very likely to affect societies and ecosystems in adverse rather than beneficial ways. Climate models and scientific understanding of

climate processes project more frequent and intense extreme events and do not rule out the possibility of exceeding systemic thresholds for climate variables causing abrupt changes, with potentially much more dangerous consequences than simple gradual shifts in the makeup of regional climatologies. Associated climatic changes and economic, biological, and health-related impacts are then to be expected, ranging from changes in agricultural yields and forced abandonment and geographical shifts of crops to disruption of ecosystems and from depletion of water resources to easier spread of vector-borne diseases.

If the qualitative nature of the problem represented by global warming is generally agreed upon by a large share of the scientific, policy-making, and stakeholders' communities, a quantification of the risks associated with the above story line is far from straightforward.

There exist considerable uncertainties, both in terms of limits to predicting future changes in the drivers of GHG emissions (population, economic growth, etc.) and incomplete scientific understanding of the response of the climate system to those emissions, and in terms of intrinsic randomness or unpredictability of the processes involved. From the size of the changes to be expected to their timing and from the significance of their impacts to the costs of adaptation and mitigation strategies, every link of the chain from GHG emissions to impacts to the effects of response strategies is affected by uncertainties. The quantification of these uncertainties using distributions of probability cannot often rely on engineering-like assessments, i.e., cannot draw on an exactly replicable experimental setting or long records of observations from stationary processes. Rather, subjective choices based on expert knowledge are often necessary in the statistical modeling of the random quantities at stake, opening the way to debates and to the formation of an additional layer of uncertainty, a meta-uncertainty, that is, over the distributions of probability so derived.

In this article we will attempt to characterize the main sources and characteristics of these uncertainties, together with the most rigorous efforts currently documented in the peer reviewed literature to quantify them and incorporate them into risk analyses.

We describe in the section titled "Uncertainties in Climate Modeling" uncertainties in modeling of the climate system and their characterization and quantification through single-model and ensemble

experiments. In the section titled "Uncertainties from Boundary Conditions: Future Scenarios of Emissions" we address uncertainties in the development of future emission scenarios, recognized as the principal driving force of climatic change. The section titled "Approaches to Risk Assessment and Risk Management" describes methods and results from state-of-the-art modeling strategies adopted by impacts science and policy analysis in the arena of risk assessment and risk management. On account of the availability of a number of comprehensive reviews of climate change science, impacts, adaptation, and mitigation studies, we choose to limit the list of references to the works that we explicitly cite in this article's discussion. In the section titled "Pointers to Sources of Information in Climate Change Science, Impacts, Mitigation, and Adaptation", we list the major sources of information available for overviews or more in-depth study of climate change.

Pointers to Sources of Information in Climate Change Science, Impacts, Mitigation, and Adaptation

Every 5–6 years since 1990, the IPCC, a United Nations (UN)/World Meteorological Organization (WMO)-sponsored international organization, has published an assessment report, compiling a critical summary of the state of research on the science of climate change, its impacts and the efforts on the front of adaptation strategies, plus economic-driven studies of mitigation policies.

The last published report at the time of this writing is the third, available in its entirety, together with the two previous reports, at <http://www.ipcc.ch>. The next report will be published in the first months of 2007, and will be available through the same website.

Within the report, especially relevant to the discussion in this article, are the chapters by Working Group 1 in "The physical science basis" on "Understanding and Attributing Climate Change" (Hegerl and Zwiers coordinating lead authors), "Global Climate Projections" (Meehl and Stocker), and "Regional Climate Projections" (Christensen and Hewitson) and the literature they cite.

Amongst the chapters by Working Group 2, in "Impacts, Adaptation and Vulnerability", we have drawn on the discussion in "New assessment methodologies and the characterization of

future conditions” (Carter, Lu, and Jones) and in “Assessing key vulnerabilities and the risk from climate change” (Patwardhan, Semenov, and Schneider) and the literature assessed therein. A Special Report on Emission Scenarios (SRES) is available through the IPCC website, and addresses the construction of the GHG emission scenarios on the basis of story lines of future economic and social world development. The United States National Assessment report was produced in 2000, sponsored by the United States Global Change Research Program, a government entity. It is available at <http://www.usgcrp.gov/usgcrp/nacc/default.htm>.

In October 2006, a review of climate change science and impacts was compiled for the UK government by Sir Nicholas Stern. Setting it apart from the IPCC reports and other national assessment reports is the economic focus of this synthesis report, where explicit figures of projected loss in global GDP and detailed scenarios of specific impacts and their costs are laid out, and prescriptions like environmental taxes are put forward. The entire report is available at http://www.hm-treasury.gov.uk/independent_reviews/stern_review_economics_climate_change/sternreview_index.cfm.

On a lighter note, “The rough guide to climate change” has been recently published. In this book, Robert Henson gives a complete but very accessible picture of the symptoms, the science, the debates, and the solutions of this issue [4]. In all these review texts, extensive bibliographies are available.

Uncertainties in Climate Modeling

Similar to what happens for day-to-day weather forecasts, projections of future climate are based on numerical model experiments that resolve or parameterize the processes relevant for the forecast time scale of interest, in our case, decades to centuries.

Weather and climate models are based on the representation of the laws of fluid dynamics that govern weather processes. They approximate these laws at a finite set of points discretizing the three physical dimensions of Earth and at a sequence of finite time steps, usually at increments on the order of a few minutes, for length of simulations ranging from hours (for weather models) to centuries (for climate models).

The relevant sources of uncertainty in constructing and applying climate models are classifiable into boundary condition, parameter, and structural

uncertainties. Initial conditions are often dismissed as a source of uncertainty in climate projections since the results from model runs are usually averaged over decades and over ensemble members (i.e., over comparable simulations from the same model under different initial conditions, or from different models).

An uncertainty from boundary conditions arises because the model experiments, which are otherwise self-contained, rely on inputs that represent external influences, in the form of estimates of solar and other natural forcings or, in an atmosphere-only models, sea surface temperatures, and sea ice cover. For our discussion, the most relevant external input, and one that is fraught with uncertainties, is the change in atmospheric concentrations of GHGs over the course of the future simulated time, which we will discuss in detail in the section titled “Uncertainties from Boundary Conditions: Future Scenarios of Emissions”.

On account of limited computing power, climate processes and their interactions can be represented in computer models only up to a certain space and time scale (the two being naturally linked to each other). The current state-of-the-art atmosphere-ocean coupled climate models have spatial resolutions on the order of a 100 km in the horizontal dimension, and about 1000 m in the vertical dimension. The effects on these scales of processes living at finer scales must be parameterized with bulk formulas that depend on large-scale variables available in the model. One such parameter is, for example, the percentage of cloud cover within a model’s grid cell. This percentage depends on a number of other quantities and processes that are simulated explicitly or implicitly by the model, and is responsible for feedbacks on the large-scale quantities. However, the formation of clouds is an extremely complex process that cannot be simply measured empirically as a function of other large-scale observable (and/or representable in a model) quantities. Thus the choice of the parameters controlling the “formation of clouds” within the model cells has to be dictated by a mixture of expert knowledge and goodness of fit arguments, based on actual model runs and a number of diagnostic measures, often conflicting with one another. Of late, parameter uncertainties have started to be systematically explored and quantified by so-called perturbed physics ensembles (PPEs), sets of simulations with a single model but different choices for various parameters. In fact, many studies that have attempted to quantify climate change in a probabilistic sense have

taken this approach. However, the PPE approach is hampered by the limited interpretability of many of the parameterization choices in complex climate models. Only for a limited number of the parameters, in fact, can either empirical evidence or scientific understanding dictate sensible ranges of variation for the sensitivity analysis, and often the tweaking of parameters takes place without considering dependencies among them.

The third type of uncertainty is introduced at a fundamental level by the choices of model design and development. From basic structural features of a particular model, like type of grid, resolution, and numerical truncation, to physical aspects, like the choice of representing land-use change or not and of making it interactive or not, there are sources of variation across model output that cannot be captured by parameter perturbations, and are thought to introduce a degree of intermodel variability of larger magnitude than any PPE experiment is able to explore. The analysis of multimodel ensembles has sought to tackle this dimension of uncertainty by taking advantage of so-called ensembles of opportunity, i.e., collecting the results of standard experiments performed by different models under comparable setup.

Even if the future projections of models cannot be validated for decades to come, there are many measurements available, of past and current records, which permit the validation of the fundamental abilities of climate models to simulate the Earth's climate. Observations of hundreds of quantities on the surface, in the vertical dimension of the atmosphere, and from the depth of the oceans are used to verify the ability of these models to replicate the characteristics of current climate (from the beginning of the instrumental records), not only in mean values but also in terms of simulated variabilities at a whole spectrum of frequencies. Paleoclimate proxies (tree rings, ice cores, corals, and sediments) are used to validate the century-scale oscillations of the simulations. Experiments that simulate volcanic eruptions test the "reaction time" of the simulated system, since observations track the sudden dip in globally average temperatures and the subsequent recovery, occasioned by major volcanic eruption in the tropical zones. Long "unforced" simulations, worth hundreds of years' time are run to test the self-consistency and stationarity of so-called control climate, i.e., time series of variables produced by simply allowing the internal variability of the system play out, without

natural (e.g., solar) or anthropogenic (e.g., changes in GHGs concentrations) forcings imposed. Thus, climate model performance can be validated under many metrics and diagnostics, and state-of-the-art models are expected to pass many tests, even if it is true that no climate model performs perfectly under the whole spectrum of measures, and different models have different strengths and weaknesses.

Perturbed Physics Ensemble Experiments and Their Analysis

Given enough computational capacity, the PPE approach is a fairly straightforward enterprise: parameters are perturbed within sensible ranges and the results of model runs, given a particular setting, are validated by observations. The experiments that deliver a better performance with respect to diagnostics of choice (several observed and modeled quantities are often tested against each other) are weighted more than those that generated less realistic results. Future climate parameters of interest are collected from all the perturbation experiments and their empirical probability distributions are derived and weighted on the basis of the results of the validation step. Recent studies have used these results to produce probability density functions (PDFs) of quantities like equilibrium climate sensitivity (defined as the amount of long-term global average warming at the time of associated with a doubling of CO₂ concentrations after the climate system reaches a new equilibrium, under the hypothesis that the GHG concentrations are maintained at that fixed level); global average warming during a transient simulation (i.e., without the requirement that the climate system reach a new equilibrium, in the presence of continuing external forcing); the present-day magnitude of the aerosol forcing, and various other projections. We should note here that the computational cost of PPEs has made it unfeasible so far to run large perturbation experiments of fully coupled climate models, which are the only direct means to derive future projections of climate change at regional scales. Either simple models [5], intermediate complexity models [6], or atmosphere-only models coupled to a simplified ocean module [7, 8] have been used. Thus, in most cases, the uncertainties quantified through PPE studies pertain to global average quantities or equilibrium rather than transient quantities, and are not straightforwardly usable in most risk assessment analyses for the purpose of policy making, which

demand regional quantities as input. An additional statistical step needs to be taken in order to link projections from PPEs to transient future changes at regional scales. This has been worked out, for example, through the assumption that the geographical patterns of change, once standardized per unit of global warming, are stable and constant across time. This method, called *simple pattern scaling*, has been tested and validated, and appears to produce accurate projections for average temperature changes and – in a less robust way, at least for increasingly finer regional changes – for average precipitation, but is less defensible if projections are sought for other climate variables (e.g., winds, sea level rise), and quantities other than averages (e.g., extremes). Unfortunately, these are often the most relevant kind of projections for risk assessment.

Work to link results from PPE experiments to runs by fully coupled General Circulation Models (GCMs) continues [9]. This area is, in fact, making use of interesting statistical modeling, for example, in order to construct “emulators” able to represent climate model output over a large space of parameters’ ranges, in the absence of actual model experiments filling in the parameter space in its entirety [10]. We need to point out that in order to explore the full range of uncertainties and for making sure that the results of this type of analysis are not dependent on the particular model used in the PPE experiments, an ensemble of different models should be considered. Unfortunately, given the computational cost of PPE approaches, only the Hadley center model, from the UK met office, has performed systematic experiments of this kind [11]. Other such efforts should be encouraged, since the resolution of the major model uncertainties has to rely on both “between model” and “within model” ensemble analyses.

Structural Uncertainty Exploration and the Analysis of Multimodel Ensembles

Results from different fields of application (e.g., public health, agriculture, and hydrology) have shown that combining models generally increases the skill, reliability, and consistency of model forecasts. In weather and climate forecast studies (especially seasonal forecasts for the latter category, where verification is more straightforward) multimodel ensembles are generally found to be better than single-model forecasts, particularly if an aggregated performance measure over many diagnostics is considered.

Multimodel ensembles are defined in these studies as a set of model simulations from structurally different models, where one or more initial condition ensembles are available from each model (if more initial conditions for each model are available the experiments are often said to constitute a super ensemble). On account of several international efforts, most importantly the IPCC-driven activities, in preparation for its multiannual assessment reports, multimodel ensembles are becoming increasingly available in the climate change analysis field as well. These joint experiments are usually performed within a systematic framework in support of model diagnosis, validation, and intercomparison. Table 1 lists the models that have contributed to the latest IPCC-AR4 and the corresponding modeling centers where the models are developed.

There are obviously different ways to combine models, and controversies exist regarding which is best, even in applications where skill or performance can be calculated by comparing model predictions to observations. The problem is rendered more challenging for climate projections, where no simple verification is attainable.

Multimodel projections for long-term climate change were used in reports of the IPCC, in the form of unweighted multimodel means, and were thought of as better best-guess estimates than the projections of single models. The underlying assumption was and is, of course, that the independent development of models makes their structural errors random and likely to cancel one another in an average. In this context, the only measure of uncertainty associated with the multimodel means was an intermodel deviation measure, consisting of a simple standard deviation measure often calculated on the basis of as few as seven to nine models.

Probabilistic projections based on multimodel ensembles are rather new in the literature and are based on a variety of statistical methods. In 1997, Raisanen [12] for the first time advocated the need for a quantitative model comparison and the importance of intermodel agreement in assigning confidence to the forecasts of different models. A few years later, in 2001, Raisanen and Palmer [13] published the first article to propose a probabilistic view of climate change on the basis of multimodel experiments as an alternative to the simple estimate of the average of the models. On the basis of 17 models, probabilities of threshold events such as “the warming at the

Table 1 Models and their originating groups

BCCR-BCM2.0	Bjerknes Centre for Climate Research	Norway
BCC-CM1	Beijing Climate Center	China
CCSM3	National Center for Atmospheric Research	USA
CGCM3	Canadian Centre for Climate Modeling and Analysis	Canada
CNRM-CM3	Meteo-France/Centre National de Recherches Météorologiques	France
CSIRO	CSIRO Atmospheric Research	Australia
ECHAM/MPI	Max Planck Institute for Meteorology	Germany
ECHO-G	Meteorological Institute of the University of Bonn/Meteorological Research Institute of KMA/Model and Data Group	Germany/Korea
FGOALS	LASG/Institute of Atmospheric Physics	China
GFDL	NOAA/Geophysical Fluid Dynamics Laboratory	USA
GISS	NASA/Goddard Institute for Space Studies	USA
INM-CM3	Institute for Numerical Mathematics	Russia
IPSL-CM4	Institut Pierre Simon Laplace	France
MIROC	Center for Climate System Research (University of Tokyo), National Institute for Environmental Studies	Japan
MRI-CGCM	Meteorological Research Institute	Japan
PCM	National Center for Atmospheric Research/Department of Energy	USA
UKMO-HadCM3	Hadley Center for Climate Prediction and Research/Met Office	UK
UKMO-HadGEM1	Hadley Center for Climate Prediction and Research/Met Office	UK

time of doubled CO_2 will be greater than $1^\circ\text{C}''$ were computed as the fraction of models that simulated such event. These probabilities were computed at the grid-point level and also averaged over the entire model grid to obtain global mean probabilities. The authors used cross-validation to test the skill of the forecast so derived, for many different events and forecast periods. They also showed how probabilistic information may be used in a decision theory framework, where a cost–benefit analysis of initiating some action may have different outcomes depending on the probability distribution of the uncertain event, from which protection is sought. Naturally, in their final discussion, the authors highlighted the importance of adopting a probabilistic framework in climate change projections, and wished it became an “established part of the analysis of ensemble integrations of future climate”. The same authors applied the same procedure to forecasts of extreme events [14].

In 2003, Giorgi and Mearns published a significantly different approach to the computation of threshold events, as a follow-up to their method

described in a first paper the year before [15, 16], by proposing the first weighted average of models, based on model performance and model agreement. Their reliability ensemble average (REA) approach rewards models with small bias and projections that agree with the ensemble “consensus”, and discount models that perform poorly in replicating observed climate and that appear as outliers with respect to the ensemble’s central tendency. This is achieved by defining model-specific weights as

$$R_i = \left\{ \left[\min \left\{ 1, \frac{\varepsilon_T}{\text{abs}(B_{T,i})} \right\} \right]^m \left[\min \left\{ 1, \frac{\varepsilon_T}{\text{abs}(D_{T,i})} \right\} \right]^n \right\}^{[1/(m \times n)]} \quad (1)$$

where the subscript i indexes the different models of the ensemble. In equation (1) the term ε_T is a measure of natural variability of, say, average seasonal temperature in a given region; $B_{T,i}$ represents the bias (the discrepancy between a model simulation and observations) in the simulation of

temperature in the region by model i , for a period for which observations are available and that constitutes the baseline with respect to which “change” is defined. The term $D_{T,i}$ represents a measure of distance of the i th projection of temperature change from the consensus estimate. The weights R_i are used to construct a weighted average of model projections for the future climate mean, with standard error also computed from the weights. Using the form (1) the weights cannot be computed upfront but are estimated by an iterative procedure, starting with a consensus estimate that is a simple average and iterating to convergence. It is easy to show that if the form of $D_{T,i}$ is an Lp norm the optimal solution to the estimation problem is $\Delta\tilde{T}$ that minimizes $\sum_i C_i |\Delta T_i - \Delta\tilde{T}|^q$, where C_i is the part of the weight that does not depend on $D_{T,i}$, and q is the function of p . In [15, 16] the authors chose m and n such that q is equal to one, making the result of the estimation amount to the choice of a robust measure of location, the weighted median of the observations. Interestingly, this fact went unnoticed by the authors of the REA approach and was recognized in [17]. Besides the innovative step of considering two criteria in the formal combination of the ensemble projections, the focus on regional analyses made the methods more directly relevant for impact studies and decision making. However, the authors stopped short of applying probability distributions to actual decision-making contexts.

The REA approach motivated the work in [18] and [19]. A Bayesian statistical model (*see Bayesian Statistics in Quantitative Risk Assessment*) was proposed as the basis for any final estimate of climate change probabilities. The Bayesian analysis treats the unknown quantities of interest (current and future climate signals, whose estimated difference is equivalent to $\Delta\tilde{T}$ in [15, 16], and model reliabilities that serve the same function as the weights R_i , above) as random variables with appropriate prior distributions. Assumptions on the statistical distribution of the data (models’ output and observations) as a function of the unknown parameters (the likelihood) are combined through *Bayes’ theorem* (*see Bayes’ Theorem and Updating of Belief*) to derive posterior distributions of all the uncertain quantities of interests, including the climate change signal. The likelihood model assumes that the current and future temperature simulations by model i , respectively, X_i and Y_i , are centered on the true climate signals, μ (for current

average temperature) and ν (for future average temperature) with a precision that is model-specific, as in:

$$X_i \sim N(\mu; \lambda_i)$$

$$Y_i|X_i \sim N(\nu + \beta(X_i - \mu); \theta\lambda_i) \quad (2)$$

The Gaussian assumption is motivated by the fact that X ’s and Y ’s are large regional and seasonal averages, and it is believed that the models will aim at approximating the true climate with errors that are normally distributed, each model with its own specific degree of precision. In equation (2) the linear term in the mean of Y_i serves the function of allowing for correlation between current and future simulated temperature by the same model. The parameter θ introduces a deflation/inflation factor in the precision of the models when comparing the accuracy of current-day simulations and future simulations. The λ_i terms represent the model-specific precisions. The unknown parameters λ_i , μ , ν , θ , and β are all given improper or highly dispersed prior distributions to represent a state of prior ignorance. The distributional form of the priors is conjugate to the likelihood model and the joint posterior distributions for all the parameters can be easily derived by Markov chain Monte carlo simulation and its marginal for the difference $\nu - \mu$ can be straightforwardly evaluated to represent the best guess at future temperature change and its uncertainty. The simple model in equation (2) is applied to single subcontinental size regional averages of temperature or precipitation. A more complex model introduced region-specific parameters in the mean and variance structure of the Gaussian likelihood, to treat more than one region at a time, thus borrowing information across space in the estimate of the model-specific and common parameters [20].

The papers [18, 19] were the first to propose a formal statistical treatment of the data available that produced regional probabilistic projections of temperature and precipitation change. The regional perspective was also adopted by Greene *et al.* [21]. In this paper, formal statistical models are adapted from the seasonal forecasting methodology to long-term future projections, basically fitting a linear model between observed (as regressand) and modeled temperatures (as regressors) from the ensemble of models. The probabilistic nature of the results is here too a natural product of adopting a Bayesian approach, with multivariate Gaussian priors for the coefficients of the regression. After determining the

posterior for the coefficients, the same are applied to the future temperatures simulated by the same models, and PDFs of temperature change at large regional scales are derived.

The methods described so far all used large regional averages of model output, producing probabilistic projections at the subcontinental scale.

Furrer *et al.* [22] attacked the statistical modeling of GCM output at the grid-point scale, using spatial statistics techniques (*see Spatial Risk Assessment*) and thus producing probabilistic projections able to jointly quantify the uncertainty over the global grid. They do so by modeling geographical features of the fields of temperature and precipitation change at varying degrees of smoothness, accounting for the spatial correlation between locations as a function of their distance. They use Gaussian random fields to model the residual terms after projecting larger scale features over a set of basic functions comprising spherical harmonics, land-ocean contrasts, and observed mean fields.

Perhaps better suited to actual impacts assessment and thus a formal risk analysis, some studies have forgone the general approach and focused on projections for specific areas and impact type. For example, Benestad [23] used statistical downscaling of a multimodel ensemble to derive probabilistic scenarios at fine resolution over northern Europe. Luo *et al.* [24] analyzed impacts over Australian wheat production by using Monte Carlo techniques to explore the range of uncertainties as quantified by the multimodel ensemble, besides addressing other sources of uncertainty like future emission scenarios.

A Single-Model Approach at Quantifying Future Warming Uncertainty

An alternative approach has developed from a paper by Myles Allen *et al.* [25] and characterizes the uncertainty in climate change projections on the basis of results from a single model. A regression performed between observed global trends and model simulated global trends (where the latter come from separate experiments by the same model using selective subsets of the forcings known to have influenced the observed temperature trends: natural, GHGs, sulfate aerosols) allows to gauge the model's accuracy in reproducing the observed trends, and the scaling needed on the coefficients of the forcings to make the climate simulations agree with observed changes.

The model is that of a simple linear regression:

$$y = \sum_{i=1}^m (x_i - v_i) \beta_i + v_0 \quad (3)$$

where the observations y are regressed over modeled response patterns (x_i) to different forcings applied separately within the model's experimental setting. Here the y 's and x 's are derived from a time series of global fields produced by the model runs through a principal component analysis, able to extract the main spatial patterns of variability from the results of each experiment's simulations and from the observed records. The response-specific error terms in equation (3), v_i , derive from the natural variability intrinsic to model runs that introduces noise in the spatial patterns.

Then, under the assumption that the underestimate or overestimate of the effect of the specific forcings within the current climate simulations remains the same in the future runs of the model, the scaling factors β_i are applied to the model's future projections, together with their estimated uncertainty:

$$y^{\text{for}} = \sum_{i=1}^m (x_i^{\text{for}} - v_i^{\text{for}}) \beta_i + v_0^{\text{for}} \quad (4)$$

where the x 's are now model predictions under future scenarios of anthropogenic and natural forcings. A distribution for y^{for} can be derived as a function of distributions estimated for the scaling factors. On account of the need of a strong signal-to-noise ratio when detecting warming and attributing it to anthropogenic causes through the estimate of equation (3), this analysis has been mainly limited to global temperature change, and only recently the same approach has been attempted at continental scales. As was expected, the lack of a strong signal in the anthropogenic component of the historic trends at regional scales, and the consequent larger uncertainty in the estimate of the corresponding scaling coefficients produces extremely large uncertainty bounds for future projections. This problem would be exacerbated if the same analysis were applied to other climate variables (e.g., precipitation change).

Uncertainties from Boundary Conditions: Future Scenarios of Emissions

The analyses described in the section titled "Uncertainties in Climate Modeling" are concerned with the quantification of the uncertainty in future projections stemming from natural variability, (which

would confound changes in climate statistics even in the absence of external forcings) and from model uncertainty (caused by our limited understanding and capacity for representation of climate processes). A third important source of uncertainty derives from not knowing what future emissions will be like. Thus, in most cases the analysis of model output is conducted conditionally on a specified scenario of future emissions, and a range of different scenarios is explored in order to bracket as large a spectrum of “possible futures” as resources permit.

The compilation of future scenarios is in itself a complex matter, with the need to take into account a number of social, political, and economic factors: population growth will determine the consumption of natural resources, economic development will require energy, and technical and scientific progress will shape the way in which alternative (clean and renewable) sources of energy will take the place of fossil fuels. The level of international collaboration will determine how advances in carrying out sustainable development will spread around the world, especially in those countries undergoing intense economic development like China and India. IPCC, in the late 1990s, commissioned a SRES charged with developing story lines of possible future world development, and translating these into emission scenarios. Six scenario groups were developed, covering wide and overlapping emission ranges. These ranges become wider over time to account for long-term uncertainties in the driving forces (different socioeconomic developments, energy sources availability, degrees of technological innovation, etc.). Some SRES scenarios show changes in the sign of trends (i.e., initial emission increases followed by decreases), and cross paths (i.e., initially emissions are higher in one scenario, but later emissions are higher in another). The range of total carbon emissions at the end of the twenty-first century goes from 770 to 2540 GtC. Emissions from land-use changes are also considered, together with emissions of hydrofluorocarbons and sulfur. As we mentioned, uncertainties in the depiction of future emissions are complicated by the interplay of natural, technological, and social aspects and the IPCC report explicitly acknowledged that: “Similar future GHG emissions can result from very different socioeconomic developments, and similar developments of driving forces can result in different future emissions. Uncertainties in the future developments of key emission driving forces create large uncertainties

in future emissions, even within the same socioeconomic development paths. Therefore, emissions from each scenario family overlap substantially with emissions from other scenario families.” There is no easily identifiable driving force, not even one that is itself an integrator of a complex set of factors, like income growth, that has a determinant influence on the size and rate of GHG emissions.

Of course, where the analysis of most climate quantities is concerned, different scenarios cause significant differences in model output, especially when evaluating long-term future changes. These differences translate not only in different intensities of the projected changes (this being the case for average temperature changes whose behavior shows direct proportionality to the concentrations of GHGs), but also in the regional patterns by which these changes are predicted to take place. Unfortunately, the range of uncertainty determined by different scenarios may be as large as the one determined by structural uncertainty, or uncertainty in climate response to external forcings. Thus, a quantitative risk assessment of the impacts of future changes that seeks to encompass the complete range of uncertainty would require the – uncomfortable for most – attribution of probabilities to future scenarios.

Some advocates of a fully probabilistic approach to climate change have proposed approaching the problem of assigning probabilities to the different future scenarios in a necessarily subjective (Bayesian) way. This proposition has encountered strong resistance, however, and so far the only attempts at integrating the uncertainty from alternative scenarios into the final range of probabilistic projections have assumed equal probability for the different alternative [5, 16]. This “default” approach has not avoided sharp critiques, fueled by the fact that some of the scenarios have been derived heuristically rather than through explicit economic and social dynamics’ modeling, and are thus thought of being considerably less likely than others. Alternatively, impact studies have chosen to develop their analysis conditionally on a given (set of) SRES scenarios, determining consequences for each separately and – in the best examples of this approach – trying to at least ensure the representation of the outer envelope of possible future paths. Among the SRES scenarios, four have been extensively utilized in computer experiments, and the results of these experiments are thought of as being representative of a large range of the uncertainty linked to

future GHGs emissions. They are identified by the labels B1, A1B, A2, and A1FI. In an appendix, we give a more detailed description of the story lines behind them.

Approaches to Risk Assessment and Risk Management

The challenges offered to risk assessment by climatic change have been summarized recently in the Stern report as existing along two dimensions. The first has to do with the uncertainties described in the previous sections of this paper: depending on the modeling uncertainties, different climate variables have very different levels of uncertainties attached to their projections, with some variables at some scales being better “predictable” in a probabilistic sense than others. If it is relatively easier to talk about global average temperature change, especially for shorter time frames before the effects of different emission scenarios start to be felt, it is much more challenging to predict temperature at smaller regional scales, and to do so independently of scenarios, if the horizon becomes longer (for example, if predictions are sought for the end of the twenty-first century as opposed to middle of the century). Even more challenging are quantitative, probabilistic predictions of other parameters (starting from precipitation) and other aspects of the climate system (e.g., extremes). The uncertainty grows further when one tackles regime changes triggered by exceedance of systemic thresholds.

The second dimension has to do with the quantification of the costs associated with impacts and policy measures to prevent them. It is relatively easy to assess market-determined costs, but often the victims of climate change are nonmarket entities (like biodiversity or human health), and the brunt of the costs are to be borne by future generations. The magnitude of the discounting rate that is appropriate when considering decisions regarding mitigation of future climate impacts involves considerations of inter-generational equity and has been matter of significant debate. Also, the actual impacts taking place in the future are dependent on socially contingent hypotheses, leaving the possibility open of positive and negative feedbacks along the way, dynamic adjustments, and socially and politically driven changes that may influence the final outcome.

Even in the midst of these challenges, impact analysis is evolving, taking advantage of several

auxiliary sources of methods and information: down-scaling methods are providing fine-resolution climate scenarios, either by statistically linking the large scales simulated by global climate models to local climate effects, or by dynamically coupling finer resolution, limited-area regional climate models to the coarser resolution, global models generating the long-term projections. Statistical analyses are increasingly performed to postprocess the results and determine probabilistic future scenarios, or at a minimum quantify ranges of uncertainties around the simple model projections. When a characterization of the uncertainty of the full range of values of a climate variable is not feasible, threshold analysis is substituted, where only relevant turning points, outside of a determined coping range for the systems concerned, are probabilistically treated. Climate impact studies apply a variety of threshold concepts, including physical, biological, social, and behavioral thresholds. Such thresholds either indicate a point beyond which the system in question fundamentally changes in response to climate change or crosses a subjectively defined “significant” level of impacts. The selection of meaningful and applicable thresholds for risk assessment is not a trivial task. Often, a bottom-up approach is taken to determine these relevant thresholds by engaging representatives of the stakeholder communities. Examples include analyses of the enhanced risks of heat waves under future global warming [26], estimates of the extent to which historical climate change has increased mortality and disease spread and of the enhanced risk of health-related detrimental effects in the future [27], and quantitative assessment of damages to agricultural yields under different scenarios in Australia [28] and Southeast Asia [29].

Finally, there is growing scientific literature regarding integrated assessment (IA) modeling that combines scientific knowledge from the natural and social sciences to provide information that is relevant for the management of climate risks. Such modeling explicitly considers policy strategies for mitigating GHG emissions. Methods include scenario analysis – discrete or probabilistic impacts analyses of alternative emissions scenarios (see section titled “Uncertainties in Climate Modeling”); guardrail analysis – inverse analyses deriving emission ranges that are compatible with predefined constraints on temperature increase, climate impacts, and/or mitigation costs; cost–benefit analysis – generally employing

optimization models, which balance the costs of mitigation with the benefits of avoided climate impacts represented in monetary terms and produce an “optimal” solution highly dependent on model assumptions; and cost-effectiveness analysis – determining policy strategies that minimize costs and are compatible with specified constraints on future climate change or its impacts.

IA modeling techniques have been applied to a wide variety of policy questions relevant to risk management, including trade-offs in terms of mitigation costs and projected damages on different sectors and regions, implications of “wait-and-see” *versus* “hedging” approaches to emission controls, costs and benefits of differential strategies for investment in mitigation, adaptation and sequestration, and the influence of potential impacts from large-scale abrupt climate changes (such as collapse of the thermohaline ocean circulation) on policy decisions. They have also been applied generally to the assessment of response strategies to avoid “dangerous anthropogenic interference with the climate system”, the stated goal of Article 2 of the United Nations Framework Convention on Climate Change. See DeCanio [30] and Fussler and Mastrandrea [31], e.g., for further discussion of IA models and methods.

References

- [1] Barnett, T., Zwiers, F., Hegerl, G. & others, The International Ad Hoc Detection Group (2005). Detecting and attributing external influences on the climate system: a review of recent advances, *Journal of Climate* **16**, 1291–1314.
- [2] Hegerl, G.C., Karl, T., Allen, M., Bindoff, N., Gillett, N., Karoly, D. & Zwiers, F. (2006). Climate change detection and attribution: beyond mean temperature signals, *Journal of Climate* **19**(20), 5058–5077.
- [3] Santer, B.D., Wigley, T.M.L., Gleckler, P.J., Bonfils, C., Wehner, M.F., AchutaRao, K., Barnett, T.P., Boyle, J.S., Brüggemann, W., Fiorino, M., Gillett, N., Hansen, J.E., Jones, P.D., Klein, S.A., Meehl, G.A., Raper, S.C.B., Reynolds, R.W., Taylor, K.E. & Washington, W.M. (2006). Forced and unforced ocean temperature changes in Atlantic and Pacific tropical cyclogenesis regions, *Proceedings of the National Academy of Sciences* **103**(38), 13905–13910.
- [4] Henson, R. (2006). *The Rough Guide to Climate Change*, Penguin, Berkeley.
- [5] Wigley, T.M.L. & Raper, S.C.B. (2001). Interpretation of high projections for global-mean warming, *Science* **293**, 451–454.
- [6] Forest, C.E., Stone, P.H. & Sokolov, A.P. (2006). Estimated PDFs of climate system properties including natural and anthropogenic forcings, *Geophysical Research Letters* **33**, DOI: 10.1039/2005GL023977.
- [7] Allen, M.A. (1999). Do it yourself climate prediction, *Nature* **401**, 642.
- [8] Stainforth, D.A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D.J., Kettleborough, J.A., Knight, S., Martin, A., Murphy, J.M., Piani, C., Sexton, D., Smith, L.A., Spicer, R.A., Thorpe, A.J. & Allen, M.R. (2005). Uncertainty in predictions of the climate response to rising levels of greenhouse gases, *Nature* **433**, 403–406.
- [9] Harris, G.R., Sexton, D.M.H., Booth, B.B.B., Collins, M., Murphy, J.M. & Webb, M.J. (2006). Frequency distributions of transient regional climate change from perturbed physics ensembles of general circulation model simulations, *Climate Dynamics* **27**(4), 357–375.
- [10] Rougier, J., Sexton, D.M.H., Murphy, J.M. & Stainforth, D. (2007). Emulating the sensitivity of the HadSM3 climate model, *Journal of Climate* (forthcoming).
- [11] Murphy, J.M., Sexton, D.M.H., Barnett, D.N., Jones, G.S., Webb, M.J., Collins, M. & Stainforth, D.A. (2004). Quantification of modeling uncertainties in a large ensemble of climate change simulations, *Nature* **429**, 768–772.
- [12] Raisanen, J. (1997). Objective comparison of patterns of CO₂ induced climate change in coupled GCM experiments, *Climate Dynamics* **13**, 197–211.
- [13] Raisanen, J. & Palmer, T.N. (2001). A probability and decision-model analysis of a multimodel ensemble of climate change simulations, *Journal of Climate* **14**, 3212–3226.
- [14] Palmer, T.N. & Raisanen, J. (2002). Quantifying the risk of extreme seasonal precipitation events in a changing climate, *Nature* **415**, 512–514.
- [15] Giorgi, F. & Mearns, L.O. (2002). Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the ‘Reliability Ensemble Averaging’ (REA) method, *Journal of Climate* **15**, 1141–1158.
- [16] Giorgi, F. & Mearns, L.O. (2003). Probability of regional climate change calculated using the Reliability Ensemble Average (REA) method, *Geophysical Research Letters* **30**, 1629–1632.
- [17] Nychka, D. & Tebaldi, C. (2003). Calculation of Average, Uncertainty Range and Reliability of Regional Climate Changes from AOGCM Simulations via the “Reliability Ensemble Averaging” (REA) method, *Journal of Climate* **16**, 883–884.
- [18] Tebaldi, C., Mearns, L.O., Nychka, D. & Smith, R.L. (2004). Regional probabilities of precipitation change: a Bayesian analysis of multimodel simulations, *Geophysical Research Letters* **31**, DOI: 10.1029/2004GL021276.
- [19] Tebaldi, C., Smith, R.L., Nychka, D. & Mearns, L.O. (2005). Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the

- analysis of multi-model ensembles, *Journal of Climate* **18**, 1524–1540.
- [20] Smith, R.L., Tebaldi, C., Nychka, D. & Mearns, L.O. (2007). Bayesian modeling of uncertainty in ensembles of climate models, *Journal of the American Statistical Association* (forthcoming).
- [21] Greene, A.M., Goddard, L. & Lall, U. (2006). Probabilistic multimodel regional temperature change projections, *Journal of Climate* **19**, 4326–4343.
- [22] Furrer, R., Sain, S.R., Nychka, D. & Meehl, G.A. (2007). Multivariate Bayesian analysis of atmosphere-ocean general circulation models, *Environmental and Ecological Statistics* **14**, 249–266.
- [23] Benestad, R.E. (2004). Tentative probabilistic temperature scenarios for Northern Europe, *Tellus* **56A**, 89–101.
- [24] Luo, Q., Jones, R.N., Williams, M., Bryan, B. & Bellotti, W. (2005). Probabilistic distributions of regional climate change and their application in risk analysis of wheat production, *Climate Research* **29**, 41–52.
- [25] Allen, M.R., Stott, P.A., Mitchell, J.F.B., Schnur, R. & Delworth, T.L. (2000). Quantifying the uncertainty in forecasts of anthropogenic climate change, *Nature* **407**, 617–620.
- [26] Stott, P.A., Stone, D.A. & Allen, M.R. (2004). Human contribution to the European heatwave of 2003, *Nature* **432**, 610–614.
- [27] WHO (2003). *Climate Change and Human Health – Risks and Responses*, A.J. McMichael, ed, Report by WHO, UNEP, WMO, Geneva at <http://www.who.int/globalchange/publications/cchhsummary/en>. See also references therein.
- [28] Naylor, R., Battisti, D., Vimont, D., Falcon, W. & Burke, M. (2007). Assessing risks of climate variability and global warming on Indonesian rice agriculture, *Proceeding of the National Academy of Science* (forthcoming).
- [29] Jones, R.N. (2001). An environmental risk assessment/management framework for climate change impact assessment, *Natural Hazards* **23**, 197–230.
- [30] DeCanio, S.J. (2003). *Economic Models of Climate Change*, Palgrave Macmillan, New York.
- [31] Füssel, H.-M. & Mastrandrea, M.D. (2007). Integrated assessment modeling of climate change, in *Climate Change Science and Policy*, S.H. Schneider, A. Rosenzanz & M.D. Mastrandrea, eds, forthcoming.
- A1 story line and scenario family: a future world of very rapid economic growth, global population that peaks in mid-century and declines thereafter, and rapid introduction of new and more efficient technologies.
 - A2 story line and scenario family: a very heterogeneous world with continuously increasing global population and regionally oriented economic growth that is more fragmented and slower than in other story lines.
 - B1 story line and scenario family: a convergent world with the same global population as in the A1 story line but with rapid changes in economic structures toward a service and information economy, with reductions in material intensity, and the introduction of clean and resource-efficient technologies.
 - B2 story line and scenario family: a world in which the emphasis is on local solutions to economic, social, and environmental sustainability, with continuously increasing population (lower than A2) and intermediate economic development.

Within each story line, quantitative projections in terms of population and economic growth were prepared, followed by full quantification of the implied emissions using integrated assessment models. The result consisted of 40 scenarios developed by six modeling teams, all considered equally valid, i.e., with no assigned probabilities of occurrence. Six groups of scenarios were drawn from the four families: one group each in the A2, B1, and B2 families, and three groups in the A1 family, characterizing alternative developments of energy technologies: A1FI (fossil intensive), A1T (predominantly non-fossil), and A1B (balanced across energy sources). Illustrative scenarios 1 were selected by the IPCC to represent each of the six scenario groups, and have formed the basis of most of the subsequent analyses of model output. In the current assessment report, three scenarios have been chosen to span the range of possible futures: A2 for high emissions, B1 for low emissions and A1B for intermediate emissions.

CLAUDIA TEBALDI, MICHAEL D.
MASTRANDREA AND RICHARD L. SMITH

Appendix

SRES Scenarios

The SRES first developed story lines by combining factors along two dimensions: one spanning the distance between strong economic values and strong environmental values, and the other between increasing globalization and increasing regionalization. Four major story lines were put forward as follows:

Group Decision

The practice of risk assessment has been substantially influenced by the development of decision theory and its sister discipline, *Bayesian statistics* (see **Bayesian Statistics in Quantitative Risk Assessment**). However, despite the romantic image of the lone decision maker which underpins, for example, Edwards [1] or Savage [2], most risk related decisions impact on and involve multiple parties who may have interests which are not congruent (see **Expert Judgment**). Consider an example from nuclear-waste management: different communities and stakeholder groups within a country may have quite different views about where to site a waste repository, and indeed, about whether such a facility is needed at all, but if the decision is to be implementable, the decision process will have to involve many of these concerned parties (see **Hazardous Waste Site(s)**).

By itself, the presence of multiple involved parties does not necessarily mean that the decision will be made by a group. It may be that one party has the power to unilaterally achieve their goals and force others to live with the consequences, at least in the short term, as may be the case within a hierarchically structured organization. However, in situations where all parties have some degree of influence in determining what happens, it is meaningful to talk of a “group decision”.

Just as the theory of the individual decision maker can be developed from either a normative or a descriptive point of view, so too can the theory of group decision. In this article, two normative theories are contrasted, game theory (see **Game Theoretic Methods**) and social choice theory (see **Societal Decision Making**), with a descriptive theory – the social psychological theory of groups. This is by no means a complete survey of disciplines relevant to the study of group decision, and even within the disciplines which have been chosen for survey, only those concepts and results that have been particularly influential are presented. Significant controversies around interpretations are often glossed over, and to keep terminology simple, mathematical results are presented in more casual forms than is standard in the technical literature, so readers are referred to the original sources for full details.

Normative Approaches: Game Theory

Game theory [3, 4] provides one normative framework for understanding group decision situations. A game theoretic model consists of a set N of n players, each of whom has actions ($a_i \in A_i$, $i \in N$) available to them. The set of actions chosen by each player can be thought of as a profile or vector $\mathbf{a}$, with one entry from each A_i (i.e., $\mathbf{a} \in \prod_{i \in N} A_i$).

Each player i has a (nonstrict) preference ranking over these action profiles denoted by $\succsim_i$ and the corresponding strict preference ranking is denoted by $\succ_i$ (see **Utility Function**). By definition, any player would want to choose the action which would maximize the value of the realized action profile.

Game theory is conventionally divided into two parts, referred to as *noncooperative* and *cooperative* game theory. In the former, parties have to decide how to act in the absence of any ability to form binding agreements, prior to taking their actions; but in the latter, it is possible to form such agreements. In our nuclear-waste management example, if parties had to submit individual sealed statements of their position to some governmental agency, the situation might resemble a noncooperative game. On the other hand, if they had the opportunity to interact amongst themselves, and with the agency, and arrive at a joint position, it could resemble a cooperative one.

Noncooperative games often have no “solution”, in the sense that there may be no clearly best action for any given player. However, concepts which offer insight and understanding are on offer. One such concept is the idea of a *Nash equilibrium*, named after the mathematician and game theorist, John Nash.

Definition. A Nash equilibrium is an action profile $\mathbf{a} = (a_1, \dots, a_i, \dots, a_n)$ such that for every player i , $\mathbf{a} \succsim_i (a_1, \dots, a'_i, \dots, a_n)$ for all $a'_i \in A_i$.

Colloquially speaking, a Nash equilibrium is a configuration of actions such that no player can achieve an improvement by unilaterally changing only their own choice of action. As binding agreements are impossible in noncooperative game theory, having reached a Nash equilibrium, a group of players might stick there for some time: a Nash equilibrium is likely to be *stable*. Stability is surely a desirable characteristic of a solution, but there are excellent reasons why the Nash equilibrium cannot be considered a “solution” to a game in any simple way.

To see why, consider the famous game known as *Prisoner's Dilemma*. This game features, as players, two suspects who are being held, in separate cells, on remand pending trial. There is evidence to convict both on a minor charge (leading to 2 years in prison) but no evidence for the major offense (leading to 10 years) with which both are charged. The prosecutor makes both prisoners the same invitation: confess and implicate your partner. If one prisoner alone confesses, the prosecution will drop all charges against that person, but the partner will receive the full penalty of 12 years for both offenses. If both confess, both will be prosecuted for the major offense alone, leading to 10 years in prison.

One can imagine a structurally similar situation in the nuclear-waste management example. Suppose the agency asks two communities to submit sealed bids to "support" or "oppose" the establishment of a nuclear repository. If one community supports, the repository will be established on their doorstep, the worst result for them, but the best result for their counterparty. If both oppose, the repository will not be established at all, the third best outcome for both. But if both support, the repository will be established at some unpopulated third location intermediate between the two, the second best outcome for both. Clearly the temptation for both communities is to oppose. But this would result in the loss of the win-win intermediate site solution.

The *payoff matrix* for Prisoner's Dilemma is shown in Table 1. The pair of numbers in the cells represent a valuation of the associated consequence to Players 1 and 2 respectively (we shall suppose that these valuations are von Neumann–Morgenstern utilities, see **Utility Function**). Checking the cells to find Nash equilibria results in the discovery that each player has a unilateral improvement from (don't confess, don't confess) and the only equilibrium is at (confess, confess). (This corresponds to (oppose, oppose) in the site selection example.) Thus, although there are potential gains from cooperation, the players

Table 1 Payoff matrix for Prisoner's Dilemma

		Player 2	
		Don't confess	Confess
Player 1	Don't confess	3, 3	0, 4
	Confess	4, 0	1, 1

Table 2 Payoff matrix for Battle of the Sexes

		Player 2	
		Theater	Opera
Player 1	Theater	2, 1	0, 0
	Opera	0, 0	1, 2

may be driven by mistrust to what is the third best/second worst outcome for both of them.

A second game, called *Battle of the Sexes*, features a couple who wish to spend an evening together, either at the theater, or at the opera. Player 1 prefers the theater to the opera, and Player 2 prefers the opera to the theater. If the communities involved in the nuclear-waste siting example feel that they will be individually better off if a waste repository is established somewhere, *even if it is near them*, than otherwise, then the siting problem may resemble Battle of the Sexes rather than Prisoner's Dilemma.

The payoff matrix for Battle of the Sexes is as shown in Table 2. It is easy to check that both players have a unilateral improvement from (theater, opera) and (opera, theater), but neither have a unilateral improvement from (theater, theater), or (opera, opera): there are two Nash equilibria in this game. (Note that this double equilibrium structure depends on both players preferring to spend the evening together rather than not: if Player 1 had a preference for going to the theater alone to going to the opera with the partner, the normatively rational – indeed, dominating – course of action would be to unilaterally go to the theater.) Even if we make the game cooperative by supposing that players can make binding agreements prior to decision, it is not obvious what players should be advised to do.

We will run with the idea of a cooperative version of Battle of the Sexes, but to take the analysis one step further, we will suppose that the players have access to some device which generates a random number between 0 and 1, and they are prepared to stake whether to go to the theater or opera on the output of this device. To make this rule operational, they will have to choose a critical value, p , such that if the random number drawn is less than p , they will attend the theater, and if greater, the opera. While this introduction of randomization does not solve the problem, it does replace a disagreement about a discrete variable with a disagreement about a continuous variable, and this opens up the possibility

of a compromise solution based on a value of p strictly between 0 and 1.

Nash also proposed a solution to this sort of problem (the “Nash bargaining solution”) based on maximizing the product of the player’s utilities [5]. In our Battle of the Sexes example, Player 1’s utility from the randomized strategy would be $(2p + (1 - p))$ and Player 2’s would be $(p + 2(1 - p))$. The product of these two quantities simplifies to $2 + p - p^2$, and setting the derivative to 0 to find the maximum yields $p = 0.5$, for an expected utility of 1.5 for each player. Thus, for players facing symmetric payoffs, the Nash formula yields a symmetric solution. This is not a coincidence, as the Nash formula can be deduced from a small number of reasonable axioms of which this symmetry condition is one. An important feature of this solution is that it makes no assumptions about interpersonal comparability of utilities: scaling the utilities of either player by a fixed constant would yield precisely the same outcome.

Analytically, the Nash solution is an interesting device, but it can be overinterpreted. Nash himself talks about “fair” solutions to the bargaining problem, but what he means by this is not clear, and it may be doubted whether an ethically fair solution can be arrived at with such a restricted information set. Also, the use of randomized decision rules for serious decisions seems hard to justify [6]. While the Battle of the Sexes couple might use a randomizing device to solve their theater-opera dilemma, it is hard to imagine a randomized strategy being adopted in siting a nuclear-waste repository.

Various conclusions can be drawn from the game theory, but one response, in light of the limitations of the Nash equilibrium and bargaining solution concepts, is to argue that such situations call for regulatory intervention by society at large, for example, to bind players’ hands in the case of Prisoner’s Dilemma, or to break the stalemate between players by making some determinate choice in the case of Battle of the Sexes. Strategically minded players may themselves accept this sort of role for the state, even though it involves constraints on individual freedom of action: Hardin [7] describes this as “mutual coercion mutually agreed upon”.

Normative Approaches: Social Choice

Attractive as this response may seem, it raises a further problem: how is “society at large” to make a

choice? In representative democracies, citizens vote infrequently for representatives with a known position on various issues, and these representatives then take decisions on behalf of society throughout their term of office, although occasionally, when a significant decision looms, citizens may vote directly in a referendum. These national elections are commonly mirrored by the use of the voting mechanism locally, within constituencies, and by elected representatives themselves in parliament and in cabinet. In modern societies, then, voting is a prominent mechanism, and the organizing normative theory for the study of voting systems is referred to as the theory of “social choice”.

Formally, the setup for this new problem of social choice is similar to the game theory setup – the people ($i \in N$) are still the same as previously (although we will now call them “citizens” rather than “players”), and each have a preference ordering $\succsim_i$. In contrast to the game theoretic world, however, these preferences are defined not over action profiles listing individual actions, but rather over collective actions ($a \in A$). Social choice theory sees voting as a process by which by which citizens’ preferences are elicited and combined to yield a collective group ranking $\succsim_g$ (or $\succ_g$ in the case of the strict part).

A solution honored by custom and practice is *majority voting*, but it has been known for more than a century (Nanson [8], as cited by Sen [9]) that this can lead to odd consequences. Consider three citizens whose preference orderings are as shown in Figure 1: citizen 1 prefers a to b to c ; citizen 2, b to c and c to a ; and citizen 3 prefers c to a and a to b . In a vote for a against b , a will receive two votes; in a vote for b against c , b will also receive two votes; and in a vote for c against a , c will receive two votes. Thus majority voting on a pairwise basis yields a cyclical or intransitive system of preferences: a beats b beats c beats a . As transitivity is a fundamental principle of rational choice (see **Utility Function**), this is a little disconcerting.

Disconcerting it may be, but not disastrous. There are many ways other than majority voting of combining preference orderings. Nanson’s paradox shows the limitations of one approach, but it also poses an intriguing question: can we find some method of combining individual preferences that is, in general, satisfactory in some sense?

To consider this, we need a little more formality. A way of combining individual preference orderings

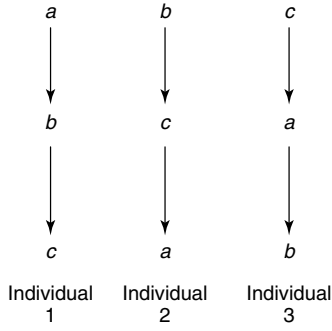


Figure 1 Preference orderings leading to intransitive social choice

$\succsim_i$ which yields a group ordering $\succsim_g$ can be called a *constitution* [6]. Note that on this definition, majority voting is not a constitution, as it does not yield a well-defined ordering. Clearly, the number of possible constitutions, logically speaking, is very great, and one might expect that it would be very hard to say anything conclusive at all. As it happens however, this, the field of social choice is dominated by a single, definitive result: Arrow's theorem [10]. This theorem shows that, for situations with at least two citizens and at least three alternatives, no constitution can be found which satisfies four seemingly very innocuous conditions.

The four key conditions are often stated as axioms.

The first axiom, *nondictatorship*, says that no citizen can dictate the group preference ordering. This is essentially a nontriviality condition: obviously dictatorships do exist, and, perhaps, there may even be situations where dictatorships may be desirable (e.g., in crisis situations where time is very scarce), but this form of preference aggregation can clearly not be considered a meaningful solution to the problem of group decision.

Axiom D Nondictatorship. *There is no i such that for all $a, b \in A$, $a \succsim_i b$ implies $a \succsim_g b$.*

The second axiom, *unrestricted domain*, says that the constitution must be defined for all profiles of preference ordering on the part of citizens. Again, this is a nontriviality condition: we must work with the preference orderings with which citizens present us, whatever they may be. If we wish to restrict the domain of preference orderings, we would have to exclude citizens who had the sort of preference orderings we regard as pathological.

Axiom U Unrestricted domain. $\succsim_g$ is defined for all possible $\succsim_1, \succsim_2, \dots, \succsim_n$.

The third axiom is the *Pareto principle*: it says that if *all* citizens prefer one action to another, then the group collectively should prefer one action to another. The Pareto principle is a foundation stone of welfare economics, and is almost completely uncontroversial (but see discussion of Sen's theorem below). Indeed, one could argue that if all citizens prefer one action to another, then there is no social dilemma: in this sense, the Pareto principle can be considered to be almost a nontriviality condition.

Axiom P Pareto principle. *If, for some $a, b \in A$, $a \succ_i b$ for all i , then $a \succ_g b$.*

To understand the fourth and final axiom, it helps to think procedurally. Suppose that we aggregate preferences, using some constitution. At some point, we find that some action we thought was available, is not, in fact, available: it has to be deleted from the action set. (In Arrow's original monograph, he gives the example of an election, during the course of which one of the candidates dies.) Should it matter to the final result whether we make this discovery *prior to* aggregation (and so all our citizens delete the action from their preference ordering) or *subsequent to* aggregation (in which case we delete the action from the derived social ordering)? If it does make a difference, given a suitable profile of preferences amongst the citizenry, an unscrupulous politician could conceivably manipulate the results of the election by adding or deleting items from the election list. This seems obviously undesirable, and so it is plausible to claim that it *should not matter* when we delete the action. This axiom is called *independence of irrelevant alternatives*.

It should be noted that some prominent, and popular systems of voting do not respect this principle. For example, consider a system, known as the *Borda method*, which assigns points to actions based on citizen ranks. Specifically, if there are six possible actions, the point score for citizen i for an action is $7 - r_i$, where r_i is the rank given by citizen i , and the collective ordering is based on the aggregate points for each action. Suppose moreover, that citizens rank actions as shown in the first panel of Figure 2: then a (13 points) is collectively strictly preferred to d (12 points) under this constitution. However, if c is deleted from the preference orderings of the citizenry,

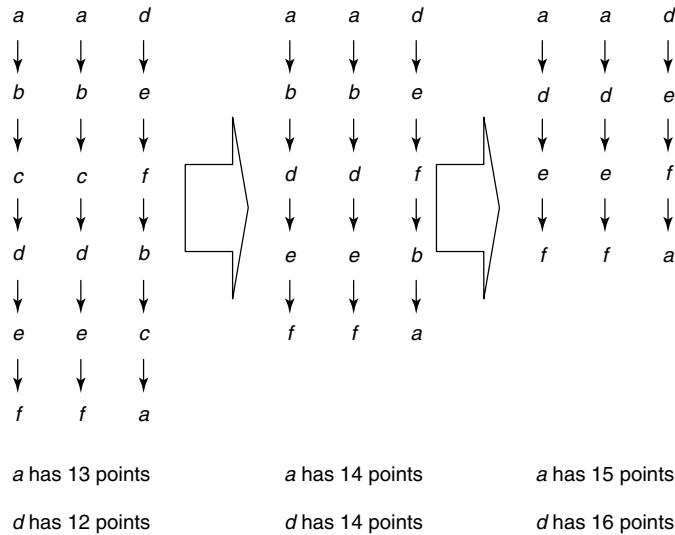


Figure 2 The social ordering from a rank points scheme can be reversed by deletion of irrelevant actions

as shown in the second panel, then a gains one point from its resulting move up the rankings, but d gains two points, with the result that a and d draw neck and neck. Deleting b leads to a system of preference orderings in which d is now the outright victor in the social ordering.

Axiom I Independence of irrelevant alternatives.

Consider two possible profiles of public preferences $(\succsim_1, \succsim_2, \dots, \succsim_n)$ and $(\succsim'_1, \succsim'_2, \dots, \succsim'_n)$, and call their aggregations under the constitution $\succsim_g$ and $\succsim'_g$ respectively. Consider also a subset A^* of A (i.e., delete one or more actions from A). If, for all citizens i , and for all pairs of alternatives $a, b \in A^*$, $a \succsim_i b \Leftrightarrow a \succsim'_i b$ (i.e., the citizen's rankings on the restriction of A to A^* are the same), then $a \succsim_g b \Leftrightarrow a \succsim'_g b$ (i.e., the collective ranking is the same).

It is now possible to present Arrow's theorem.

Theorem 1 (Arrow's theorem). *There is no constitution which combines $\succsim_1, \succsim_2, \dots, \succsim_n$ to yield $\succsim_g$ in a manner consistent with all of D, U, P, and I.*

Versions of the proof of Arrow's theorem are presented in Arrow [10], Sen [9], and French [6].

Arrow's theorem packs a powerful punch. Although the theorem is weakened if the citizen's preference orderings can be constrained to those of a certain structure ("single peakedness"), there is no easy way to defeat the result. The most promising

avenue is to expand the frame, including more information which allow statements relating to the relative interpersonal intensity of preferences to be made ("X prefers a to b more strongly than Y prefers c to d "). This raises the question of who is to make such profoundly difficult judgments in the absence of some supra decision maker (see **Supra Decision Maker**).

A further depressing result in the same vein, the impossibility of a Paretian Liberal (IPL), is due to Sen [9]. Sen observes that an attractive principle of many choice rules is liberalism: for choices between actions which pertain to me and me alone (e.g., my private religious observances or sexual practices), the social ordering should respect my preferences. In the case of the nuclear-waste siting example, a liberal position might be that local communities should have to give their consent to having a repository (perhaps supplemented with certain inducements) sited in their locality. This idea is formalized through a further axiom.

Axiom L Liberalism. *For each person i , there is a nonempty subset A_i containing at least two actions, and for $a, b \in A_i$, if $a \succ_i b$, then $a \succ_g b$.*

Obviously, the practical implication of this axiom rather depends on the precise content of A_i (smoking in restaurants, making provocative statements about others' religious beliefs). But formally, the interesting

thing about L is that it opens the way to another impossibility theorem.

Theorem 2 (IPL). *There is no constitution which combines $\succsim_1, \succsim_2, \dots, \succsim_n$ to yield $\succsim_g$ in a manner consistent with all of U , P , and L .*

The full proof of IPL is presented in Sen [9], but to give a flavor, consider the following variation on the nuclear-waste siting example. Geological investigation has determined that two localities are suitable for siting a repository, around Alphatown and around Betaville. The social action set contains three actions, siting the repository at Alphatown (a), siting the repository at Betaville (b), and not building a repository at all (c). The inhabitants of Alphatown do not favor building a repository as a solution to the nuclear-waste management problem. However, they are not economically well-off, and the community would benefit from the investment if a repository were to be built locally, so if it is to be built, they would rather it be built nearby. In Betaville, on the other hand, there is enthusiasm for the repository solution in principle, but Betaville residents do not lack for the material things of life and would prefer the repository to be sited elsewhere. These preferences are shown in Figure 3, with the actions in each respective A_i shown in bold.

The stage is now set for the main trick. Liberalism requires that since Betaville prefers b to c , $b \succ_g c$, and since Alphaville prefers c to a , $c \succ_g a$, and the Pareto principle requires that since both prefer a to b , $a \succ_g b$. But this leads to $b \succ_g c \succ_g a \succ_g b$, a cycle, and so the relation $\succ_g$ is not a true ordering. This unsettling result illustrates that the

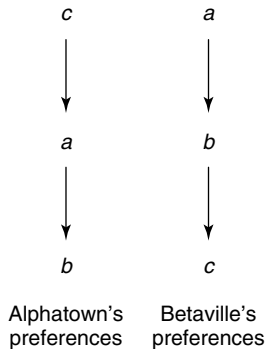


Figure 3 Preference orderings in the Paretian Liberal example

liberalism principle and the Pareto principle can be incompatible: no well-defined constitution exists which can be guaranteed to always do both.

Looking back across game theory and social choice, the reader may feel a sense of disappointment. The normative theory of the individual decision maker gives us subjective expected utility theory (see **Subjective Expected Utility**) and multiattribute value theory (see **Multiattribute Value Functions**), but the normative theory of group decision gives us no uniquely satisfying theoretic framework for making group decisions, like our nuclear-waste repository siting decision. One might expect that this would mean that the scope for descriptive research on group decision is limited, since what has driven the corresponding research program on individual decision making since its inception [1] is the contrast between the behavior of the idealized “perfectly rational” decision maker and what real people actually do.

Descriptive Approaches: Social Psychology

Yet, descriptive research on group decision and performance, more generally, has flourished within the discipline of social psychology [11, 12]. Although the explicit connection between the normative and descriptive aspects of group decision are weaker than in the case of individual decision, there is considerable resonance between the key concepts of the two areas of research. Indeed, the social choice concept of a constitution has an analogue in the descriptive theory in the concept of a *decision rule*, which “specifies, for any given individual set of preferences regarding some set of alternatives, what the group preference or decision is regarding the alternatives” [13, p. 327].

A concept related to (but distinct from) that of the decision rule is a *decision scheme matrix (DSM)* [14–16], a notion which has been particularly popular in the study of jury decision making. These matrices (called *DSMs* here) capture *descriptively* how individual preferences, stated prior to interaction, relate to a group decision. A DSM thus is an input–output model and can be fitted to particular data sets obtained from experiments with decision making groups (for example, mock juries) in laboratory settings.

The structure of a DSM is as follows. The columns are indexed after the set of actions A , and the rows of the matrix are indexed by possible *ex ante*

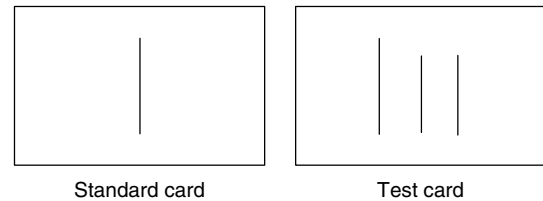
Table 3 Examples of decision scheme matrices

Distribution of support		D_1		D_2		D_3		D_4		
a	b	a	b	a	b	a	b	a	b	Hang
5	0	1	0	1	0	1	0	1	0	0
4	1	1	0	0.8	0.2	1	0	0	0	1
3	2	1	0	0.6	0.4	1	0	0	0	1
2	3	0	1	0.4	0.6	1	0	0	0	1
1	4	0	1	0.2	0.8	1	0	0	0	1
0	5	0	1	0	1	0	1	0	1	0

distributions of support given by group members for each action $a \in A$. The quantities in the cells represent probabilities of the group *ex post* choosing a , given such and such a distribution of support. A number of DSMs for five persons and two actions are shown in Table 3. D_1 describes a situation where the majority preference wins out. D_2 describes a “proportionality” model, where the probability of the final decision being some given action is proportional to the support for that action. In D_3 , as long as at least one person chooses a , the social decision is a : this rule is sometimes called *truth wins* and is appropriate for decisions where the correct answer is obvious as soon as one person has found it (e.g., in a task like the taking of square roots). D_4 models a situation where unless there is complete prior consensus on one action or another, the group will “hang” and fail to come to a decision.

The relationship between DSMs and decision rules is not one-to-one. D_1 , for example, could summarize behavior in a group using majority voting. It could also summarize behavior in a group which does not use an explicit decision rule at all, but comes to a decision based on the chair’s assessment of the “balance of opinion”. It could even summarize the behavior of a jury which has been instructed to use a consensus decision rule, but where, through the process of deliberation, preferences change, and eventually all group members converge on the same verdict.

Understanding how preferences may change in interacting groups has been a central aim of social psychological research on groups. Two possible ways of accounting for preference shift are in terms of *information processing* (as the group has shared all relevant information about the problem at hand, there is less room for disagreement); and in terms of *conformity to majority opinion*. A seminal example

**Figure 4** Cards in Asch’s experiment

of experimental work on the latter is that of Asch [17]. Subjects in Asch’s experiments were told that they were participating in a test of visual abilities. They were then placed in a group of apparent fellow subjects (actually confederates of the experimenter), and in a series of eighteen trials, they were presented with two cards, a standard card and a test card, as illustrated in Figure 4. Group members were asked to declare in sequence which of the lines on the test card was equal in length to the line on the standard card, with the actual subject declaring last. On twelve of the trials, the confederates would systematically declare a wrong line (e.g., the rightmost line on the test card in Figure 4) to be equal to the standard line. Although the right answer to the task is obvious, and subjects had no difficulty in arriving at the correct answer when tested alone, 76% of subjects gave at least one incorrect answer in the group setting, demonstrating the hold which majority opinion can have on an isolated (but correct) subject.

Conformity to majority opinion is a recognizable phenomenon. However, if everyone conformed, society would rapidly degenerate into an intellectual monoculture, and insofar as this has not happened, the conformity account of opinion change is incomplete. Alongside the conformity research, a separate research tradition has studied the ways in which a

minority can influence majority opinion. A key lesson from this research is the importance of behavioral style. In particular, experimental studies show that a minority which reliably calls blue lights green can eventually change majority definitions of green and blue, but this effect does not exist when the minority is not consistent [18]. Although conformity to majority opinion and minority influence may appear to be opposite, they are in fact closely linked: the reason that minorities can be influential is precisely because majorities recognize that minority status is uncomfortable, and so conclude that obstinate minorities must have some good reason for their persistence.

Both majority and minority influence are examples of the general phenomenon of *social influence*, “a change in the judgments, opinions, and attitudes of an individual as a result of being exposed to the judgments, opinions, and attitudes of other individuals” [19]. Researchers distinguish two different types of influence, *normative* and *informational* influence. Normative influence arises from the suppression of a view through a reluctance to express overt dissent, and may lead to *public compliance* with the influencers, with no change in underlying opinion. Informational influence occurs when the influencer concludes that it is more likely that his own judgment is faulty than that everyone else is wrong, and may lead to *private acceptance* of a new belief.

Studies of majority and minority influence shed some light on the mechanisms of opinion formation in groups, but do not give a sense of how these effects might play out in a freely interacting group. Yet, systematic patterns do exist. One counterintuitive, yet robust, example is *polarization*, which relates to a tendency for group decisions arising out of unstructured interaction to be more extreme than would be suggested by individual attitudes elicited prior to interaction. For example, if a group is asked to select the maximum level of riskiness at which they would consider it reasonable to submit to some curative surgical procedure, and group members tend to be at the more cautious (more risky) end of the scale, the level of risk of the option preferred by the group will tend to be lesser (greater) than the average level of risk of the preferred options of individual group members.

Explanations for polarization based on both a *persuasive arguments* and a *mere exposure* rationale have been suggested [20], although other modes of

explanation do exist [21]. The idea behind the persuasive arguments account is that if group members tend to lean to one direction, the pooling of the arguments behind these individual preferences will tend to lead to a more pronounced group position. Mere exposure accounts, on the other hand, are predicated on one of two mechanisms, either the *removal of pluralistic ignorance*, or *one-upmanship*. Removal of pluralistic ignorance accounts emphasize that group members’ originally stated preferences may underrepresent the true extremity of their true preferences for purposes of conformance: as it becomes clear that the center of opinion is off-center in one direction, this caution is muted. One-upmanship effects, in contrast, are predicated on the notion that people tend to favor positions which are more extreme in a socially valued direction. Experimental evidence for all these effects exists, and polarization seems to be a composite phenomenon, arising out of a number of different processes, the relative contribution of which depend on local context.

A well-known application of social psychological ideas such as social influence and polarization to the understanding of policy making is that of Janis [22], who coined the popular notion of “groupthink” and used it to analyze a number of policy “fiascoes” such as Kennedy’s Bay of Pigs invasion of Cuba, and the failure of the US Navy to prepare for attack at Pearl Harbor. Janis’ argument is that in these cases, a combination of high group cohesion, time pressure, strongly directive leadership, and isolation (frequently self-imposed) from other sources of information, led to gross overconfidence on the basis of manifestly inadequate analysis.

In the context of nuclear-waste management, it is interesting to note that current practice has moved away from the decide-announce-defend (DAD) model, where a small group of nuclear insiders would propose a management solution based on a technical analysis. Instead the fashion is for more consultative processes designed to be open to external scrutiny and to involve a broad spectrum of society with a diverse range of viewpoints (see Pescatore and Vári [23] for a review of practice in a number of OECD countries). Although the reasons given for this shift are typically in terms of bolstering public trust, from a social psychological point of view, DAD seems set up to produce groupthink. Consequently, a more participative model of decision making in such situations may not only lead to greater public confidence, but

may also lead to an improvement in the quality of the decisions made.

In this section, we have surveyed some key ideas in the social psychology of group decision. Although the *explicit* connection between this body of research and the normative research described in previous sections is not particularly strong, descriptive research can be seen as calling into question the notion of a fixed system of preferences which the normative theory postulates. Some recent research, for example, the growing body of work on behavioral game theory [24, 25], enriches the formal frameworks of the normative research with behavioral insights about endogenous preference change. There is a parallel here with the way in which the theory of the individual decision maker has developed: prospect theory and its competitors attempt to build a formalized descriptive theory which generalizes *subjective expected utility theory* (see **Subjective Expected Utility**). However, a review of this work is beyond the scope of the current article.

Conclusion

This review article has canvassed some highlights of normative and descriptive research on group decision, sketching some key ideas from game theory and social choice on the one hand, and from the social psychological study of groups, on the other. A central contrast has been between a normative theory which takes as exogenously given an action set and a system of individual (“player”, “citizen”) preferences over these actions, and a descriptive theory, which focuses on how preferences change as a result of group interaction.

A second contrast has been between the theories of individual and group decision making. The decision theory of the individual decision maker, at least as interpreted by its proponents, provides in subjective expected utility and multiattribute value theory a satisfying account of how decisions *should* be made. The theory of group decision, on the other hand, does not offer any such comfort. Rather, it throws into focus the drawbacks of particular attractive solution concepts such as the Nash equilibrium and bargaining solution in the case of game theory, or majority voting and the Pareto and liberalism principles, in the case of social choice.

In practice, group decisions about how risks are to be managed do get made, but often they are made

badly in ways that the theory warns us about: they may be unfair, unsustainable, arbitrary, or based on poorly shared information, or deliberately or accidentally suppressed dissent. Fortunately, a growing body of research and practice in the management of societal risks studies show that it is possible to design decision processes which make appropriate use of risk and decision analytic technologies, but which are also responsive to the social context in which the decision takes place. The reader is referred to other articles in this encyclopedia for a more thorough discussion (see **Decision Conferencing/Facilitated Workshops**).

Acknowledgment

The author acknowledges helpful discussions with Larry Phillips and Mara Airoidi.

References

- [1] Edwards, W. (1954). The theory of decision making, *Psychological Bulletin* **51**, 380–417.
- [2] Savage, L.J. (1954). *The Foundations of Statistics*, John Wiley & Sons, New York.
- [3] Luce, R.D. & Raiffa, H. (1957). *Games and Decisions*, John Wiley & Sons, New York.
- [4] Osborne, M.J. & Rubinstein, A. (1991). *A Course in Game Theory*, MIT Press, Cambridge.
- [5] Nash, J.F. (1950). The bargaining problem, *Econometrica* **18**, 155–162.
- [6] French, S. (1986). *Decision Theory: An Introduction to the Mathematics of Rationality*, Ellis Horwood, Chichester.
- [7] Hardin, G. (1968). The tragedy of the commons, *Science* **162**, 1243–1248.
- [8] Nanson, E.J. (1882). Methods of elections, *Proceedings of the Royal Society of Victoria* **18**, 197–240.
- [9] Sen, A.K. (1970). *Collective Choice and Social Welfare*, Holden-Day, San Francisco.
- [10] Arrow, K.J. (1951). *Social Choice and Individual Values*, Yale University Press, New Haven.
- [11] Stangor, C. (2004). *Social Groups in Action and Interaction*, Psychology Press, New York.
- [12] McGrath, J.E. (1984). *Groups: Interaction and Performance*, Prentice Hall, Englewood Cliffs.
- [13] Miller, C.E. (1989). The social psychological effects of group decision rules, in *Psychology of Group Influence*, P.B. Paulus, ed, Lawrence Erlbaum, Hillsdale.
- [14] Davis, J.H. (1973). Group decision and social interaction: a theory of decision schemes, *Psychological Review* **80**, 97–125.
- [15] Davis, J.H. (1980). Group decision and procedural justice, in *Progress in Social Psychology*, M. Fishbein, ed, Lawrence Erlbaum, Hillsdale, Vol. 1.

- [16] Laughlin, P.R. (1980). Social combination processes of cooperative problem-solving groups on verbal intellectual tasks, in *Progress in Social Psychology*, M. Fishbein, ed, Lawrence Erlbaum, Hillsdale, Vol. 1.
- [17] Asch, S.E. (1956). Studies of independence and conformity: a majority of one against a unanimous majority, *Psychological Monographs: General and Applied* **70**, 1–70.
- [18] Moscovici, S., Lage, E. & Naffrechoux, M. (1969). Influence of a consistent minority on the responses of a majority in a color perception task, *Sociometry* **32**, 365–380.
- [19] Van Avermaert, E. (1988). Social influence in small groups, in *Introduction to Social Psychology: A European Perspective*, M. Hewstone, W. Stroebe, J.-P. Codol & G.M. Stephenson, eds, Blackwell Science, Oxford.
- [20] Isenberg, D.J. (1986). Group polarization: a critical review and meta-analysis, *Journal of Personality and Social Psychology* **50**, 1141–1151.
- [21] Friedkin, N.E. (1999). Choice shift and group polarization, *American Sociological Review* **64**, 856–875.
- [22] Janis, I.L. (1982). *Groupthink: Psychological Studies of Policy Decisions and Fiascos*, Houghton Mifflin Company, Dallas.
- [23] Pescatore, C. & Vári, A. (2006). Stepwise approach to the long-term management of nuclear waste, *Journal of Risk Research* **9**, 13–40.
- [24] Camerer, C. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton University Press, Princeton.
- [25] Rabin, M. (1993). Incorporating fairness into game theory and economics, *American Economic Review* **83**, 1281–1302.

Related Articles

Expert Judgment

Public Participation

Decision Conferencing/Facilitated Workshops

ALEC MORTON

Hazard and Hazard Ratio

General Discussion of Mathematical Concept of Hazard

Hazard or the hazard function arises in the context of *survival analysis*, also known as *failure analysis*, or more generally analysis of the time to event. Examples include months of survival after treatment for cancer, the time to first repair of consumer products, time from installation to first leak of a roof, or time to wear out of a new automobile tire. The time variable may be age, calendar time, time in use, duty cycles, miles driven, or any other variable that tracks the aging process of the system under study. The hazard function describes the current chance of failure for the population that has not yet failed.

In general, let T represent a random variable that is the time to event. Further, let F be the cumulative probability distribution function of time, or age, to failure, i.e., $F_T(t) = \text{probability}\{T \leq t\}$. The survivor function $S_T(t) = 1 - F_T(t)$ is the proportion of the population that survives up till time t ; $S_T(t) = \text{probability}\{T > t\}$. In engineering applications this is sometimes referred to as the *reliability function*.

The hazard function at time t is the conditional probability of failure, conditional upon survival up to time t . Many of the failure time models are more easily described through the hazard function, h_T .

$$\text{Specifically, } h_T(t) = \lim_{\Delta \rightarrow 0} \frac{1}{\Delta} \frac{[F_T(t + \Delta) - F_T(t)]}{F_T(t)} \quad (1)$$

If the hazard function is constant, it means that the probability of failure is constant; it does not depend upon age, time in service, or number of duty cycles. If the hazard function is increasing with time, it describes an aging, or wear-out, process, i.e., the older the units under study, the higher the failure rate of the remaining units. If the hazard function is decreasing, it describes an infant mortality or burn-in process, i.e., the longer the units have survived, the lower the failure rate for the remaining units.

Many processes have a U-shaped hazard function. This is called the *bathtub* curve and describes the fact that many processes, including human life, are subject

both to infant mortality and to aging and wear-out. There is a high failure rate early on in the process (in humans a high death rate in the first month of life), followed by a period of low mortality or low failure, until a certain age is attained and the mortality rate or hazard function accelerates. Figure 1 illustrates these hazard patterns.

Hazard Functions for Some Typical Survival or Failure Time Models

The exponential, Weibull, and lognormal distributions are three of the commonly used statistical models of time to failure. The hazard function describes the failure process and provides a convenient way to visualize how the risk of failure changes with age or time.

Exponential

The exponential model is the simplest of the time to failure models. It is the memory-less process. The hazard does not change with time. The probability of failure for the surviving units does not depend upon previous history of the units. The survivor function is $S_T(t) = \exp(-\lambda t)$, and the probability distribution function is $1 - \exp(-\lambda t)$. The hazard function is given by

$$\begin{aligned} h_T(t) &= \lim_{\Delta \rightarrow 0} \frac{1}{\Delta} \frac{[F_T(t + \Delta) - F_T(t)]}{F_T(t)} \\ &= \frac{[d\{1 - \exp(-\lambda t)\}/dt]}{\{1 - \exp(-\lambda t)\}} = \lambda \end{aligned} \quad (2)$$

Weibull

The Weibull family of distributions is a flexible family of time to event distributions that allows both increasing and decreasing hazard functions. It is the most popular parametric model of time to failure in engineering applications. Despite its flexibility, it does not permit a hazard function that is both decreasing over part of the range and increasing over part of the range, as would be needed to model the bathtub curve type of hazard function. The bathtub curve can be modeled as a mixture of two Weibull functions, one with a decreasing hazard function and one with an increasing hazard function (see Figure 2).

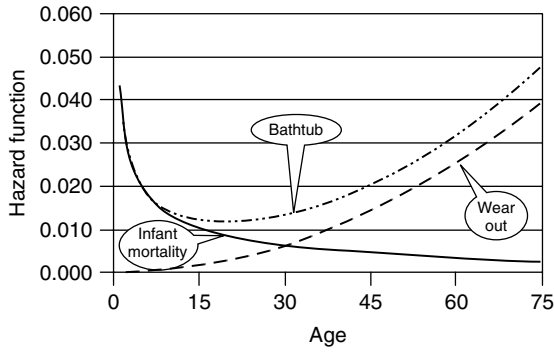


Figure 1 Three typical hazard functions

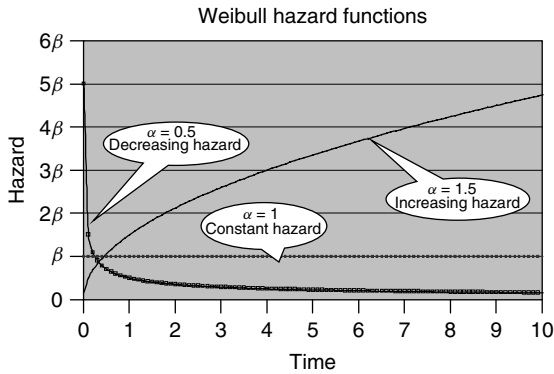


Figure 2 Hazard functions for the Weibull family

The survivor function for the two parameter Weibull family of distributions is $S_T(t) = \exp[-(t/\beta)^\alpha]$, and the hazard function is $\alpha/\beta[t/\beta]^{(\alpha-1)}$. The parameter β is a scale parameter and α is the shape parameter. If α is less than 1, the hazard function is decreasing; if α is greater than 1, the hazard function is increasing. When $\alpha = 1$ the Weibull function reduces to the exponential function.

In some applications a third parameter, τ_0 , the threshold, is added. In this case, it is assumed that there is no risk of failure prior to τ_0 , i.e., the hazard is zero prior to τ_0 . This is equivalent to shifting the timescale by τ_0 .

Lognormal

Here the logarithm of the time to failure follows the normal or Gaussian distribution. The survivor function is $S_T(t) = \Phi[-\log(t) - \alpha]/\sigma]$, and the hazard

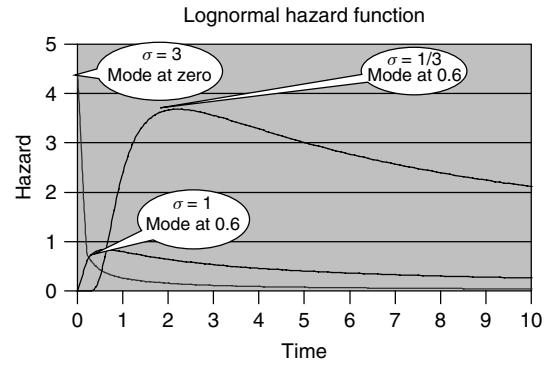


Figure 3 Hazard functions for the lognormal family

function is $\phi[\log(t) - \alpha]/\sigma]/\{\Phi[-\log(t) - \alpha]/\sigma] \times [(\log(t) - \alpha) \times \sigma]$, where Φ is the standard normal probability distribution function and ϕ is the standard normal density function. Hazard functions in the lognormal family are unimodal. If σ is greater than 1, the peak will be at zero and the hazard will be strictly decreasing. If σ is 1 or less, the hazard will be zero at $t = 0$, rise to a peak and then decrease (see Figure 3).

Nonparametric and Semiparametric Models

The life table method can be used to estimate hazard functions with no predetermined shape. This method is frequently used in medicine and epidemiology with some application in the fields of finance and economics. It provides an extension of the ideas of linear regression to the realm of survival analysis. The parameters of the model have a natural interpretation as relative risks.

The hazard function is estimated as a piecewise linear function, with constant hazard at each age group in the life table. The precision of this type of analysis is limited only by the precision of the age groupings.

The Cox proportional hazards method models the hazard function as follows:

$$h_T(t) = \lambda_0(t) \exp(\gamma + X\beta) \quad (3)$$

where $\lambda_0(t)$ is a base hazard function of unspecified shape, X is a vector of risk factors measured on each subject, and β is a vector of parameters describing the relative risk associated with the risk factors. This method may be considered semiparametric as no assumptions are made about the base hazard

function, but the effect of the risk factors is assumed to be linear in the log of the hazard function. If the risk factors are indicator functions, then there is no restriction and the model is fully nonparametric.

Hazard Ratio

The hazard ratio is a way of comparing the relative risk of failure associated with two groups. It is perhaps easiest to understand in the context of a simple Cox proportional hazards model. Assume that there is only one risk factor X , which takes the value 1 if the high risk condition is present and 0 if not. In this case, it is natural to compare the hazard function for the high risk condition, i.e., $h_T(t|X = 1) = \lambda_0(t) \exp(\gamma + \beta)$, to the hazard function for the low risk condition, $h_T(t|X = 0) = \lambda_0(t) \exp(\gamma)$.

The hazard ratio is simply given by

$$\frac{h_T(t|X = 1)}{h_T(t|X = 0)} = \exp(\beta) \quad (4)$$

This represents the instantaneous relative risk conditional upon survival until time t .

This may be the most direct way to measure relative risk in the time to failure context. Relative risk, as measured by the probability of failure in the high risk group compared to the probability of failure in the low risk group, will of course depend upon the time period during which the two groups are exposed to risk of failure. All groups will eventually reach end of life, and in the absence of competing risks, the relative risk will tend to 1.0 as the time increases. If we choose a particular time τ , the relative risk of failure up until time τ can be calculated from the hazard function as

$$RR = \frac{F_T(t|X = 1)}{F_T(t|X = 0)} \quad (5)$$

$$RR = \frac{\left\{ 1 - \exp \left[- \int_0^t \lambda_0(u) \exp(\gamma + X\beta) du \right] \right\}}{\left\{ 1 - \exp \left[- \int_0^t \lambda_0(u) \exp(\gamma) du \right] \right\}} \quad (6)$$

for the proportional hazards model.

In the typical proportional hazards analysis, the baseline hazard, λ_0 , is neither known nor estimated from the data. Thus, it is not possible to estimate or

calculate a direct measure of relative risk, or odds ratio, for any particular time t .

In general, the relative risk at time t is given by

$$RR = \frac{\left\{ 1 - \exp \left[- \int_0^t h_T(u|X = 1) du \right] \right\}}{\left\{ 1 - \exp \left[- \int_0^t h_T(u|X = 0) du \right] \right\}} \quad (7)$$

If the hazard ratio is known or estimable from the data, it will be possible to calculate the relative risk using this formula.

Hazard Ratio Patterns

Constant Hazard Ratio. The Cox proportional hazard model is one form of constant hazard ratio – that is, the hazard ratio is constant, conditional upon the value of the vector of covariates. If the hazard is assumed to depend only upon the current value of the covariates, the covariates can be allowed to change with time without loss of generality in the model.

In any model in which the hazard ratio is assumed to be constant, the interpretation of the hazard ratio as instantaneous relative risk is understandable. Another interpretation is the relative “force or mortality”, i.e., the relative probability of transitioning from the prefailure to the failure state.

Average Hazard Ratio. In many situations, the assumption of proportional hazard ratios may be unrealistic. Kabfleish and Prentice [1] developed a simple noniterative estimate of the average hazard ratio for the two-sample problem. The idea is to take a weighted average of the hazard ratio:

$$\text{Average hazard ratio} = \int_0^\infty \frac{h_T(u|X = 1)}{h_T(u|X = 0)} dG(u) \quad (8)$$

where G is a suitable weight function. G can be chosen to put the most weight during periods that are *a priori* determined to be the most important.

If the hazard ratio is always greater than 1.0 (or always less than 1.0), then group $X = 0$ has better (or worse) survivability than group $X = 1$ for all time periods. In this case, the average hazard ratio has a fairly straightforward interpretation. However, if the hazard ratio is greater than 1.0 for some ages and less than 1.0 for other ages, the average hazard ratio is

difficult to interpret. A group with the average hazard ratio less than 1.0 can have poorer survivability than the comparison group. For example, a group with high infant mortality and low hazard for older ages may have average hazard ratio less than 1.0 when compared to a group with low infant mortality and higher hazard for older ages, and yet at each age a higher proportion of the comparison group has survived. In this case, the high infant mortality outweighs the low hazard at older ages. The result is a misleading average hazard ratio.

Statistical Methods

Here we describe a general nonparametric method for describing the relative risk of failure for two or more populations. The method is very flexible in that no assumptions about the underlying form of hazard function are needed.

Rank Tests for Comparison of Survival Curves

For ease of interpretation, we start with a comparison of two groups, e.g., a treated group and a control group. The rank tests that we describe can be then generalized for the comparison of several groups. Here n_1 is the sample size from the first population, n_2 is the sample size from the second population, and $n_1 + n_2 = m$. Let h_{Ti} be the hazard function for population i .

If $h_{T1}(t) = h_{T2}(t)$ for all t , the survival distributions are the same for the two populations, and *vice versa*. Consequently, a test for equality of haz-

no censoring, a two-sample rank test such as the Wilcoxon could be used. Typically, the data does include censoring, i.e., not all subjects are observed till failure; for some subjects we only know that the time to failure is greater than the observed time under study. Generalization of the Wilcoxon test for this situation was proposed by Gehan [2] in 1965 and Breslow [3] in 1970. Mantel [4] and Cox [5] proposed the log-rank test for this problem in 1966 and 1972, respectively.

Let $t_1 < t_2 < t_3 < \dots < t_k$ be distinct times at which failures were observed. Further, let O_{ij} be the number of failures observed in the i th group at time t_j and R_{ij} be the number of observations in the i th group at risk immediately prior to time t_j . Let R_j be the total observations in both groups at risk immediately prior to time t_j and O_j be the total number that fail at time j .

Tarone and Ware [6] showed that both the log-rank test and the Gehan modification of the Wilcoxon test are part of a general class of statistics based upon comparing the observed number of failures at each time to the number of failures expected under the null hypothesis of equal distribution of time to failure, and summarizing across the k distinct failure times *via* a χ^2 statistic.

Specifically, let $v' = (v_1, v_2)$ where

$$v_i = \sum_{j=1}^k W(t_j) \times \{O_{ij} - E_{ij}\} \quad (9)$$

where E_{ij} is the expected number of failures in group i at time $t_j = O_j \times R_{ij}/R_j$.

Let $\mathbf{V}$ be the estimated covariance matrix of v :

$$\mathbf{V}_{i1} = \sum_{j=1}^k W^2(t_j) \times \left\{ \frac{[O_j \times (R_j - O_j) \times (R_j \times R_{ij}d_{i1} - R_{ij} \times R_{1j})]}{[R_j^2 \times (R_j - 1)]} \right\} \quad (10)$$

and functions is equivalent to a test of equality of survivor distributions.

Rank tests for comparing two survival curves are similar to other two-sample rank tests. We rank the times to failure for the pooled sample. If the two populations have the same distribution of time to failure and there is no censoring, the values of the ranks for group 1 will be like a random sample, without replacement of a sample of size n_1 from the integers $1, 2, \dots, m$. In the case of

Then the test statistic $S_w = v'\mathbf{V}^{-1}v$ will have a χ^2 distribution with one degree of freedom, where $\mathbf{V}^{-1}$ is the generalized inverse of $\mathbf{V}$.

W is a weight function that specified the test. The commonly used examples of this type of test are as follows:

$$\text{Wilcoxon: } W(t_j) = R_j \quad (11)$$

$$\text{Log rank: } W(t_j) = 1.0 \quad (12)$$

$$\text{Tarone-Ware: } W(t_j) = R_j^{1/2} \quad (13)$$

$$\text{Peto-Peto: } W(t_j) = \prod_{l=1}^{j-1} \left\{ \frac{1 - O_l}{(R_l - 1)} \right\} \quad (14)$$

This shows that the commonly used rank methods are all in the same class of tests. This general mathematical formulation facilitates the derivation of the properties of the tests.

This method can be generalized to comparisons of k populations, in which case the test statistic, S_w , will have a χ^2 distribution with degrees of freedom equal to the rank of $\mathbf{V}$ under the null hypothesis that all of the populations have the same distribution of time to failure.

These χ^2 tests are similar to the test for partial association in contingency tables (see Gart [7]).

The log-rank test is more powerful when the hazard ratio is, in fact, constant as in the Cox proportional hazards model. These tests are easily generalized for covariates that are more general indicators of the population from which the observations were drawn (see Kabfleish and Prentice [8, chapter 6]).

Examples

Medicine and Public Health

The Cox proportional hazard model or other hazard ratio models are commonly used in clinical trials and epidemiology. The proportional hazards model can be used in retrospective or case-control studies as well as in traditional prospective time to failure analyses.

One recent example of use of hazard ratio models is the study of survival time after the death or hospitalization of a spouse, by Christakis and Allison [9]. They studied more than half a million couples over age 65 who were enrolled in Medicare in the United States in 1993. This was a prospective study with nine years of follow-up.

Higher risk of death was found after the death of a spouse for both men and women. The risk of death was also increased following hospitalization of the spouse for dementia, psychiatric disease, hip fracture, or stroke; risk of death was not statistically significantly elevated after hospitalization of the spouse for colon cancer.

Environment

Survival analysis and the hazard ratio are used in environmental and ecological applications (see **Environmental Health Risk; Environmental Hazard**). In a recent edition of Fisheries and Aquatics Bulletin, scientists for the US Geological Survey [10] used the estimated hazard ratio to develop a disease infection model. The hazard ratio was used to compare mortality rates for rainbow trout, walleye, and catfish that were infected with *Flavobacterium columnare* to mortality rates of noninfected controls of the same species.

Finance

Mehran and Peristani [11] used the Cox proportional hazards model to evaluate the decision to return to private status among small businesses that had recently gone public with an initial public stock offering. Firms with low stock turnover, small institutional ownership of the stock, and small growth in coverage by financial analysts were seen to be more likely to return to private ownership, and to do so in a shorter period of time, than firms with high analysis coverage, institutional ownership, or high volume of trading.

Summary

The hazard function is a tool for describing the process of time to failure. It is used in epidemiology, medicine, ecology, engineering, and manufacturing settings, as well as in economics and finance. The formula for the hazard function is simpler and easier to interpret than the survivor function or the probability distribution of time to failure.

The hazard ratio is a general way of examining relative risk. In particular, Cox proportional hazards models are flexible and allow introduction of covariates to model the failure process. The parameters of the Cox model are usually interpretable as relative risk.

References

- [1] Kalbfleisch, J. & Prentice, R. (1981). Estimation of the average hazard ratio, *Biometrika* **68**, 105–112.

- [2] Gehan, E. (1965). A generalized Wilcoxon test for comparing arbitrarily single censored samples, *Biometrika* **52**, 203–223.
- [3] Breslow, N. (1970). A generalized Kruskal-Wallis test for comparing k samples subject to unequal patterns of censorship, *Biometrika* **57**, 579–594.
- [4] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports* **50**, 163–170.
- [5] Cox, D. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society B* **26**, 103–110.
- [6] Tarone, R. & Ware, J. (1977). On distribution-free tests for equality of survival distributions, *Biometrika* **64**, 156–160.
- [7] Gart, J. (1972). Contribution to the discussion on the paper by D.R. Cox, *Journal of the Royal Statistical Society B* **26**, 212–213.
- [8] Kalbfleisch, J. & Prentice, R. (1980). *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, New York.
- [9] Christakis, N. & Allison, P. (2006). Mortality after the hospitalization of a spouse, *The New England Journal of Medicine* **354**, 719–730.
- [10] U.S. Geological Survey. (2005). Fish passage – fishways are being evaluated and improved for passage of American shad in the northeast, *Fisheries and Aquatics Bulletin* **IV**(1), 1.
- [11] Mehran, H. & Peristiani, S. (2006). *Financial Visibility and the Decision to go Private*, Federal Reserve Bank of New York, http://www.crsp.uchicago.edu/forum/papers/small_company_focus/Financial%20Visibility%20and%20the%20Decision-%20formerly-%20Analyst%20Coverage%20and%20the%20Decision%20to%20Go%20Private.pdf.

Further Reading

- Abrahamowicz, M., MacKenzie, T. & Esdaile, J. (1996). Time-dependent hazard ratio: modeling and hypothesis testing with application in *Lupus Nephritis*, *Journal of the American Statistical Association* **91**, 1432–1439.
- Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion), *Journal of the Royal Statistical Society A* **135**, 185–206.
- Schoenfeld, D. & Tsiatis, A. (1987). A modified log-rank test for highly stratified data, *Biometrika* **74**, 167–175.

ROSE M. RAY

Hazardous Waste Site(s)

In general terms, hazardous waste sites are portions of land contaminated by voluntary or involuntary introduction of hazardous substances, such as improper management or disposal of waste and accidental spills or voluntary discharge of substances. In regulatory terms, the same terminology may have different meaning from country to country. In the United States “hazardous waste sites” are regulated under the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), commonly known as *Superfund*, enacted by the US Congress in 1980. A waste is considered hazardous if it exhibits one or more of the following characteristics: ignitability, corrosivity, reactivity, or toxicity (*see What are Hazardous Materials?*). In the European Union regulatory framework, although classifications of hazardous waste and dangerous substances are provided (Directive 91/689/EEC, Directive 67/548/EEC), the term *hazardous waste sites* is not used. The more general term *contaminated sites* is used to indicate local soil contamination.

The most common contaminants found at hazardous waste sites are as follows:

- some heavy metals and metalloids, like *lead* (*see Lead*), manganese, *arsenic* (*see Arsenic*), copper, nickel, cadmium, chromium, zinc, barium, *mercury* (*see Mercury/Methylmercury Risk*), antimony, beryllium;
- aromatic hydrocarbons, like *benzene* (*see Benzene*), toluene;
- polycyclic aromatic hydrocarbons (PAHs), like naphthalene and benzo(a)pyrene;
- chlorinated hydrocarbons, like vinyl chloride, dichloroethane, trichloroethane, trichloroethylene, tetrachloroethylene, carbon tetrachloride;
- polychlorinated biphenyls (PCBs) (*see Polychlorinated Biphenyls*).

The worldwide dimension of the problem is very relevant. For example, the number of potentially contaminated sites in Europe is estimated at approximately 3.5 million with 0.5 million sites being really contaminated and needing remediation [1]. The cost of remediation per site has been estimated between €19 500 and 73 500, with an overall cost for the remediation of all sites around €28 billion. Comparing

these numbers with other countries, e.g., the United States, can be misleading, because estimations greatly change according to the definition of contaminated site. Notwithstanding, it can be reasonably said that the situation is similar in all industrialized countries, and even worse in underdeveloped countries.

Due to the scale of the problem, cleanup of all hazardous waste sites in a short or medium time frame, e.g., 25 years, is not technically and economically feasible. In this context, cleanup interventions should be addressed to the most relevant sites and cleanup goals should be tailored to site-specific conditions.

Within this scope, risk assessment methodologies have been developed. The risk assessment of contaminated sites is a technical procedure to analyze the likelihood and magnitude of adverse effects on human health or the ecosystem due to the direct or indirect exposure to hazardous substances in soil. It is based on characterization of the site and on sound scientific evidence concerning the environmental behavior and toxicity of the contaminants. It follows that this is a retrospective risk assessment, because environmental and human exposure to the contaminants has been occurring. As shown in Figure 1, potential exposure pathways of human beings to hazardous substances in soil or waste include e.g., direct contact, inhalation, ingestion of particles, drinking of contaminated groundwater, etc. Similar exposure scenarios can be described for plants, soil organisms, wildlife, and fishes. In the study of hazardous waste sites, three types of retrospective applications of risk assessment can be operationally distinguished:

1. risk-based ranking of sites at regional scale,
2. screening by application of generic soil quality standards (SQSs),
3. site-specific risk assessment, i.e., the assessment based on specific environmental and exposure conditions at the site.

Generic SQSs are usually based on risk assessment estimates for standard scenarios. When the comparison with SQSs indicates a potential significant risk, the application of a site-specific risk assessment is usually recommended. In the following paragraphs, an overview is given of the risk-based ranking and site-specific risk assessment methodologies. A further insight on risk assessment methodologies to derive generic SQSs is provided in **Soil Contamination Risk**.

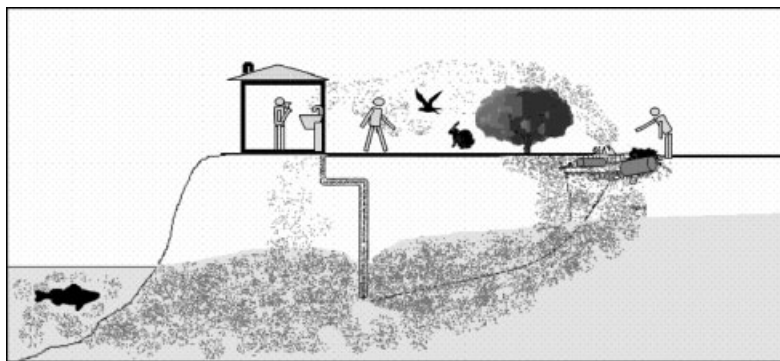


Figure 1 Exposure of human and ecological receptors to contaminants (plumes represent contaminant molecules) from hazardous waste sites [Reproduced from [2]. Environmental Protection Agency, 1989.]

Ranking Hazardous Waste Sites

Since the early 1980s, technical procedures have been developed in several countries to rank the hundreds or thousands of hazardous waste sites registered at regional or national levels. One of the most popular systems is the Hazard Ranking System adopted by the United States Environmental Protection Agency [2]. More recently, the European Environmental Agency has collected information about 25 systems worldwide and proposed a new system, named *PRAMS*, to rank areas under risk for soil contamination in Europe [3]. In general, all these procedures are qualitative scoring systems based on source, pathway, and receptor indicators. The source indicator represents the toxic potential of the waste, e.g., toxicity and waste quantity. The pathway indicator represents the likelihood that the hazardous substances are released and transported in the environment. Different pathway indicators can be defined through the soil (direct contact), air, groundwater and surface water media, respectively. The receptor indicators refer to the presence of human or sensitive environments that can be affected by the release. The risk depends on the concurrence of source, pathways, and receptors. Different risk scores can be obtained for each exposure pathway, e.g., on the basis of the product of source, pathway, and receptor indicators. The overall risk score for the site can be derived from the risk scores for all the pathways, by selecting the maximum value or the root mean square of all the scores.

The scoring of each indicator is usually built on structured scoring systems of subtended factors and parameters. For example, in the USEPA's

Hazard Ranking System the source indicator is the sum of two factors, waste characteristics and containment measures, respectively. In turn, the waste characteristics factor depends on parameters like hazardous waste quantity, toxicity, mobility, persistence, and bioaccumulation. For each parameter, scores are defined for quantitative intervals or presence/absence. In case the information is not available, the highest score is generally recommended. Thanks to the capability and widespread use of geographical information system (GIS) technologies, the development of a new spatially resolved ranking system can be foreseen. The integration of ranking systems within GIS allows the proper consideration of spatial correlation among sources and receptors. It also means that the contribution of multiple sources on the same receptors can be identified, as well as the most impacted areas at regional scale.

Site-Specific Risk Assessment

Site-specific risk assessment is based on quantitative modeling of exposure and toxicological effects to estimate the risk for human health or ecological receptors associated with contamination of a specific site. In a reverse application, acceptable risk levels are set to derive acceptable contaminant concentrations, i.e., site-specific remediation goals. The site-specific risk assessment differs from the screening risk assessment, which is applied to standard scenarios to derive general SQSs, in that quantitative modeling is applied to site-specific conditions and, whenever possible, supported by site monitoring.

Thus, in site-specific risk assessment the characterization of the site plays a key role. In the following paragraphs the fundamentals of human health and ecological risk assessment (ERA) for hazardous waste sites are presented.

Human Health Site-Specific Risk Assessment

As mentioned above, characterization of the site is at the very core of the risk assessment. The risk assessment can then be described in four traditional phases: (a) hazard identification; (b) exposure assessment; (c) toxicity assessment; and (d) risk characterization.

Characterization of the Site

The proper characterization of the site is the prerequisite for reliable risk estimates. It should provide a site conceptual model (SCM) and all the data necessary for modeling the fate and the transport of the contaminants [4].

The SCM is a scheme illustrating contaminant distributions, potential or observed released mechanisms, exposure pathways and migration routes and potential receptors. It usually includes maps, cross sections, and diagrams. The development of the SCM is composed of three steps [2]:

1. data gathering on site operation history and development of a preliminary conceptual model of the site;
2. site investigations;
3. refinement of the SCM.

Due to the spatial heterogeneity of the contamination, the definition of effective sampling strategies is a very sensitive issue. A major question is *where* and *how many* samples may be representative of the site contamination. The sampling strategy usually has the objective to detect all the contamination hot spots present at the site. The optimal number of samples may be determined on a statistical basis. In particular, statistical methods able to incorporate available knowledge and expert judgment on the sub-areas that are most likely to contain a hot spot were proven to be more effective and efficient [5, 6]. The use of screening analytical techniques, in particular those supporting real-time results on the field, has

been recently promoted in combination with traditional methods to optimize sampling strategies and reduce spatial uncertainties [7, 8]. The basic idea is that screening methods on the field allow for less sensitive but more dense sampling and to adapt the sampling plan as new information become available (dynamic work strategies) [8].

Hazard Assessment

The objective of hazard assessment is the identification of substances, named *contaminants of concern* (CoCs) present at the site that need to be included in the risk analysis. Since the number of hazardous substances present at the site can be very high, the hazard assessment has also the objective to focus the analysis on a manageable subset of representative CoCs. Scoring systems are traditionally applied. The most popular is that proposed by the United States Environmental Protection Agency [2] and named *concentration-toxicity screening*. It consists of scoring the chemicals according to their concentration and toxicity to obtain a hazard factor HF_{ij} , as follows:

$$HF_{ij} = C_{ij} \times T_{ij} \quad (1)$$

where

HF_{ij} : risk factor for chemical i and medium j

C_{ij} : maximum concentration of chemical i in medium j

T_{ij} : toxicity value for chemical i in medium j .

Separate scores should be calculated for each environmental medium (e.g., soil, groundwater, etc.). Hazard factors are summed to obtain the total hazard factor for all chemicals in a medium and the subset of chemicals accounting for a cumulative percentage of e.g., 99% are retained.

In general, the selection might include additional factors like spread distribution, frequency of detection, mobility, persistence in the environment, chemicals associated with site operation and treatability.

Toxicological Assessment

The objective of the toxicological assessment is the evaluation of the exposure dose–health effect relationship for each contaminant. For the toxicological properties, chronic, subchronic, and acute toxic effects can be distinguished, but in the case of human

exposure to hazardous waste sites the chronic exposure is usually the most relevant. Threshold and nonthreshold effects are distinguished. For threshold effects contaminants, an exposure dose below which adverse effects are negligible can be derived. The toxicity is expressed by human health reference doses (RfD) ($\text{mg/kg}_{\text{body weight}^{-\text{day}}}$) derived on the basis of toxicological tests data (i.e., no observed adverse effect levels or benchmark doses) and uncertainty/assessment factors. In the case of nonthreshold effects substances, e.g., genotoxic carcinogens, it is assumed that even at very low levels of exposure residual risk cannot be excluded. The toxicity is expressed by potency factors, i.e., estimated incremental incidence of cancer versus daily intake in a lifetime [$\text{mg/kg}_{\text{body weight}^{-\text{day}}}$] $^{-1}$. Different potency factors and RfDs can be estimated for oral, inhalation, and dermal exposure routes.

Reference doses and potency factors have been derived for the most common contaminants by national and international competent authorities. Popular sources of toxicity data available on the web are the integrated risk information system (IRIS, <http://www.epa.gov/iris>) managed by the United States Environmental Protection Agency, and the risk assessment information system (RAIS, <http://rais.ornl.gov/>) managed by the United States Department of Energy (DoE).

Exposure Assessment

The exposure analysis represents the core of risk assessment. The objectives are the quantification of the magnitude (exposure concentration), frequency, duration and routes of exposure. Three steps can be defined. The first step is characterization of the exposure setting [2]. Basic site characteristics, such as climate, vegetation, groundwater hydrology, and the presence and location of surface water are identified in this step. Potentially exposed populations are identified and their vulnerability is described in terms of e.g., distance from the site, activity patterns, and the presence of sensitive subpopulation groups. The future use of the site and exposed populations should also be considered. The second step is identification of exposure pathways. Exposure diagrams are usually developed that indicate contamination sources, release mechanisms, transports medium, exposure routes and receptors. Exposure scenarios can be

referred to the worker, the trespasser, the residential user, the recreational user, and the constructions on the site. The third step is quantification of the exposure, which comprises the estimation of exposure concentrations and the calculation of intakes. Given the exposure concentration, the administered dose (intake) can be calculated by the following general equation:

$$I = \frac{(C \times CR \times EF \times ED)}{(BW \times AT)} \quad (2)$$

where

I : intake ($\text{mg/kg}_{\text{body weight}^{-\text{day}}}$)

C : concentration at exposure point (e.g., mg l^{-1} in water or mg m^{-3} in air)

CR : contact rate (e.g., 1 day^{-1} or $\text{m}^3 \text{ day}^{-1}$)

EF : frequency (day yr^{-1})

ED : exposure duration (years)

BW : body weight (kg)

AT : average time (days).

This equation is modified for specific exposure pathways, by including factors such as the rate of respiration for inhalation or the amount of skin exposed for dermal contact. The transport processes that are commonly modeled in human exposure assessment to hazardous waste sites are represented in Table 1. A broad range of fate and transport models is available in a number of software packages. If refined estimation of the exposure is needed, a tiered approach is recommended which starts from application of conservative simple modeling, like steady-state equilibrium equations, and moves to more complex models, like numerical and probabilistic models. Among the simpler modeling tools, the algorithms proposed for petroleum release sites by the American Society of Testing and Materials have encountered a wide popularity [9]. It is important to consider that exposure modeling requires a wide range of site characterization and chemical-specific data, other than exposure data. The list of input values in fate and transport modeling might include more than 60 parameters in the following categories:

- compound-specific physicochemical properties;
- meteorological parameters;
- surface parameters;
- hydrogeological parameters;
- characterization of buildings at the site;
- human receptors parameters.

Table 1 Cross-media transfers, transport and fate and intake routes commonly modeled in the assessment of human exposure at hazardous waste sites

Source	Cross-media routes	Transport and fate	Intake routes
<u>Direct exposure pathways</u>			
Surface soil	–	–	Ingestion of soil Dermal exposure
<u>Air exposure pathways</u>			
Surface soil	Volatilization and particulate emission	Air dispersion	Ambient air inhalation of soil vapors and soil derived dust. Dermal exposure to dust
Subsurface soil	Volatilization	Air dispersion	Ambient air inhalation of soil vapors
Subsurface soil	Enclosed space volatilization	–	Indoor inhalation of soil originated vapors
Groundwater	Volatilization	Air dispersion	Ambient air inhalation of soil vapors
Groundwater	Enclosed space volatilization	–	Indoor inhalation of soil originated vapors
<u>Groundwater exposure pathways</u>			
Soil	Leaching to groundwater and dilution	Groundwater dispersion, transport and attenuation	Groundwater ingestion: inhalation of volatilized domestic water, showering (dermal contact and inhalation)
<u>Diet exposure</u>			
Soil (e.g., contaminated from run off)	Uptake of homegrown vegetables and fruits	–	Consumption of homegrown vegetables (and attached soil), fruit, meat, dairy products
<u>Surface water exposure</u>			
Soil	Leaching to groundwater, contact with surface water and dilution	Transport in the unsaturated zone, groundwater dispersion, transport, and attenuation	Dermal contact and water ingestion while swimming, consumption of fish and shellfish

Risk Characterization

This phase combines the results of the exposure analysis and the toxicological assessment. The risk for threshold effects contaminants is defined as the quotient of the estimated intake dose to the RfD, also named the *hazard index*. The risk for nonthreshold effect contaminants (e.g., carcinogenetic) is defined as the chronic daily intake dose multiplied by the carcinogenic potency factor. The result is the lifetime incremental risk associated with exposure to the hazardous waste site. The calculated (incremental)

risk can be compared to the risk caused by a variety of other sources, such as accidents, cigarette smoking, etc.

If acceptable risk thresholds are defined, reverse calculation allows for derivation of contaminant concentrations at the site below which the risk is considered not relevant. These concentration thresholds can be used as remediation goals. Acceptable risk thresholds are usually taken as the hazard index equal to 1 for threshold effects contaminants and the incremental lifetime risk between 10^{-6} and 10^{-4} for nonthreshold effects contaminants. However, the

acceptable risk thresholds should not be regarded as quantitative statements about risk, i.e., cancer affecting one person in 1 million. They should be considered conventional and operational indicators of “negligible risk” closely related to the extent of the conservative assumptions adopted in the risk estimation [10].

Risk characterization should always be supported by estimation of related uncertainties. Epistemological uncertainty, those related to ignorance about a poorly characterized variable or phenomenon, and uncertainty associated with the intrinsic variability of the variable or phenomenon, are usually considered together. However, they should be distinguished to the extent possible, because they have different implications in risk assessment: epistemological uncertainty is reducible by further measurements or studies, whereas variability is not. Uncertainty regards models assumptions and prediction capabilities as well as model input values. Uncertainty related to model input values can be quantified by probabilistic methods, like multilevel interval analysis, Monte Carlo and Latin hypercube simulations. Monte Carlo simulation implies the description of input parameters by means of probability distribution functions and the generation of a probability distribution of risk estimation.

Risk assessment of hazardous waste sites is usually a stepwise approach: it starts from conservative assumptions to verify if site conditions satisfy the criteria for a quick regulatory closure or warrant a more site-specific and accurate, but also more expensive and time consuming, risk characterization. The appropriate, i.e., cost-effective, level of investigation and modeling depends on the complexity of the site and remediation costs.

Site-Specific Ecological Risk Assessment

Historically, at hazardous waste sites the ecological concern has been second to the human health concern. Nevertheless, in particular where large areas are involved and ecologically sensitive environments are affected, ecological considerations can play a driving role in the definition of remediation interventions (see **Ecological Risk Assessment**). This is often the case of mining waste sites or in proximities of surface water bodies (see **Water Pollution Risk**). The USEPA guidance for the application of ERA to hazardous waste sites [11] defines a three-phase process,

i.e., (a) problem formulation, (b) stressor response and exposure analysis, and (c) risk characterization. Essential for all steps are negotiation and agreement of the need for further action between the risk assessor, the risk manager and other stakeholders.

In ERA, effect assessment is more complex than for human health risk assessment. This is mainly because protection of the terrestrial ecosystem would impact multiple species and their relations with the abiotic environment. In the scope, toxicity effects to the microbial community, plants and soil fauna are usually considered, but it is rather difficult to predict effects to the population, community and ecosystem levels, as well as impairments to soil functioning. In general, a simplification is adopted which consists of identification of specific assessment endpoints. Assessment endpoints are explicit expressions of the actual environmental value that is to be protected. Two elements are required to define assessment endpoints. The first is identification of the specific valued ecological entity. This can be a species (e.g., earthworms like *Eisenia andrei* or specific plants), a functional group of species (e.g., soil dwelling organisms), a community (e.g., soil fauna), an ecosystem, a specific habitat or others. The second is the characteristic about the entity that is important to protect and is potentially at risk (e.g., density of soil dwelling organisms). Criteria for the selection of assessment endpoints are their ecological relevance, their susceptibility to potential stressors and relevance to management goals.

Major uncertainties in ERA (see **Ecological Risk Assessment**) are the extrapolation of ecosystem relevance of selected assessment endpoints, the modeling of exposure and effects, the inclusion of factors like bioavailability and biomagnifications. In order to deal with conceptual uncertainties in a pragmatic way, it has been proposed that weight of evidence approaches be used. The rational is that many independent ways to arrive at one conclusion will provide a stronger evidence for ecological effects, making ERA less uncertain [12]. In the so-called TRIAD approach, three lines of evidence about ecological impairments associated with the contamination can be defined [13]:

- chemical, i.e., the concentration of chemicals, their bioavailability and bioaccumulation;
- eco-toxicological, i.e., results of bioassays carried on site samples;

- ecological, i.e., field ecological observations at the contaminated site.

Risk Management

It should be borne in mind that contamination of the soil is not only an issue of human and environmental health, but also a barrier to the redevelopment of the sites. This close connection between environmental and redevelopment problems is at the core of risk management. The benefits associated with the reuse of the site are often the main driver of detailed risk assessment and remediation interventions. Risk-management strategies can be based on the elimination or reduction of the hazardous waste source, or the control of exposure pathways or receptors. Besides the technical estimation of the risk, the perception and societal value of the risk play an utmost important role. For this reason, risk communication strategies are relevant aspects of risk management. Decision support systems have been developed to compare environmental and socioeconomic benefits and drawbacks behind different remediation options [14, 15].

References

- [1] European Commission (2006). Impact assessment of the thematic strategy on soil protection, *Thematic Strategy for Soil Protection*, European Commission Joint Research Centre, SEC, p. 1165.
- [2] USEPA (1989). *Risk Assessment Guidance for Superfund. Vol. 1: Human Health Evaluation Manual, Part A, Interim Final*, EPA/540/1-89/002, Office of Emergency and Remedial Response, Washington, DC.
- [3] EEA, European Environmental Agency (2006). *Towards an Eea Europe-Wide Assessment of Areas Under Risk for Soil Contamination*, <http://eea.eionet.europa.eu> (accessed Oct 2006).
- [4] Ferguson, C.C., Darmendrail, D., Freier, K., Jensen, J., Kasamas, H., Urzelai, A. & Vecter, J. (1998). *Risk Assessment for Contaminated Sites in Europe*, LQM Press Edition, Nottingham.
- [5] Ferguson, C.C. & Abbachi, A. (1993). Incorporating expert judgment into statistical sampling designs for contaminated sites, *Land Contamination and Reclamation* **1**, 135–142.
- [6] Carlon, C., Nathanail, C.P., Critto, C. & Marcomini, A. (2004). Bayesian statistics-based procedure for sampling of contaminated sites, *Soil and Sediment Contamination* **3**(4), 329–345.
- [7] van Ree, D. & Carlon, C. (2003). New technologies and future developments “Is there a truth in site characterisation and monitoring?” *Soil Contamination and Reclamation Journal* **11**(1), 37–47.
- [8] Crumbling, D.M., Groenjes, C., Lesnik, B., Lynch, K., Shockley, J., Van Ee, J., Howe, R.A., Keith, L.H. & McKenna, J. (2001). Managing uncertainty in environmental decisions: applying the concept of effective data at contaminated sites could reduce costs and improve cleanups, *Environmental Science and Technology* **35**(9), 404A–409A.
- [9] ASTM (1995). *Standard Guide for Risk-Based Corrective Action Applied at Petroleum Release Sites*, RBCA, E 1739-95.
- [10] Ferguson, C.C. & Denner, J. (1994). Developing guideline (trigger) values for contaminants in soil: underlying risk analysis and risk management concepts, *Land Contamination and Reclamation* **2**, 117–123.
- [11] USEPA (1998). *Guidelines for Ecological Risk Assessment*, EPA/630/R-95/002F, Washington, DC.
- [12] Suter II, G.W., Efromymson, R.A., Sample, B.E. & Jones, D.S. (2000). *Ecological Risk Assessment for Contaminated Sites*, CRC Press, Lewis Publishers, Boca Raton.
- [13] Jensen, J. & Mesman, M. (eds) (2006). *Ecological Risk Assessment of Contaminated Land-decision Support for Site Specific Investigations*, report no. 711701047, RIVM.
- [14] Pollard, S.J.T., Brookes, A., Earl, N., Lowe, J., Kearny, T. & Nathanail, C.P. (2004). Integrating decision tools for the sustainable management of land contamination, *Science of the Total Environment* **325**, 15–28.
- [15] Carlon, C., Critto, A., Ramieri, E. & Marcomini, A. (2007). Desyre decision support system for the rehabilitation of contaminated megasites, *Integrated Environmental Assessment and Management* **3**(2), 211–222.

Further Reading

US-EPA (2001). *Current Perspectives in Site Remediation and Monitoring: Using the Triad Approach to Improve the Cost-Effectiveness of Hazardous Waste Site Cleanups*, EPA 542-R-01-016, Office of Solid Waste and Emergency Response, Washington, DC.

Related Articles

Conflicts, Choices, and Solutions: Informing Risky Choices

Comparative Risk Assessment

Environmental Hazard

Environmental Risks

Hazards Insurance: A Brief History

Long before the development of insurance as such, early civilizations evolved various kinds of insurance-like devices for transferring or sharing risk. For example, early trading merchants are believed to have adopted the practice of distributing their goods among various boats, camels, and caravans, so each merchant would sustain only a partial loss of his goods if some of the boats were sunk or some of the camels and caravans were plundered. This concept survives today in the common insurance practice of avoiding overconcentration of properties in any one area, and by spreading risk through reinsurance arrangements (see **Reinsurance**).

As long ago as the Code of Hammurabi in 1950 B.C., commercial shipping included the practice of bottomry, whereby a ship's owner binds the vessel as security for the repayment of money advanced or lent for the journey. If the ship is lost at sea, the lender loses the money, but if the ship arrives safely, he receives both the loan repayment and a premium, specified in advance, which exceeds the legal rate of interest. Bottomry is one of the oldest forms of insurance, and was used throughout the ancient world [1]. By 750 B.C., the practice of bottomry was highly developed, particularly in ancient Greece, with risk premiums reflecting the danger of the venture. Mutual insurance also developed at this time, whereby all parties to the contract shared in any loss suffered by any trader who paid a premium.

The decay and disintegration of the Roman Empire in the fifth century A.D. was followed by the development of small, isolated, self-sufficient, and self-contained communities. Since international commerce practically ceased, there was little need for sophisticated risk-sharing devices such as insurance. However, the revival of international commerce in the thirteenth and fourteenth centuries revitalized use of insurance-type mechanisms. In England, Lloyd's Coffee House provided a gathering place for individual marine underwriters, leading to the formation of Lloyd's of London in 1769. London then became the center of the global marine insurance market.

Emergence of Fire Insurance

Modern fire insurance was developed after the Great Fire of London in 1666 destroyed over three quarters of the city's buildings. In the early eighteenth century, Philadelphia took the first steps toward the protection of property from the fire peril; the city possessed seven fire-extinguishing companies. Philadelphia also passed regulations concerning the nature and location of buildings in the city. Led by Benjamin Franklin, Philadelphians organized the first fire insurance company in America, the Philadelphia Contributionship for the Insurance of Houses from Loss by Fire. It was followed by the creation of four more insurance companies before the end of the eighteenth century, and by an increasing number of fire insurance companies in other cities in the first half of the nineteenth century. The financial problems associated with catastrophic losses are well illustrated by the great New York fire of 1835, which swept most of the New York fire insurance companies out of existence.

Conflagrations such as the great New York fire of 1835 and the Chicago fire of 1871 focused attention on fire prevention over other natural hazards. Established in 1866, the National Board of Underwriters in New York City concentrated on protecting the interests of fire insurance companies by establishing safety standards in building construction and working to repress incendiarism and arson. In 1894, fire insurance companies created the Underwriters Laboratories as a not-for-profit organization to test materials for public safety. Founded in 1896, the National Fire Protection Association still plays a major role in reducing the number of fires by encouraging proper construction. Today's engineering and technical schools conduct research and educate and train fire safety personnel.

Factory Mutual Insurance Companies Provided Incentives and Standards for Improved Risk Management

As an important ingredient for risk spreading and risk reduction, insurance is exemplified by the factory mutual insurance companies founded in the early nineteenth century in New England. These companies offered factories an opportunity to pay a small premium in exchange for protection against potentially large losses from fire, reflecting the risk spreading

objective of insurance. The mutuals also required inspections of factories both before and after issuing a policy. Poor risks had their policies canceled, and premium reductions were given to factories that instituted loss prevention measures.

For example, the Boston Manufacturers worked with lantern manufacturers to encourage them to develop safer designs and then advised all policyholders that they had to purchase lanterns from those companies whose products met their specifications. In many cases, insurance would only be provided to companies that adopted specific loss prevention methods. For example, one company, the Spinners Mutual, only insured risks where automatic sprinkler systems were installed. The Manufacturers Mutual in Providence, Rhode Island, developed specifications for fire hoses and advised mills to buy only from companies that met these standards [2].

Wind Damage and the Rise of Extended Coverage

Insurance against wind damage was first written in the second half of the nineteenth century, but only by a limited number of companies. In its earliest form, the policy contract provided for insurance against loss by fire or storm. Until 1880, this type of insurance was sold by local or farmers' mutual insurance companies on rural risks, rather than city risks, and was confined to the East. The first insurance covering windstorm by itself was initiated in the Midwest and was called *tornado insurance*. It continued as a separate policy until 1930.

Windstorm insurance began to evolve in its present form in 1930 when the stock fire insurance companies filed an "Additional Hazards Supplemental Contract" in conjunction with the standard fire policy. This new contract provided a single rate for coverages that had traditionally been offered as separate policies for tornado, explosion, riot and civil commotion, and aircraft damage insurance. Incorporating this array of coverage in one contract was seen as a radical step in a generally conservative industry. In October 1930, the contract was expanded even further when the New England Insurance Exchange revised it to include perils from hail and motor vehicles, as well as all types of windstorms, not just tornadoes. In 1938, the Additional Hazards Supplemental Contract was renamed as *extended coverage* (EC), and smoke damage was added to the coverage.

When it was introduced in the East, the EC endorsement was at first purchased by few individuals, and was viewed as a luxury. Comparatively little of this insurance was sold in Massachusetts until after the 1938 hurricane, the first to hit New England in a century. This event stimulated sale of the EC endorsement in part, because many banks then required that it must be added to fire insurance on mortgaged property. Although there were some changes in the perils covered during the 1930s and 1940s, the basic package has been retained and continues to be used today [3].

Property-Casualty Insurance Today

The insurance industry today is divided into two broad areas: property-casualty insurance companies, and life and health insurance companies. Property-casualty companies predominate, with close to 4000 property-casualty companies and about half as many life insurers operating in 1992. The insurance laws of the various states recognize the sharp distinction between these different types of insurance. The property-casualty companies are the main source of coverage against natural hazards in the United States and hence bear most of the insured losses from disasters.

There is a significant difference between the property and casualty lines, although both are written by property-casualty companies. Property insurance reimburses policyholders directly for their insured losses, and thus is labeled as *first-party coverage*. Casualty insurance, for the most part, protects its policyholders against financial losses involving third parties, and consequently is called *third-party insurance*. For example, an automobile liability policy typically pays for damage caused by negligent action on the part of its insured that involves a vehicle owned by a third party. Other important casualty lines of insurance are those of general liability and workers' compensation.

Agents or brokers sell most insurance policies. An agent legally represents the insurer and has the authority to act on the company's behalf. In property insurance, independent agents can represent several companies, while exclusive agents represent only a single company. Insurance is also marketed by direct writers, who are employees of the insurer, and by mail-order companies. The emergence of

electronic commerce is making the World Wide Web an increasingly important source for marketing insurance. Brokers legally represent the insured, normally a corporation. They play a key role in commercial property insurance today, providing risk management and loss control services to companies as well as arranging their insurance contracts (*see also Nonlife Insurance Markets; Risk Classification in Nonlife Insurance*).

References

- [1] Covello, V. & Mumpower, J. (1985). Risk analysis and risk management: an historical perspective, *Risk Analysis* 5, 103–120.

- [2] Bainbridge, J. (1952). *Biography of an Idea: The Story of Mutual Fire and Casualty Insurance*, Doubleday, Garden City, NY.
- [3] Kunreuther, H. (1998). Insurability Conditions and the Supply of Coverage, *Paying the Price: The Status and Role of Insurance Against Natural Disasters in the United States*, H. Kunreuther & R. Roth, Sr, eds, Joseph Henry Press, Washington, DC, Chapter 2.

Related Articles

Nonlife Insurance Markets

Risk Classification in Nonlife Insurance

HOWARD KUNREUTHER

Health Hazards Posed by Dioxin

Since the late 1970s, regular meetings of international groups of scientists from nearly every developed country have been convened to discuss the developing scientific evidence on 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) exposure and toxicity (the 1986 Banbury Conference, and the Annual International Symposium on Halogenated Environmental Organic Pollutants and POPs). Most recently, a rodent cancer bioassay evaluating TCDD and 2,3,4,7,8-pentachlorodibenzofuran (PeCDF) has been completed [1] with alternative potency estimates made [2–4]; tolerable intake estimates have been established by several groups [5–8]; and several cancer potency estimates based on occupational epidemiology studies (*see Occupational Cohort Studies*) have been proposed [9–17, 18].

Hazard Identification

Hazard identification is the process of evaluating the available toxicity studies and determining the range of toxic endpoints relevant to humans, as well as identifying any significant data gaps relating to the primary studies relied upon in the derivation of the toxicity criteria (e.g., cancer potency factor or reference dose (RfD)) [19].

Noncancer Hazard at Low Doses

There are dozens of studies that address the toxicology or epidemiology of TCDD and some dioxin-like compounds [20–34]. Animal studies demonstrate a relatively diverse range of dose-dependent noncarcinogenic adverse responses to TCDD that vary considerably between species including wasting syndrome, reproductive toxicity, developmental effects, and commonly observed toxic effects on the liver, kidney, gastrointestinal tract, and certain endocrine organs [21, 34]. Chloracne has been observed in studies of humans with excessive exposure to TCDD from occupational or accidental contact [35–37]. Clinical tests have suggested subtle changes of metabolism, endocrine function, and developmental effects in humans [26, 33, 38], but such effects have not been

conclusively demonstrated at the low concentrations to which humans are routinely exposed.

Research organizations, regulatory agencies, and individual scientists have relied on measures of different kinds of toxicity to determine “safe” exposure levels. In general, scientists examine the toxicologic literature, identify the adverse effect relevant to humans that occurs at the lowest dose (*see Low-Dose Extrapolation*), and calculate the likelihood that adverse effect will occur at various exposure levels. For instance, Agency for Toxic Substances and Disease Registry (ATSDR) scientists relied upon developmental studies in rodents and monkeys to set its intermediate duration and chronic minimal risk levels (MRLs). ATSDR relied upon immunological toxicity studies in guinea pigs [39] for their intermediate MRL and developmental toxicity studies in monkeys [40] for their chronic MRL.

In a review of the TCDD literature concerning noncancer effects in animals and humans, Greene *et al.* [41] identified chloracne in children exposed during the Seveso trichlorophenol reactor explosion incident [42–46] as the best documented and lowest-dose disease endpoint in humans. Greene *et al.* [41] also identified developmental studies of TCDD as providing the best documented and lowest-dose toxicity endpoint in animals [41, 47–53]. Scientists at the World Health Organization [54] and the Joint FAO/WHO Expert Committee on Food Additives [31] relied on these same developmental toxicity studies to derive their tolerable intake estimates for TCDD. These scientists, like Greene *et al.* [41], believed their estimates of tolerable intake for the noncancer effects of TCDD would place the cancer hazard at negligible or tolerable levels [31]. WHO [31] identified tolerable estimates of intake for toxicity equivalents (TEQ) in the range of $1\text{--}5\text{ pg kg}^{-1}\text{ day}^{-1}$.

Interestingly, after a very long and exhaustive analysis, the United Nations working group on polychlorinated dibenzo-*p*-dioxins and furans (PCDD/Fs) reached the conclusion that doses in the current Western diet should not be expected to produce adverse health effects [55]. The current average intake of dioxin/furan TEQ in the United States diet is about $1\text{--}3\text{ pg kg}^{-1}\text{ day}^{-1}$ [31, 34]. Conversely, the US EPA has stated that they believe background doses due to diet for Americans are potentially hazardous [18]. A peer review of the so-called EPA reassessment of dioxin (a nearly 20-year-old process) by the

National Academy of Science within the United States recommended that the EPA update their current view of this chemical and place greater weight on describing the uncertainty in their predictions regarding the significance of the current risk of dietary exposure [56].

Cancer Hazard at Low Doses

TCDD has long been known as one of the most potent rodent carcinogens among the chemicals regulated by US EPA. Different researchers and regulatory scientists have calculated the cancer potency of TCDD to range from 40 to 1 400 000 (mg/kg/day)⁻¹ based on findings from the same animal study by Kociba *et al.* [57]. This 2-year cancer bioassay, which involved dietary ingestion of TCDD by rats, has been the basis for most of the published cancer potency estimates for TCDD and for the other 2,3,7,8-substituted PCDD/Fs (*via* the TCDD toxic equivalents approach) [58]. Different interpretations of the data correspond to use of different assumptions about mechanism of action (e.g., nongenotoxic or genotoxic), endpoints to be modeled (e.g., neoplastic nodules *versus* tumors *versus* carcinomas) and extrapolation models (e.g., linearized multistage *versus* safety factor approach) [59]. Recent developments include a further evaluation of the cancer potency of TCDD and 2,3,4,7,8-PeCDF in a chronic rodent bioassay [1], and proposals from US EPA [34] and others [9, 10, 12, 13] rely on occupational cancer epidemiology studies to define PCDD/F cancer potency.

International Agency for Research on Cancer (IARC) [20] evaluated the epidemiological literature on TCDD, noting the generally low magnitude of increased risk, the absence of any consistent pattern of site-specific cancer excess, and the lack of clear dose-response trends. The IARC workgroup classified TCDD as “carcinogenic to humans (Group 1)” on the basis of this limited epidemiological evidence, sufficient animal evidence, and the fact that the presumed mechanism (Ah receptor activation) is known to occur in humans and experimental animals. However, a direct correlation between Ah receptor binding affinity and cancer response has not been clearly demonstrated in animals or humans for TCDD and other PCDD/Fs, and significant quantitative and qualitative differences between humans and animals almost certainly exist [60].

The US EPA Science Advisory Board (SAB) that reviewed the US EPA “reassessment” in 2000 was fundamentally undecided about whether the available evidence was sufficient for TCDD to be considered a human carcinogen at any dose; at least half of the members concluded that the evidence was inadequate to support a “known human carcinogen” designation [61]. This group within the SAB offered several lines of support for its views. One included an analysis that reported that only 2 of the 12 key cohort studies (*see Cohort Studies*) had significantly elevated total cancer mortality rates, and there was a flat dose-response trend (i.e., no dose-response relationship) when cancer rates were plotted against average body burden level for the various cohorts [24, 30]. Moreover, several groups of workers who had chloracne (which likely requires peak blood lipid TCDD levels above 5000 ppt; [41]) did not have an increased cancer risk [62]).

Recent papers put into question the scientific foundation for US EPA’s use of worker epidemiology studies to define their proposed TCDD cancer potency factor of 1 000 000 (mg/kg/day)⁻¹ [63]. Aylward *et al.* [63] showed (as others have suggested in the past) that the human half-life of TCDD is much shorter at high dose levels, on the basis of a pharmacokinetic model of TCDD elimination in exposed Seveso residents and workers exposed to high levels of TCDD in a laboratory accident. This finding is significant because it invalidates the assumption in each of the epidemiology-based dose-response analyses (*see Dose-Response Analysis*) that body burdens of workers can be calculated using a constant half-life for TCDD. Aylward *et al.* [63] correctly note that if the TCDD half-life is shorter at high body burdens, determinations of dose based on calculations that incorporate a constant half-life are underestimated (and potency is thereby overestimated). This has enormous implications in the derivation of a cancer slope factor from these occupational studies. Until the half-life issue is addressed for these occupational cohorts, it doesn’t seem appropriate to these cancer potency estimates in quantitative risk assessments of TCDD.

PCDD/Fs are generally found in environmental media as chemical mixtures, and thus, humans are more likely exposed to mixtures rather than individual congeners. This is an obvious complication of the classic risk assessment paradigm because detailed dose-response studies are largely limited to TCDD.

To address this challenge, toxicity equivalency factors (TEFs) were developed to facilitate the quantification of PCDD/F dose and potential health risks. TEFs are relative potency factors assigned to each dioxin-like chemical to approximate the total mixture potency relative to the well-studied and reasonably well-understood toxicity of TCDD in experimental animals. The process involves assigning individual TEFs to each of the 17 2,3,7,8-chlorinated congeners. There is a scientific consensus on the general mechanism through which these congeners begin to exert their effects e.g., they first bind with the Ah receptor in the cytosol of a cell, the receptor–ligand complex then moves to the nucleus of a cell where it binds to dioxin response elements in the regulatory portion of genes. This concept is reflected in the current regulatory approach which relates the potency of other 2,3,7,8-substituted congeners to TCDD in comparable tests (TCDD, by definition, has a TEF of 1). The other 2,3,7,8-substituted congeners have TEF values ranging from 1 to 0.0001. Congeners without 2,3,7,8-substitution have been assigned TEF values of zero and are therefore not included in the analysis of TEQ. The most current TEF values were recently reported by the WHO following re-evaluation of the 1998-TEF values [64]. Despite a broad scientific consensus that use of this approach for risk assessment purposes is appropriate, there are substantial data gaps and scientific uncertainties associated with the use of the TEF approach [65, 66]. As such, some investigators have proposed utilizing distributions of relative potency values rather than point estimate TEFs [66, 67].

Although TCDD is the most widely studied of the PCDD/Fs, many studies have examined the toxicological properties of other congeners. The common underlying mechanism of action for all dioxin-like compounds is assumed to be that the chemical first binds to the Ah receptor [34, 58, 68, 69]. This assertion has been widely adopted for regulatory purposes to implicate all of the PCDD/Fs as multiorgan toxicants even though the evidence remains limited as to whether the non-TCDD congeners exhibit the same broad range of effects as TCDD. For example, IARC [20] concluded that there is sufficient evidence in animals and humans to designate TCDD as “carcinogenic to humans (Group 1)”, while all other PCDDs and PCDFs are “not classifiable as to their carcinogenicity to humans (Group 3)”. There is limited but growing evidence to support the assumption that other 2,3,7,8-substituted congeners have the capacity

to cause the effects that have been documented in animals treated with TCDD [2, 30, 70, 71]. However, the receptor-mediated mechanism of action for TCDD is subject to competitive inhibition by other dioxin-like congeners as well as other environmental chemicals with varying degrees of Ah receptor affinity [24, 27, 65, 69, 70, 72–79]. The potential impact of such inhibition should not be discounted, especially at low environmental doses at issue for risk assessment.

Others have suggested that tumor rates below control levels in the lowest TCDD dose group of the [57] study may indicate competitive inhibition of TCDD-induced response and/or a hormetic effect, i.e., depression of background cancer response at very low doses [61, 80–82]. Thus, the net effect of the usual low level exposure of humans to mixtures of dioxin-like compounds in the environment may present a much smaller human health hazard than that indicated by linear extrapolation models applied to animal studies of TCDD alone.

Noncancer Hazard Assessment

A number of different estimates of the so-called safe dose for noncancer effects have been published by regulatory agencies and other researchers over the past 20 years concerning PCDD/Fs. Table 1 provides a summary of the noncancer toxicity criteria that have been published since 1983. Most these estimates fall in the range of $1\text{--}5\text{ pg kg}^{-1}\text{ day}^{-1}$.

US EPA had previously established a noncancer RfD of $1\text{ pg kg}^{-1}\text{ day}^{-1}$ for TCDD. This value was withdrawn in 1989 and has now been replaced with a “margin of exposure (MOE)” approach which examines the source-related contribution to daily dose and/or body burden in comparison to background exposures and/or other no-effect-level or low-effect-level dose benchmarks [34, 83]. US EPA currently asserts that an appropriately defined RfD for TCDD would be of no practical benefit because this “safe” dose would fall below background exposure levels (i.e., below daily intake in the foods of those in western society who have no unusual source of exposure) [84]. For example, based on Portier [68], estimates of the effective dose (ED) of TCDD at 1% effect incidence (ED_{01}) are as low as $0.013\text{ pg kg}^{-1}\text{ day}^{-1}$; this corresponds to a steady-state human body burden of 0.025 ng kg^{-1} (ng of TCDD/kg of body weight). Following a trend of decrease over the past two decades,

Table 1 Procedures used by regulatory agencies or scientific bodies to estimate virtually safe or tolerable doses for humans based on dose–response data for noncancer effects of TCDD

Agency/group	Basis for extrapolation to humans	Acceptable intake rate
Japan, 1983 [84]	Yusho disease NOAEL in humans (1 ng kg ⁻¹ day ⁻¹) with 10-fold safety factor for sensitive humans	100 pg kg ⁻¹ day ⁻¹
Germany, 1985 [84]	Reproductive toxicity NOAEL = 1 ng kg ⁻¹ day ⁻¹ [94] with safety factor of 100–10000	1–10 pg kg ⁻¹ day ⁻¹
Nordic Group, 1987 [85]	Reproductive toxicity NOAEL = 1 ng kg ⁻¹ day ⁻¹ [94] with safety factor of 100	10 pg kg ⁻¹ day ⁻¹
United States, 1989 [86]	Reproductive toxicity NOAEL = 1 ng kg ⁻¹ day ⁻¹ [94] with safety factor of 1000	<i>RfD</i> = 1 pg kg ⁻¹ day ⁻¹
World Health Organization, 1990 [87]	Combined consideration of cancer, liver toxicity, reproductive and immune toxicity NOAELs = 1 ng kg ⁻¹ day ⁻¹ with 100-fold safety factor. Also adopted by UK, New Zealand, and the Netherlands	10 pg kg ⁻¹ day ⁻¹
United States, 1992 [88]	Reproductive toxicity NOAEL = 1 ng kg ⁻¹ day ⁻¹ [94] with 1000 safety factor for chronic/intermediate minimal risk level (MRL)	Chronic MRL = 1 pg kg ⁻¹ day ⁻¹ Intermediate MRL = 1 pg kg ⁻¹ day ⁻¹
The Netherlands, 1996 [89]	Reproductive toxicity LOAEL = 0.1 ng kg ⁻¹ day ⁻¹ in monkey studies with 100-fold safety factor	1 pg kg ⁻¹ day ⁻¹
Japan, 1997 [90]	Combined consideration of reproductive toxicity in monkeys [95] and carcinogenicity	5 pg kg ⁻¹ day ⁻¹
ATSDR, 1998 [22, 23, 91, 92]	Chronic MRL: Reproductive toxicity in monkeys with 120-fold safety factor applied to LOAEL [40] Intermediate MRL: 90-day immunotoxicity study in guinea pigs with 30-fold safety factor [39]	Chronic MRL = 1 pg kg ⁻¹ day ⁻¹ Intermediate MRL = 20 pg kg ⁻¹ day ⁻¹
U.S. EPA, 2000 [68]	ED-01 body burden of 0.025 ng kg ⁻¹ in rats based on [47] sperm effects, converted to human daily dose assuming 50% bioavailability and 7.6-year half-life	0.013 pg kg ⁻¹ day ⁻¹ margin of exposure approach: < 0.1 pg kg ⁻¹ day ⁻¹ including background
World Health Organization, 2000 [54]	Reproductive toxicity in rats with 10-fold safety factor applied to LOAEL [48, 49, 96, 97] calculated from maternal body burden with half-life of 8.5 years	Tolerable daily intake: 1–4 pg kg ⁻¹ day ⁻¹
European Commission, 2001 [93]	Reproductive toxicity in rats with 9.6-fold safety factor applied to NOAEL for male rat offspring [51, 53] calculated from maternal body burden with half-life of 7.6 years	Tolerable weekly intake: 14 pg/kg/week or 2 pg kg ⁻¹ day ⁻¹
United Kingdom, 2001 [8]	Reproductive toxicity in rats with 9.6-fold safety factor applied to NOAEL for male rat offspring [51] calculated from maternal body burden with half-life of 7.5 years	Tolerable daily intake: 2 pg kg ⁻¹ day ⁻¹
World Health Organization (JECFA), 2001 [31]	Reproductive toxicity in rats with 9.6-fold safety factor applied to NOAEL for male rat offspring [51, 53] calculated from maternal body burden with 7.6-year half-life.	Provisional tolerable monthly intake: 70 pg/kg/month or 2.3 pg kg ⁻¹ day ⁻¹

current background human body burdens of TCDD appear to average around 3 ppt in blood lipid or about 0.75 ng/kg of body weight for a 60-kg person with 25% body fat, with total lipid TEQ from PCDD/Fs being about 15 to 30 from a public health to set an RfD that is well below average body burdens of TCDD and total TCDD toxic equivalents for the general public.

The issue of background dietary exposures becomes a potentially important concern for defining a proper noncancer toxicity criterion. The RfD criterion is defined as a safe dose level determined by taking a no-effect level or a low-effect level defined in human or animal studies and dividing that level by appropriate safety/uncertainty factors to arrive at a conservative level that can be compared to doses received from a particular source, e.g., contaminated soils in a residential area.

US EPA [34] asserts that the low-effect levels observed in certain animal studies could plausibly occur at PCDD/F body burdens within 10- to 100-fold of the TEQ average in the background population. However, there is no confirmation that humans experience such effects, even in studies of humans with much higher body burdens. Several reviewers have pointed to the inconclusive nature of the studies that US EPA [98] cited as evidence of human effects of PCDD/Fs at doses or body burdens near background levels [22, 23, 25, 26, 28, 30, 32, 33, 38, 99]. The absence of findings confirming excess disease in humans at or near background body burdens may reflect the difficulties in finding proper control (unexposed) populations to distinguish them. However, it seems equally likely that, given the limited range of toxicity in humans at very high doses, alternative explanations may apply to risks at or near background body burdens of PCDD/Fs. These may include the following:

1. Humans may be less sensitive compared with test species in regards to experiencing the adverse effects under study.
2. The steady-state body burden of TCDD may not be an appropriate dose metric for comparisons between these animal studies and humans.
3. Studies of TCDD alone in animals may not be representative of the human population which experiences predominant exposures to PCDD/F mixtures of which TCDD is only a small fraction (e.g., competitive inhibition of Ah receptor).

4. It may not be possible to extrapolate higher-dose studies in animals to humans in a meaningful way because of a combination of factors which may include nonlinear (threshold-dependent) dose-response relationships, predominant influences of competitive inhibition of Ah receptor activation from environmental mixtures [78], and/or hormetic effects of background environmental exposures in humans [82] that do not occur at higher/TCDD-only doses in animals.

Regardless of these considerations, the US EPA withdrew its original RfD of $1 \text{ pg kg}^{-1} \text{ day}^{-1}$ for TCDD and proposed a MOE approach as an alternative to evaluating noncancer risks. The MOE is calculated by dividing a “point of departure” for extrapolation purposes at the low end of the range of observation in human or animal studies (e.g., the ED_{01}) by the human exposure or body burden of interest (predicted dose). They propose that MOE values in excess of 100–1000 for background plus site-related TEQ doses “are adequate to rule out the likelihood of significant effects occurring in humans based on sensitive animal responses or results from epidemiologic studies”. However, the practical application of this approach is hampered by many variables and uncertainties that will potentially take many years to sort out, including the reliable estimation of background exposures and the validity of many possible choices for the “point of departure” to be assessed [100–102]. Others have proposed that pharmacokinetic models similar to those used for risk assessment of lead exposures may be appropriate for dioxin-like compounds [103, 104]. MOE analysis was not done in this assessment, instead practical surrogate values for TCDD RfD were used to evaluate noncancer hazard.

International regulatory authorities have expressed noncancer toxicity criteria as tolerable daily intake (TDI) [54], tolerable weekly intake (TWI) [93], or provisional tolerable monthly intake (PTMI) [31] expressed as a single value or a range (Table 1). These values are based on scientific panel reviews of the currently available literature on TCDD and other PCDD/Fs. These tolerable intake estimates are considered by these authorities to be protective against cancer and noncancer health effects of PCDD/Fs.

ATSDR’s estimated safe dose for noncancer effects of TCDD consider duration of exposure. ATSDR has established acute, intermediate, and

chronic MRLs. An MRL is defined as “an estimate of the daily human exposure to a hazardous substance that is likely to be without appreciable risk of adverse noncancer effects over a specified duration of exposure”. The scientific basis for prior TCDD MRLs [88] and current MRLs [21] has been reviewed previously [22, 23]. The MRL values for acute ($200 \text{ pg kg}^{-1} \text{ day}^{-1}$) and intermediate ($20 \text{ pg kg}^{-1} \text{ day}^{-1}$) exposures are higher now (less conservative) than they were on the basis of the available studies in ATSDR’s earlier assessment [88]. Indeed, the current MRL for intermediate oral exposures is 20-fold higher than the 1992 value, whereas the chronic MRL ($1 \text{ pg kg}^{-1} \text{ day}^{-1}$) remained consistent with the former chronic MRL and the former US EPA RfD for TCDD [105]. ATSDR explained that these MRL changes are the result of greater availability of human studies and animal studies that addressed certain uncertainties in the earlier MRL determinations [22, 23]. This reduced uncertainty translated into plausible justifications for the use of smaller safety margins on the MRL [22, 23, 32].

Closing

All in all, in spite of the hundreds of millions of dollars in research devoted to studying PCDD/PCDFs over the past 30 years, the ability of the scientific community to quantitatively characterize the risks to humans of current environmental doses is not nearly as certain as one would expect. The precise mechanism through which the chemical acts in humans is still not well understood and the role of other chemicals that compete for the Ah receptor remains unclear with respect to predicting both the cancer and non-cancer risk at very low doses. One is hard pressed to identify another chemical which has been so difficult to characterize in the history of toxicology and pharmacology research.

References

- [1] National Toxicology Program (2004). Draft NTP technical report on the toxicology and carcinogenesis studies of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) (CAS No. 1746-01-6) in female Harlan Sprague-Dawley rats (Gavage study), NTP TR 521, NIH Publication No. 04-4455, U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health.
- [2] Walker, N.J., Crockett, P.W., Nyska, A., Brix, A.E., Jokinen, M.P., Sells, D.M., Hailey, J.R., Easterling, M., Haseman, J.K., Yin, M., Wyde, M.E., Bucher, J.R. & Portier, C.J. (2005). Dose-additive carcinogenicity of a defined mixture of “dioxin-like compounds”, *Environmental Health Perspectives* **113**(1), 43–48.
- [3] Popp, J.A., Crouch, E. & McConnell, E.E. (2006). A weight-of-evidence analysis of the cancer dose-response characteristics of 2,3,7,8-tetrachlorodibenzo dioxin (TCDD), *Toxicological Sciences* **89**(2), 361–369.
- [4] Faust, J.B. & Zeise, L. (2004). A reassessment of the carcinogenic potency of 2,3,7,8-tetrachlorodibenzo-p-dioxin based upon a recent long-term study in female rats, *Proceedings of the Society for Risk Analysis Annual Meeting*, Palm Springs, December 5–8, 2004.
- [5] van Leeuwen, F.X., Feeley, M., Schrenk, D., Larsen, J.C., Farland, W. & Younes, M. (2000). Dioxins: WHO’s tolerable daily intake (TDI) revisited, *Chemosphere* **40**(9–11), 1095–1101.
- [6] Schecter, A., Cramer, P., Boggess, K., Stanley, J., Papke, O., Olson, J., Silver, A. & Schmitz, M. (2001). Intake of dioxins and related compounds from food in the U.S. population, *Journal of Toxicology and Environmental Health, Part A* **63**(1), 1–18.
- [7] European Commission Scientific Committee on Food (2001). *Opinion of the Scientific Committee on Food on the Risk Assessment of Dioxins and Dioxin-like Food*, Health and Consumer Protection Directorate General, Brussels, CS/CNTM/DIOXIN/20final, at http://europa.eu.int/comm/food/fs/sc/scf/reports_en.html.
- [8] CoT (2001). *Committee on Toxicity and Chemicals in Food, Consumer Products and the Environment, Statement on the Tolerable Daily Intake for Dioxins and Dioxin-like Polychlorinated Biphenyls*, COT/2001/07, at <http://www.food.gov.uk/science/ouradvisors/toxicity/statements/coststatements2001/dioxinsstate>.
- [9] Becher, H. & Flesch-Janys, D. (1998). Dioxins and furans: epidemiologic assessment of cancer risks and other human health effects, *Environmental Health Perspectives* **106**(Suppl. 2), 623–624.
- [10] Becher, H., Steindorf, K. & Flesch-Janys, D. (1998). Quantitative cancer risk assessment for dioxins using an occupational cohort, *Environmental Health Perspectives* **106**(Suppl. 2), 663–670.
- [11] Starr, T.B. (2001). Significant shortcomings of the U.S. environmental protection agency’s latest draft risk characterization for dioxin-like compounds, *Toxicological Sciences* **64**(1), 7–13.
- [12] Steenland, K., Calvert, G., Ketchum, N. & Michalek, J. (2001). Dioxin and diabetes mellitus: an analysis of the combined NIOSH and ranch hand data, *Occupational and Environmental Medicine* **58**(10), 641–648.
- [13] Steenland, K., Deddens, J. & Piacitelli, L. (2001). Risk assessment for 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) based on an epidemiologic study, *American Journal of Epidemiology* **154**(5), 451–458.

- [14] Crump, K.S., Canady, R. & Kogevinas, M. (2003). Meta-analysis of dioxin cancer dose response for three occupational cohorts, *Environmental Health Perspectives* **111**(5), 681–687.
- [15] Aylward, L.L., Brunet, R.C., Carrier, G., Hays, S.M., Cushing, C.A., Needham, L.L., Patterson, D.G., Gerthoux, P.M., Brambilla, P. & Mocarelli, P. (2005). Concentration-dependent TCDD elimination kinetics in humans: toxicokinetic modeling for moderately to highly exposed adults from Seveso, Italy, and Vienna, Austria, and impact on dose estimates for the NIOSH cohort, *Journal of Exposure Analysis and Environmental Epidemiology* **15**(1), 51–65.
- [16] Aylward, L.L., Brunet, R.C., Starr, T.B., Carrier, G., Delzell, E., Cheng, H. & Beall, C. (2005). Exposure reconstruction for the TCDD-exposed NIOSH cohort using a concentration- and age-dependent model of elimination, *Risk Analysis* **25**(4), 945–956.
- [17] Aylward, L.L., Lamb, J.C. & Lewis, S.C. (2005). Issues in risk assessment for developmental effects of 2,3,7,8-tetrachlorodibenzo-p-dioxin and related compounds, *Toxicological Sciences* **87**(1), 3–10.
- [18] U.S. Environmental Protection Agency (2003). *Draft Exposure and Human Health Risk Assessment of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) and Related Compounds, Parts I, II, and III*, Washington, DC, at <http://www.epa.gov/ncea/pdfs/dioxin/nas-review/>.
- [19] Williams, P. & Paustenbach, D.J. (2002). Risk characterization: principles and practice, *Journal of Toxicology and Environmental Health, Part B: Critical Reviews* **5**(4), 337–406.
- [20] IARC (1997). Polychlorinated dibenzo-para-dioxins and polychlorinated dibenzofurans, *IARC Monograph Evaluating Carcinogen Risks in Humans*.
- [21] Agency for Toxic Substances and Disease Registry (ATSDR) (1998). *Toxicological Profile for Chlorinated Dibenzo-p-dioxins*, Atlanta.
- [22] De Rosa, C.T., Brown, D., Dhara, R., Garrett, W., Hansen, H., Holler, J., Jones, D., Jordan-Izaguirre, D., O'Conner, R., Pohl, H. & Xintaras, C. (1999). Dioxin and dioxin-like compounds in soil, part II: technical support document for ATSDR policy guideline, *Toxicology and Industrial Health* **6**, 558–576.
- [23] De Rosa, C.T., Brown, D., Dhara, R., Garrett, W., Hansen, H., Holler, J., Jones, D., Jordan-Izaguirre, D., O'Conner, R., Pohl, H. & Xintaras, C. (1999). Dioxin and dioxin-like compounds in soil, part I: ATSDR policy guideline, *Toxicology and Industrial Health* **15**(6), 552–557.
- [24] Adami, H.O., Cole, P., Mandel, J., Pastides, H., Starr, T.B. & Trichopoulos, D. (2000). Dioxin and cancer, *CCC Sponsored Dioxin Workshop*, IS RTP and the American Bar Association.
- [25] Feeley, M. & Brouwer, A. (2000). Health risks to infants from exposure to PCBs, PCDDs and PCDFs, *Food Additives and Contaminants* **17**(4), 325–333.
- [26] Sweeney, M.H. & Mocarelli, P. (2000). Human health effects after exposure to 2,3,7,8-TCDD, *Food Additives and Contaminants* **17**, 303–316.
- [27] Adami, H.O., Day, N.E., Trichopoulos, D. & Willett, W.C. (2001). Primary and secondary prevention in the reduction of cancer morbidity and mortality, *European Journal of Cancer* **37**(Suppl. 8), S118–S127.
- [28] Kogevinas, M. (2001). Human health effects of dioxins: cancer, reproductive and endocrine system effects, *Human Reproduction Update* **7**(3), 331–339.
- [29] Smith, A.H. & Lopipero, P. (2001). Invited commentary: how do the Seveso findings affect conclusions concerning TCDD as a human carcinogen? *American Journal of Epidemiology* **153**, 1045–1047.
- [30] Starr, T.B. (2001). Significant shortcomings of the U.S. environmental protection agency's latest draft risk characterization for dioxin-like compounds, *Toxicological Sciences* **64**, 7–13.
- [31] WHO (2001). *Joint FAO/WHO Expert Committee on Food Additives (JECFA)*, Food and Agriculture Organization of the United Nations, Rome.
- [32] Pohl, H., Hicks, H.E., Jones, D.E., Hansen, H. & De Rosa, C.T. (2002). Public health perspectives on dioxin risks: two decades of evaluations, *Human and Ecological Risk Assessment* **8**(2), 233–250.
- [33] Greene, J.F., Hays, S.M. & Paustenbach, D. (2003). Basis for a proposed reference dose (RfD) for dioxin of 1–10 pg/kg-day: a weight of evidence evaluation of the human and animal studies, *Journal of Toxicology and Environmental Health, Part B* **6**, 115–159.
- [34] USEPA (2003). *Draft Exposure and Human Health Risk Assessment of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) and Related Compounds, Parts I, II and III*, Office of Research and Development, National Center for Environmental Assessment, Exposure Assessment and Risk Characterization Group, Washington, DC.
- [35] Poland, A.P., Smith, D. & Metter, G. (1971). A health survey of workers in a 2,4-D and 2,4,5-T plant with special attention to chloracne, porphyria cutanea tarda, and psychologic parameters, *Archives of Environmental Health* **22**(3), 316–327.
- [36] Caramaschi, F., del Corno, G., Favaretti, C., Giambelluca, S.E., Montesarchio, E. & Fara, G.M. (1981). Chloracne following environmental contamination by TCDD in Seveso, Italy, *International Journal of Epidemiology* **10**(2), 135–143.
- [37] Baccarelli, A., Pesatori, A.C., Consonni, D., Mocarelli, P., Patterson, D.G., Caporaso, N.E., Bertazzi, P.A. & Landi, M.T. (2005). Health status and plasma dioxin levels in chloracne cases 20 years after the Seveso, Italy accident, *British Journal of Dermatology* **152**(3), 459–465.
- [38] Bertazzi, P.A., Pesatori, A.C. & Zocchetti, C. (1998). Seveso-dioxin: an example of environmental medicine. Epidemiologic data as guidelines for health programming, *Giornale Italiano di Medicina del Lavoro ed Ergonomia* **20**, 194–196.

- [39] DeCaprio, A.P., McMartin, D.N., O'Keefe, P.W., Rej, R., Silkworth, J.B. & Kaminsky, L.S. (1986). Subchronic oral toxicity of 2,3,7,8-tetrachlorodibenzo-p-dioxin in the guinea pig: comparisons with a PCB-containing transformer fluid pyrolysate, *Fundamental and Applied Toxicology* **6**, 454–463.
- [40] Schantz, S.L., Ferguson, S.A. & Bowman, R.E. (1992). Effects of 2,3,7,8-tetrachlorodibenzo-p-dioxin on behavior of monkey in peer groups, *Neurotoxicology and Teratology* **14**, 433–446.
- [41] Greene, J.F., Hays, S. & Paustenbach, D. (2003). Basis for a proposed reference dose (RfD) for dioxin of 1–10 pg/kg-day: a weight of evidence evaluation of the human and animal studies, *Journal of Toxicology and Environmental Health, Part B: Critical Reviews* **6**(2), 115–159.
- [42] Reggiani, G. (1978). Medical problems raised by the TCDD contamination in Seveso, Italy, *Archives of Toxicology* **40**, 161–188.
- [43] Caramaschi, F., del Corno, G., Favaretti, C., Giambelluca, S.E., Montesarchio, E. & Fara, G.M. (1981). Chloracne following environmental contamination by TCDD in Seveso, Italy, *International Journal of Epidemiology* **10**, 135–143.
- [44] Ideo, G., Bellati, D., Bellobuono, A. & Bissanti, L. (1985). Urinary D-glucaric acid excretion in the Seveso area, polluted by tetrachlorodibenzo-p-dioxin (TCDD): five years of experience, *Environmental Health Perspectives* **60**, 151–157.
- [45] Mocarelli, P., Marocchi, A., Brambilla, P., Gerthoux, P.M., Young, D.S. & Mantel, N. (1986). Clinical laboratory manifestations of exposure to dioxin in children. A six-year study of the effects of an environmental disaster near Seveso, Italy, *Journal of the American Medical Association* **256**, 2687–2695.
- [46] Assennato, G., Cervino, D., Emmett, E.A., Longo, G. & Merlo, F. (1989). Follow-up of subjects who developed chloracne following TCDD exposure at Seveso, *American Journal of Industrial Medicine* **16**, 119–125.
- [47] Mably, T.A., Bjerke, D.L., Moore, R.W., Gendron-Fitzpatrick, A. & Peterson, R.E. (1992). In utero and lactational exposure of male rats to 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD). 3. Effects on spermatogenesis and reproductive capability, *Toxicology and Applied Pharmacology* **114**, 118–126.
- [48] Gray, L.E., Ostby, J.S. & Kelce, W.R. (1997). A dose-response analysis of the reproductive effects of a single gestational dose of 2,3,7,8-tetrachlorodibenzo-p-dioxin in male Long Evans hooded rat offspring, *Toxicology and Applied Pharmacology* **146**, 11–20.
- [49] Gray, L.E., Wolf, C., Mann, P. & Ostby, J.S. (1997). In utero exposure to low doses of 2,3,7,8-tetrachlorodibenzo-p-dioxin alters reproductive development of female Long Evans hooded rat offspring, *Toxicology and Applied Pharmacology* **146**, 237–244.
- [50] Faqi, A.S. & Chahoud, I. (1998). Antiestrogenic effects of low doses of 2,3,7,8-TCDD in offspring of female rats exposed throughout pregnancy and lactation, *Bulletin of Environmental Contamination and Toxicology* **61**, 462–469.
- [51] Faqi, A.S., Dalsenter, P.R., Merker, H.J. & Chahoud, I. (1998). Reproductive toxicity and tissue concentrations of low doses of 2,3,7,8-tetrachlorodibenzo-p-dioxin in male offspring rats exposed throughout pregnancy and lactation, *Toxicology and Applied Pharmacology* **150**, 383–392.
- [52] Ostby, J.S., Price, M., Huey, O., Hurst, C., Birnbaum, L. & Gray Jr, L.E. (1999). Developmental and reproductive effects of low-dose, steady-state maternal 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) administration, *Society of Toxicology 38th Annual Meeting*, March 14–18, 1999, New Orleans.
- [53] Ohsako, S., Miyabara, Y., Nishimura, N., Kurosawa, S., Sakaue, M., Ishimura, R., Sato, M., Takeda, K., Aoki, Y., Sone, H., Tohyama, C. & Yonemoto, J. (2001). Maternal exposure to a low dose of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) suppressed the development of reproductive organs of male rats: dose-dependent increase of mRNA levels of 5-alpha-reductase type 2 in contrast to decrease of androgen receptor in the pubertalventral prostate, *Toxicological Sciences* **60**, 132–143.
- [54] WHO (2000). Assessment of the health risk of dioxins: re-evaluation of the tolerable daily intake (TDI), in *Organized by WHO European Centre for Environment and Health and International Programme on Chemical Safety, Food Additives and Contaminants*, F.X.R. van Leeuwen & M.M. Younes, eds, Taylor & Francis, London, Vol. 17, pp. 233–369.
- [55] Renwick, A. (2004). Recent risk assessments of dioxin, *24th International Symposium on Halogenated Environmental Organic Pollutants and POPs*, Berlin.
- [56] NAS (2006). *Health Risks from Dioxin and Related Compounds: Evaluation of the EPA Reassessment*, National Academy Press, Washington, DC.
- [57] Kociba, R.J., Keyes, D.G., Lisowe, R.W., Kalnins, R.P., Dittenber, D.D., Wade, C.E., Gorzinski, S.J., Mahle, N.H. & Schwetz, B.A. (1978). Results of a two-year chronic toxicity and oncogenicity study of 2,3,7,8-tetrachlorodibenzo-p-dioxin in rats, *Toxicology and Applied Pharmacology* **46**(2), 279–303.
- [58] van den Berg, M., Birnbaum, L., Bosveld, A.T.C., Brunstrom, B., Cook, P., Feeley, M., Giesy, J.P., Hanberg, A., Hasegawa, R., Kennedy, S.W., Kubiak, T., Larsen, J.C., van Leeuwen, F.X.R., Liem, A.K.D., Nolt, C., Peterson, R.E., Poellinger, L., Safe, S., Schrenk, D., Tillitt, D., Tysklind, M., Younes, M., Waern, F. & Zacharewski, T. (1998). Toxic equivalency factors (TEFs) for PCBs, PCDDs, PCDFs for humans and wildlife, *Environmental Health Perspectives* **106**(12), 775–792.
- [59] Sielken, R.L. (1987). Quantitative cancer risk assessments for 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD), *Food and Chemical Toxicology* **25**(3), 257–267.

- [60] Connor, K. & Aylward, L.L. (2005). Human response to dioxin: aryl hydrocarbon receptor (AhR) molecular structure, function, and dose-response data for enzyme induction indicate an impaired human AhR, *Journal of Toxicology and Environmental Health, Part B* **9**(2), 147–171.
- [61] Paustenbach, D.J. (2002). The USEPA science advisory board evaluation of the EPA (2001) dioxin reassessment, *Regulatory Toxicology and Pharmacology* **36**, 211–219.
- [62] Bodner, K.M., Collins, J.J., Bloemen, L.J. & Carson, M.L. (2003). Cancer risk for chemical workers exposed 2,3,7,8-tetrachlorodibenzo-p-dioxin, *Occupational and Environmental Medicine* **60**(9), 672–675.
- [63] Aylward, L.L., Brunet, R.C., Carrier, G., Hays, S.M., Cushing, C.A., Needham, L.L., Patterson, D.G., Gerthoux, P.M., Brambilla, P. & Mocarelli, P. (2005). Concentration-dependent TCDD elimination kinetics in humans: toxicokinetic modeling for moderately to highly exposed adults from Seveso, Italy, and Vienna, Austria, and impact on dose estimates for the NIOSH cohort, *Journal of Exposure Analysis and Environmental Epidemiology* **15**(1), 51–65.
- [64] van den berg, M., Birnbaum, L.S., Denison, M., De Vito, M., Farland, W., Feeley, M., Fiedler, H., Hakanson, H., Hanberg, A., Haws, L., Rose, M., Safe, S., Schrenk, D., Toyama, C., Tritscher, A., Tuomisto, J., Tysklind, M., Walker, N. & Peterson, R.E. (2006). The 2005 World Health Organization reevaluation of human and mammalian toxic equivalency factors for dioxins and dioxin-like compounds, *Toxicological Sciences* **93**(2), 223–241.
- [65] Safe, S. (1998). Limitations of the toxic equivalency factor approach for risk assessment of TCDD and related compounds, *Teratogenesis, Carcinogenesis, and Mutagenesis* **17**, 285–304.
- [66] Finley, B.L., Connor, K.T. & Scott, P.K. (2003). The use of toxic equivalency factor distributions in probabilistic risk assessments for dioxins, furans, and PCBs, *Journal of Toxicology and Environmental Health* **66**(6), 533–550.
- [67] Haws, L., Harris, M., Su, S., Birnbaum, L., DeVito, M.J., Farland, W., Walker, N., Connor, K., Santamaria, A. & Finley, B. (2004). Development of a refined database of relative potency estimates to facilitate better characterization of variability and uncertainty in the current mammalian TEFs for PCDDs, PCDFs and dioxin-like PCBs, *Organohalogen Compounds* **66**, 3426–3432.
- [68] Portier, C. (2000). Risk ranges for various endpoints following exposure to 2,3,7,8-TCDD, *Food Additives and Contaminants* **17**(4), 335–346.
- [69] Safe, S. (2001). Molecular biology of the Ah receptor and its role in carcinogenesis, *Toxicology Letters* **120**(1–3), 1–7.
- [70] Starr, T.B., Greenlee, W.F., Neal, R.A., Poland, A. & Sutter, T.R. (1999). The trouble with TEFs, *Environmental Health Perspectives* **107**, A492–A493.
- [71] Yoshizawa, K., Walker, N.J., Jokinen, M.P., Brix, A.E., Sells, D.M., Marsh, T., Wyde, M.E., Orzech, D., Haseman, J.K. & Nyska, A. (2005). Gingival carcinogenicity in female Harlan Sprague-Dawley rats following two-year oral treatment with 2,3,7,8-tetrachlorodibenzo-p-dioxin and dioxin-like compounds, *Toxicological Sciences* **83**(1), 64–77.
- [72] Bannister, R., Davis, D., Zacharewski, T., Tizard, I. & Safe, S. (1987). Aroclor 1254 as a 2,3,7,8-tetrachlorodibenzo-p-dioxin antagonist: effect on enzyme induction and immunotoxicity, *Toxicology* **46**(1), 29–42.
- [73] Bannister, R., Biegel, L., Davis, D., Astroff, B. & Safe, S. (1989). 6-Methyl-1,3,8-trichlorodibenzofuran (MCDF) as a 2,3,7,8-tetrachlorodibenzo-p-dioxin antagonist in C57BL/6 mice, *Toxicology and Applied Pharmacology* **54**(2), 139–150.
- [74] Biegel, L., Harris, M., Davis, D., Rosengren, R., Safe, L. & Safe, S. (1989). 2,2',4,4',5,5'-Hexachlorobiphenyl as a 2,3,7,8-tetrachlorodibenzo-p-dioxin antagonist in C57BL/6J mice, *Toxicology and Applied Pharmacology* **97**(3), 561–571.
- [75] Davis, D. & Safe, S. (1989). Dose-response immunotoxicities of commercial polychlorinated biphenyls (PCBs) and their interaction with 2,3,7,8-tetrachlorodibenzo-p-dioxin, *Toxicology Letters* **48**(1), 35–43.
- [76] Davis, D. & Safe, S. (1990). Immunosuppressive activities of polychlorinated biphenyl in C57BL/6N mice: structure-activity relationships as Ah receptor agonists and partial antagonists, *Toxicology* **63**(1), 97–111.
- [77] Harper, N., Connor, K., Steinberg, M. & Safe, S. (1995). Immunosuppressive activity of polychlorinated biphenyl mixtures and congeners: nonadditive (antagonistic) interactions, *Fundamental and Applied Toxicology* **27**(1), 131–139.
- [78] Safe, S. (1998). Hazard and risk assessment of chemical mixtures using the toxic equivalency factor approach, *Environmental Health Perspectives* **106**(Suppl. 4), 1051–1058.
- [79] Connor, K., Harris, M., Edwards, M., Chu, A., Clark, G. & Finley, B. (2004). Estimating the total TEQ in human blood from naturally-occurring vs. anthropogenic dioxins: a dietary study, *Organohalogen Compounds* **66**, 3360–3365.
- [80] Calabrese, E. & Baldwin, L.A. (2001). The frequency of U-shaped dose responses in the toxicological literature, *Toxicological Sciences* **62**, 330–338.
- [81] Calabrese, E. & Baldwin, L.A. (2001). Scientific foundations of hormesis, *Critical Reviews in Toxicology* **31**, 4–5.
- [82] Calabrese, E. (2002). Defining hormesis, *Human and Experimental Toxicology* **21**, 91–97.
- [83] USEPA (2000). *Draft Exposure and Human Health Risk Assessment of 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD) and Related Compounds, Parts I, II and III*, Office of Research and Development, National Center for Environmental Assessment, Exposure Assessment and Risk Characterization Group, Washington, DC.

- [84] Larsen, J.C., Farland, W. & Winters, D. (2000). Current risk assessment approaches in different countries, *Food Additives and Contaminants* **17**(4), 359–369.
- [85] Ahlborg, U., Hakansson, H., Wern, F. & Hanberg, A. (1988). Nordisk dioxin risk bedöorning, Rapport från en expertgroup, Nord 4:9(Miljörapport):7, Nordisk Ministerråd, København.
- [86] Environmental Protection Agency (1989). *Integrated Risk Information System (IRIS) On-line Database Entry for 2,3,7,8-tetrachlorodibenzo-p-dioxin*, January 1.
- [87] World Health Organization (1991). *Consultation on Tolerable Daily Intake from Food of PCDDs and PCDFs*, Copenhagen, Summary Report, Bilthoven, EUR/IPC/PCS 030(S), December 4–7, 1990.
- [88] Agency for Toxic Substances and Disease Registry (ATSDR) (1992). *Toxicological Profile for Chlorinated Dibenzo-p-dioxins*, Atlanta.
- [89] Health Council of the Netherlands (1996). Committee on the risk evaluation of substances dioxins, *Dioxins, Polychlorinated Dibenzo-p-dioxins Dibenzofurans and Dioxin-like Polychlorinated Biphenyls*, Publication No: 1996/10, Rijswijks.
- [90] Environmental Agency Japan (1997). *Report of Ad Hoc Committee on Dioxin Risk Assessment (Summary in English)*, at <http://www.chem.unep.ch/pops/POPs Inc/proceedings/bangkok/KIMURA.html>.
- [91] Agency for Toxic Substances and Disease Registry (1997). *Interim Policy Guideline: Dioxin and Dioxin-like Compounds in Soil*, Atlanta.
- [92] Agency for Toxic Substances and Disease Registry (1997). *Technical Support Document for ATSDR Interim Policy Guideline: Dioxin and Dioxin-like Compounds in Soil*, Atlanta.
- [93] European Commission, SCoFS (2001). *Opinion of the Scientific Committee on Food on the Risk Assessment of Dioxins and Dioxin-Like Food*, European Commission, Health and Consumer Protection Directorate General, Brussels.
- [94] Murray, F.J., Smith, F.A., Nitschke, K.D., Humiston, C.G., Kociba, R.J. & Schwetz, B.A. (1979). Three generation reproduction study of rats given 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) in the diet, *Toxicology and Applied Pharmacology* **50**, 241–251.
- [95] Rier, S.E., Martin, D.C., Bowman, R.E., Dmowski, W.P. & Becker, J.L. (1993). Endometriosis in rhesus monkeys (*Macaca mulatta*) following chronic exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin, *Fundamental and Applied Toxicology* **21**, 433–441.
- [96] Gehrs, B.C. & Smialowicz, R.J. (1997). Alterations in the developing immune system of the F344 rat after perinatal exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin. I. Effects on the fetus and the neonate, *Toxicology* **122**, 219–228.
- [97] Gehrs, B.C., Riddlge, M.M., Williams, W.C. & Smialowicz, R.J. (1997). Alterations in the developing immune systems of the F344 rat after perinatal exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin. II. Effects on the pup and the adult, *Toxicology* **122**, 229–240.
- [98] U.S. Environmental Protection Agency (2000). *Draft Exposure and Human Health Risk Assessment of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) and Related Compounds, Parts I, II, and III*, USEPA/600/P-00/001Bg, Washington, DC, at <http://cfpub.epa.gov/ncea/cfm/part1and2.cfm?AcType=default>.
- [99] Cole, P., Trichopoulos, D., Pestides, H., Starr, T.B. & Mandel, J.S. (2003). Dioxin and cancer: a critical review, *Regulatory Toxicology and Pharmacology* **38**, 378–388.
- [100] Aylward, L.L. & Hays, S.M. (2002). Temporal trends in human TCDD body burden: decreases over three decades and implications for exposure levels, *Journal of Exposure Analysis and Environmental Epidemiology* **12**(5), 319–328.
- [101] Gaylor, D.W. & Aylward, L.L. (2004). An evaluation of benchmark dose methodology for non-cancer continuous-data health effects in animals due to exposures to dioxin (TCDD), *Regulatory Toxicology and Pharmacology* **40**(1), 9–17.
- [102] U.S. Environmental Protection Agency-Science Advisory Board (2001). *Dioxin Reassessment: An SAB Review of the Office of Research and Development's Reassessment of Dioxin*, USEPA-SAB-EC-01-006, Washington, DC, at [http://yosemite.epa.gov/sab/SABPRODUCT.NSF/C3B2E34A9CD7E9388525718D005FD3D2/\\$File/ec01006.pdf](http://yosemite.epa.gov/sab/SABPRODUCT.NSF/C3B2E34A9CD7E9388525718D005FD3D2/$File/ec01006.pdf).
- [103] Kerger, B.D., Suder, D.R., Schmidt, C.E. & Paustenbach, D.J. (2005). Airborne exposure to trihalomethanes from tap water in homes with refrigeration-type and evaporative cooling systems, *Journal of Toxicology and Environmental Health A* **68**(6), 401–429.
- [104] Paustenbach, D.J., Leung, H.W., Scott, P.K. & Kerger, B.D. (2004). An approach to calculating childhood body burdens of dibenzodioxins and dibenzofurans which accounts for age-dependent biological half-lives, *Organohalogen Compounds* **66**, 2714–2721.
- [105] Environmental Protection Agency (1989). *Risk Assessment Guidance for Superfund, Volume 1: Human Health Evaluation Manual Part A*, USEPA/540/1-89/002, Washington, DC.

Further Reading

Aylward, L.L. & Hays, S.M. (2002). Temporal trends in human TCDD body burden: decreases over three decades and implications for exposure levels, *Journal of Exposure Analysis and Environmental Epidemiology* **12**(5), 319–328.

Related Articles

Cancer Risk Evaluation from Animal Studies
Environmental Hazard
What are Hazardous Materials?

DENNIS J. PAUSTENBACH

Hexavalent Chromium

Chromium may exist in nine valence states; however, only two are generally relevant for environmental and occupational risk assessment: trivalent chromium [Cr(III)] and hexavalent chromium [Cr(VI)]. It is important to speciate the valences of chromium for health risk assessment as the valences have distinctively different properties that are important for characterizing toxicity and exposure [1]. While Cr(III) is an essential micronutrient with very low potential to cause toxicity, Cr(VI) is better absorbed into cells with far greater toxic potential, causing damage at the site of exposure, particularly the skin, gastrointestinal tract (GI), and respiratory system. Upon high level exposures, Cr(VI) may cause systemic effects in the kidney, liver, and anemia [2]. Cr(VI) is also recognized as a human carcinogen, causing lung cancer in workers exposed to high concentrations in certain industries. Reduction of Cr(VI) to Cr(III) is a detoxification mechanism when occurring extracellularly such as in fluids and tissues of the body, including saliva, gastric acid, the liver, and blood [3]. Cr(VI) has not been shown to cause cancer outside the respiratory tract in humans; however, a recent study has shown that very high exposures to Cr(VI) in drinking water can cause cancer in the oral cavity and small intestine of rats and mice, respectively [2].

Cr(III) is the thermodynamically stable form of chromium in most environmental media and the form that occurs naturally in soil. Cr(III) primarily occurs as relatively insoluble oxides or hydroxides at concentrations ranging from 1 to 2000 mg kg⁻¹ [4]. Cr(VI) is produced from chromite ore, which is rich in naturally occurring Cr(III), in the chromate chemical production industry. While very low concentrations of Cr(VI), typically less than 10 ppb, may occur naturally in soil and ground water [5], exposure to Cr(VI) primarily occurs in association with anthropogenic activity, such as in certain industries or from environmental pollution. Because Cr(III) is relatively insoluble at neutral pH, chromium in solution in filtered ground water and drinking water is frequently primarily in the hexavalent state [5, 6].

In addition to chromate production, workers in many industries, including wood treatment, leather tanning, refractory production, ferrochromium and stainless steel production, metal plating and anodizing, pigment production, and those that use chromium

in pigments, paints, and primers and those involving contact with wet cement, have exposure to chromium [7, 8]. The health effects of Cr(VI) are well recognized as a result of high concentration airborne and dermal exposures to Cr(VI) in certain industries [9]. Historical occupational exposures occurring in chromate production, pigment production, and chrome plating have been associated with an increased risk of lung cancer, and several quantitative risk assessments have been derived for lung cancer risk associated with inhalation exposure in the chromate production industry [8]. Other health effects that are recognized to be associated with occupational exposure to Cr(VI) include allergic contact dermatitis (ACD) from contact with wet cement, which contains Cr(VI); it is also highly alkaline and irritating to the skin. Also, historical exposures in some industries have been associated with mild to severe irritation of the upper respiratory tract resulting in tissue damage, which has at extreme exposures, caused nasal septum perforation [10].

Studies of populations with environmental exposures to chromium have mostly been negative [11–13]. While one study of Chinese villagers reported an increased risk of stomach cancer in association with drinking Cr(VI)-contaminated well water [14], there are significant limitations with the exposure and mortality data for this cohort rendering the conclusions questionable [15]. Although it is generally more difficult to study disease outcome in environmentally exposed populations, health risk assessment methods have been developed over the past 20 years to predict the potential hazards associated with Cr(VI) from low level environmental exposures and set cleanup goals and exposure limits for environmental media.

This article describes important health risk assessment principles for evaluation of Cr(VI), based on the current state of the science, and includes discussions of hazard identification, exposure assessment, and toxicity assessment wherein the quantitative dose–response relationships (*see Dose–Response Analysis*) are described. Approaches used to characterize the risk associated with environmental and occupational exposure and set exposure limits are quantitatively described.

Hazard Identification

The health effects associated with oral, dermal, and inhalation exposures that are used as the basis for risk

assessment are discussed here. Extensive discussions of the potential hazards associated with chromium have been provided in agency reviews [4, 8, 9, 16].

Oral

Oral exposure to Cr(III) has generally demonstrated very low bioavailability and toxicity [16]. Because low exposures (e.g., 1 mg l^{-1}) to Cr(VI) are thought to be reduced to Cr(III) in the stomach [17], Cr(VI) toxicity from ingestion exposure has not been reported, except at very high doses. However, at exposures that overwhelm these defenses, tissue damage in the GI, liver, and kidney occur. The most definitive studies of the potential for Cr(VI) to cause reproductive and developmental toxicity in rodents were negative [18, 19].

A recently completed study by the National Toxicology Program (NTP) found that exposures to Cr(VI) at high concentrations can result in an increase in oral cavity cancer in rats and small intestinal cancer in mice [2]. Interestingly, the effects were species specific, e.g., small intestinal tumors did not occur in rats and oral cavity tumors did not occur in mice. Further, increases in tumor occurrence were exclusive to the oral cavity and GI, even though very high levels of chromium circulated systemically in this study. The oral cavity tumors occurred at concentrations of Cr(VI) exceeding 60 mg l^{-1} , and the small intestinal tumors occurred at exposures that exceeded 20 mg l^{-1} . These concentrations of Cr(VI) are highly discolored, and it is not likely that humans would chronically consume Cr(VI) at these levels. Decreased water consumption and body weight were reported in the NTP study for the highest dose groups. These levels of Cr(VI) are also hundreds to thousands of times higher than current human to Cr(VI) in drinking water (Figure 1).

Dermal

Cr(VI) is also a dermal sensitizing agent and can cause ACD among individuals who are allergic to it. Cr(VI)-induced ACD is a Type IV cell-mediated allergic reaction [20]. It is manifested as eczematous erythema (redness), pruritus (itching), and the formation of vesicles and papules on the skin, accompanied by scaling. The reaction is of delayed onset and, depending on the severity, can continue for several weeks if left untreated. Cr(VI)-induced ACD

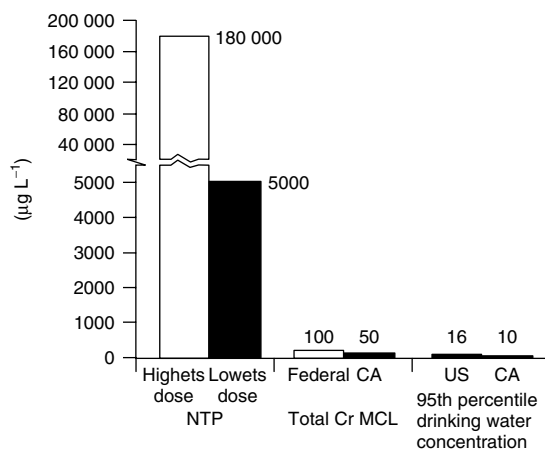


Figure 1 Comparison of Cr(VI) drinking water exposure concentrations in National Toxicology Program rat and mouse cancer bioassay with US total chromium maximum contaminant level (MCL) and 95th percentile drinking water concentrations in United States and California

has been observed in occupational settings, including chrome plating, chromite-ore processing, lithography, leather tanning, and work involving repeated dermal contact with wet cement [21–23]. Because leather is tanned with chromium chemicals, the primary source of nonoccupational dermatitis due to chromium is contact with leather products, such as in footwear and watchbands [24].

ACD is typically described in two phases: induction (or sensitization) and elicitation [25]. In the induction phase, the chemical penetrates into the skin and is taken up and processed by antigen-presenting cells, to activate an allergic response. In the second phase, called elicitation, reexposure to the same allergen activates the memory T cells to trigger an inflammatory response, which results in the clinical symptoms of ACD [25]. Elicitation of ACD is a threshold response in which single or repeated exposures to low doses of Cr(VI) may not produce an allergic response if the threshold dose is not reached [23]. It is generally believed that it takes a higher dose to induce sensitization than to elicit symptoms of ACD (i.e., the elicitation threshold is lower than the induction threshold) [26]. Therefore, dermal standards protective of elicitation of ACD are also protective of the induction of sensitization to Cr(VI).

Cr(VI)-induced ACD is not life threatening, and the effect is generally limited to the skin. The other

type of skin reaction that can occur in response to chemical exposure is irritant contact dermatitis. A cutaneous irritant causes direct damage to the skin without prior sensitization [20, 27]. By comparison, irritant dermatitis is nonimmunological, has a more rapid onset than ACD, and exhibits a more rapid recovery on discontinuation of exposure [27].

Although no survey of the US general population has been conducted to assess the prevalence of allergy to Cr(VI), the prevalence of Cr(VI) allergy among clinical populations – individuals visiting the dermatologist for symptoms of contact allergy – suggests that only a very small fraction of the general population is likely to be allergic to Cr(VI) [23, 28, 29]. Of the more than 17 000 individuals patch tested by physician members of the North American Allergic Contact Dermatitis Group (NACDG), the percent with positive reactions to Cr(VI) has been low: 2.0% in 1994–1996, 2.8% in 1996–1998, 5.8% in 1998–2000, and 4.3% 2000–2002, and of these only approximately 50–70% of the positive responses were considered relevant by the treating physicians [30]. The prevalence rate in the general population, as compared to these data for clinical population patch tested for Cr(VI) allergy as reported in Pratt *et al.* [30], is obviously expected to be much lower [29].

Dermal contact with Cr(VI) has not been shown to cause cancer in animals or humans.

Inhalation

Respiratory tract effects are the most sensitive endpoint for inhalation exposure to Cr(VI), and most importantly, workers exposed to high concentrations of Cr(VI) have experienced an increased risk of respiratory tract cancer, primarily lung cancer [8, 31]. Cr(VI) has also been shown to cause lung cancer in animals upon inhalation and intrabronchial instillation [32, 33]. Cancer of the respiratory system is typically associated with significant tissue irritation and inflammation. Cr(III) is generally not regarded as an inhalation carcinogen and studies of workers exposed to chromium primarily in the trivalent state (e.g., ferrochrome production, tanning) have not demonstrated an increased risk of cancer [34, 35]. Inhalation of Cr(III) at extremely high concentrations ($>4 \text{ mg m}^{-3}$) has caused inflammation effects in the respiratory tract in rats [32, 36, 37].

While inhalation exposure to Cr(VI) has been associated with lung cancer among workers of certain industries, cancer outside the respiratory tract

has typically not been reported [15]. In a recent meta-analysis of chromium exposure and risk of cancer of kidney, stomach, prostate, central nervous system, leukemia, Hodgkin's disease, and lymphohematopoietic cancers, the only significant positive association was found for lung cancer [38]. This is thought to be due to the reduction of Cr(VI) to Cr(III) in the blood and liver following inhalation exposure [3]. Cr(VI) is a respiratory carcinogen that causes cancer at the site of exposure, but is detoxified through reduction in the red blood cells (RBCs) and liver, so that tumors are not observed at locations distant from the site of exposure [3, 15]. Physical and chemical properties of Cr(VI) are important in understanding carcinogenicity. Forms of Cr(VI) that are sparingly soluble – such as calcium and zinc chromate, have a longer biological half-life in the lung and greater carcinogenic potency than forms that are freely soluble – such as chromic acid and sodium dichromate [33, 39]. Also, particle size is important. For example, Cr(VI) in the chromate production industry involves exposure to ultrafine particles (particulate matter (PM) of $<2.5 \mu\text{m}$), and many studies have found an increased risk of lung cancer in this industry [31]. However, among aerospace workers exposed to particles of Cr(VI) in painting and priming, exposures are to much larger particles ($10\text{--}20 \mu\text{m}$), and no study of aerospace workers has reported an increased lung cancer risk [40, 41].

Exposure Assessment

Important considerations in the exposure assessment for chromium are (a) speciating valence state in environmental media, (b) environmental conditions such as pH and oxidation–reduction potential (ORP) dictated valence and bioavailability, (c) Cr(VI) is unstable in the presence of reducing agents, such as organic matter, reduced iron and sulfides, and is converted to Cr(III) – this reaction occurs more rapidly in acidic conditions, and (d) natural sources of both Cr(III) and Cr(VI) should be differentiated from anthropogenic sources. Exposure assessment in air, soil and sediment, and water are discussed in greater detail below.

Air

Because Cr(VI) is recognized as an inhalation carcinogen and Cr(III) is not, it is very important

to speciate Cr(VI) from total chromium in airborne samples for risk assessment. It is also important to characterize the particle size to understand the respirable fraction that can be inhaled and distributed to the lung.

Cr(VI), at ambient conditions, exists as a particulate, not as a vapor. Air sampling and analysis methods for total and hexavalent chromium in occupational and environmental settings have been developed by National Institute for Occupational Safety and Health (e.g., NIOSH Method 7604), and US Environmental Protection Agency (EPA) (e.g., Method 68-D-00-246). As airborne concentrations of Cr(VI) in environmental settings are generally less than 1 ng m^{-3} , it is important to use a method with a sufficiently low limit of detection, generally achieved through use of an ion exchange separation column. For example, concentrations of Cr(VI) in southern California range from 0.03 to 23 ng m^{-3} , with annual mean values by monitoring station that range from 0.04 to 1.24 ng m^{-3} [42].

Soil and Sediment

Because there are important differences in the toxicity of Cr(VI) and Cr(III), it is important to speciate Cr(VI) from total chromium in soils and other solid media, including sediment. Further, because Cr(III) is naturally occurring in soil and sediments at concentrations that typically range between 20 and 200 mg kg^{-1} , it is important to be able to differentiate naturally occurring Cr(III) from Cr(VI), which is not expected to be naturally occurring in soil or sediment, for risk assessment. In some cases the level of concern for Cr(VI) may be below naturally occurring levels of Cr(III). Further, it is important to consider that Cr(VI) may pose an inhalation cancer risk in soil because inhalation exposure might occur by suspension of Cr(VI)-bound to soil through wind erosion of vehicle traffic on unpaved soils. Methods for risk assessment of Cr(VI) in suspended soil have been described in Scott and Proctor [43] and EPA [44].

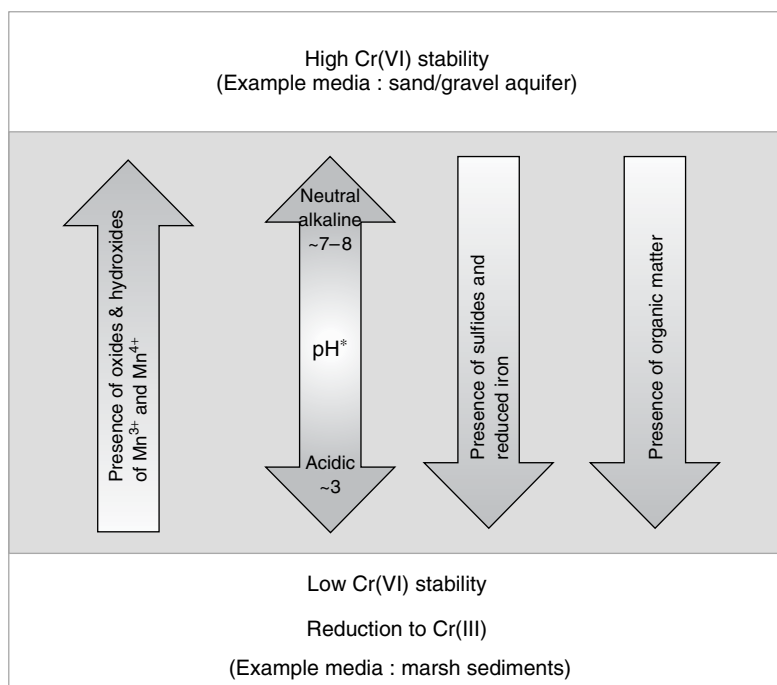
Cr(VI) is typically not present in sediments owing to the abundance of reducing agents; thus total chromium in sediments is generally not a health or environmental concern. Berry *et al.* [45] recently developed the "Cr hypothesis" for evaluating chromium in sediments. The hypothesis states that chromium is of low toxicity in sediments with measurable acid volatile sulfides (AVS) because the conditions are reducing and chromium exists as Cr(III)

which has very low bioavailability and toxicity. This hypothesis has been demonstrated to be correct in several environmental settings [46, 47].

Similar to sediments, the bioavailability or bioaccessibility of chromium in soils is also an important consideration in risk assessment [1]. Chromium that cannot be removed from the soil matrix by biological media is not available for absorption into cells and does not pose a health hazard. Although Cr(VI) is generally soluble in soil, there is also likely to be a fraction that is insoluble and not bioavailable. By comparison, Cr(III) in the oxide or hydroxide form is relatively insoluble, and this characteristic renders most Cr(III) in the environment of negligible bioavailability and toxicity [48]. Two exceptions to this rule involve Cr(III) in sewage sludge and tannery waste [48]. Forms of Cr(III) in these waste streams may be bound to organic molecules or in solution in acidic environments such that their potential bioavailability is much greater. While soluble forms of Cr(III) in soil can be oxidized to Cr(VI) in the presence of manganese oxide, at the same time (a) the reaction is cyclical (i.e., Cr(VI) oxidized from Cr(III) is reduced by reducing agents present in the environment); (b) it is facilitated at lower pH; (c) only soluble forms of Cr(III) may be oxidized to Cr(VI); and (d) only a low percentage, less than 5% of soluble Cr(III), is oxidized [48]. The conditions dictating speciation in soil and sediment are presented in Figure 2. Bioavailability of Cr(VI) in soils and sediments is most appropriately evaluated on a site-specific basis.

Water

In neutral conditions, Cr(III) is typically insoluble and falls out of water as a particulate. Cr(VI) is soluble in neutral and alkaline conditions, and predominates in filtered water. California Department of Health Services (CDHS) [6] has evaluated the levels of Cr(VI) in drinking water from more than 7000 drinking water sources in California and found low levels of Cr(VI), typically ranging from 1 to 10 ppb , in approximately one-third. Most Cr(VI) in CA drinking water is believed to be from natural sources [6]. Similarly, other researchers have found naturally occurring Cr(VI) in groundwater [5, 49]. Levels of naturally occurring Cr(VI) are typically very low ($<50 \text{ } \mu\text{g l}^{-1}$). An extensive data set of potential drinking water data developed by the United



* Rate and occurrence of oxidation/reduction reactions are pH dependent

Figure 2 Factors affecting Cr(VI) stability in environmental media

States Geological Survey (USGS), EPA, and the State of Texas of Cr(VI) and total chromium in ambient groundwater wells includes approximately 25 600 samples from locations across the United States. Cr(VI) was not detected in 48% of the samples in which it was analyzed. The geometric mean and 95th percentile Cr(VI) concentrations were 1.5 and 16 $\mu g\ l^{-1}$, respectively, and those for total chromium were 5.8 and 50 $\mu g\ l^{-1}$, respectively [50].

The current drinking water standard, maximum contaminant level (MCL), in the United States is for total chromium, but is designed to be protective of total chromium as Cr(VI). The MCL is protective of noncancer health effects as quantified using the Cr(VI) reference dose (RfD), which is discussed in the Toxicity Assessment section of this chapter. The MCL was originally set at 50 $\mu g\ l^{-1}$ in the 1970s, then revised to 100 $\mu g\ l^{-1}$ in 1991 [51]. The standard is for total chromium rather than for Cr(VI) because the laboratory analysis for total chromium is simpler, less expensive to conduct and, until improvements take place in the analytical chemistry, achieves a lower limit of detection.

Toxicity Assessment

The methods used to quantify the dose–response relationship for chromium are valence specific and different for cancer and noncancer health effects. Noncancer health effects are quantified using a RfD or RfC (reference concentration), which are typically calculated by dividing a no-observed-adverse-effect level (NOAEL) by uncertainty factors to account for limitations in the NOAEL for describing a threshold dose for sensitive humans. For carcinogenic effects, the relationship between dose and response is typically described using a unit risk factor or cancer slope factor, which describes the relationship between the increased risk or rate of cancer with increasing dose. For inhalation cancer risk assessment of Cr(VI), it is typically assumed that the dose–response observed at the high dose range can be extrapolated to the low dose range with the assumption of linearity and no threshold [16, 52, 53]. No biologically based dose–response model for Cr(VI) currently exists.

Noncancer Toxicity Assessment

Oral. The EPA RfDs for both Cr(III) and Cr(VI) are based on NOAELs observed in rats. The Cr(VI) RfD is based on a drinking water study in which no effects were observed at doses up to $2.5 \text{ mg kg}^{-1} \text{ day}^{-1}$ for 1 year, the highest dose administered [54]. To the NOAEL, a 300-fold uncertainty factor was applied to account for variability between species, sensitive subpopulations, the less than lifetime exposures in the principle study, and uncertainties regarding the reproductive toxicity of Cr(VI) to result in an RfD of $0.003 \text{ mg kg}^{-1} \text{ day}^{-1}$. The EPA RfD for Cr(VI) may be revised in the future based on the results of the 1996, 1997, and 2007 NTP studies [2, 18, 19]. While the NTP studies demonstrated that the uncertainty factor used to account for inadequate information regarding reproductive toxicity is not necessary, the results suggest a much lower effect level in mice as compared to rats (lowest-observed-adverse-effect level (LOAEL) of $0.4 \text{ mg kg}^{-1} \text{ day}^{-1}$ in mice). The NTP study results for rats are consistent with the current RfD, but if the findings in the mouse are deemed of greater relevance for human exposures, the RfD could be decreased accordingly.

Inhalation. Inhalation RfCs are used for noncancer toxicity assessment in a manner similar to RfDs. RfCs are the concentration of a chemical in air, which can be inhaled continuously over the course of a lifetime, and not pose deleterious effects, even among sensitive human subpopulations. EPA has not set an RfC for Cr(III) because there is inadequate data to do so, and the information that does exist suggests that Cr(III) is of limited bioavailability and toxicity [16]. However, for Cr(VI), RfCs have been developed for different forms of Cr(VI), chromic acid mists and Cr(VI) soluble salts/particulates. Separate values have developed because of the recognized differences in toxicity of certain Cr(VI) chemicals. Because of its highly irritating properties, an RfC of 8 ng m^{-3} has been set for chromic acid mists. Chromic acid mists might occur in proximity to a chrome-plating facility, but the environmental half-life is relatively short in this chemical form. For most environmental exposures, Cr(VI) is expected to be in the form of a chromate or dichromate salt [55]. The Cr(VI) RfC for soluble salts and particulates is 100 ng m^{-3} [16]. This RfC is based on a benchmark dose for pulmonary

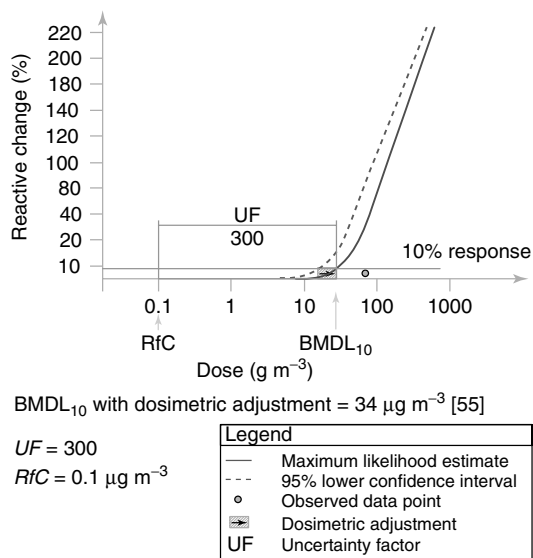


Figure 3 Benchmark dose analysis used to develop the Cr(VI) reference concentration for particulates

inflammation in rats and uses a 300-fold uncertainty factor (Figure 3).

Cancer Risk Assessment

The cancer risk assessment for chromium focuses on the potential cancer risks associated with inhalation exposure to Cr(VI). Until the 2007 NTP rodent study, the weight of evidence indicated that Cr(VI) did not poses a cancer risk from ingestion exposure, and Cr(III) has not been considered carcinogenic *via* any route of exposure [15].

Oral Cancer Risk Assessment. While no quantitative cancer risk assessment for oral exposures to Cr(VI) have been developed to date, there are several important considerations when using the NTP rat oral cavity cancer and mouse small intestinal cancer data for human risk assessment.

First, ingested Cr(VI) is reduced to the trivalent state in the stomach [3, 17, 56]. This is a detoxification process, and it is reasonable to expect that it may be overwhelmed at high doses. The reduction capacity of the mouse stomach may be considerably less than that of the rat and that of a human, potentially explaining the occurrence of small intestine tumors in mice, but not in rats. The tissue data collected from the NTP study suggests that much higher

levels of Cr(VI) were passed from the stomach of the mouse to the small intestine, as compared to that of the rat, because much higher levels of Cr were observed in the mouse liver (Figure 4). Humans are expected to have even greater capacity to reduce Cr(VI) to Cr(III) in the stomach because of the much lower pH of the human stomach (fasting pH of ~ 1.5 – 2) as compared to a mouse or rat (glandular stomach pH of ~ 3) [57]. As a result, the dose necessary to induce tumors of the small intestine of a mouse may be far less than that which could induce tumors in humans and a physiologically based pharmacokinetic (PBPK) model should be used for interspecies extrapolations.

Second, many NTP studies for other chemicals have found that rats are more sensitive to oral cavity tumors than mice, and the rat oral cavity tumors were found in absence of non-neoplastic lesions that are suggestive of chronic tissue damage [2]. For this reason, it may be presumed that the mode of action (MOA) for oral cavity cancers is more complicated than Cr(VI) causing tumors from site-of-contact mutagenic activity or because of chronic hyperplasia. More research is needed to better characterize the MOA in rats, and it is possible that the observations in rats are species specific or only occur at extraordinarily high exposures.

Finally, it is important to consider that the doses to which the NTP rodents were chronically exposed far exceeds the levels of Cr(VI) to which human could be exposed to in drinking water (Figure 1). It is probable that drinking water exposures at the current exposure limit ($100 \mu\text{g l}^{-1}$ of total chromium) can be detoxified through reduction to Cr(III) before posing an oral cancer hazard.

Inhalation Cancer Risk Assessment. The dose–response for increased cancer risk associated with inhalation exposure to Cr(VI) has been quantitatively evaluated in several studies [8, 52, 58]. Because several epidemiological studies present increased lung cancer risk with increasing exposure, animal studies have not typically been used to quantify dose–response. Dose–response is quantified using a unit risk factor that is the increased risk associated with continuous exposure to $1 \mu\text{g m}^{-3}$ of Cr(VI) over an exposure period.

The unit risk factors calculated for occupational and environmental populations demonstrate general consistency between studies. For continuous environmental exposures (24 h day^{-1} for a lifetime) EPA calculated a unit risk factor of $0.0117 \mu\text{g}^{-1} \text{ m}^{-3}$ using the Mancuso [59] study, and using updated

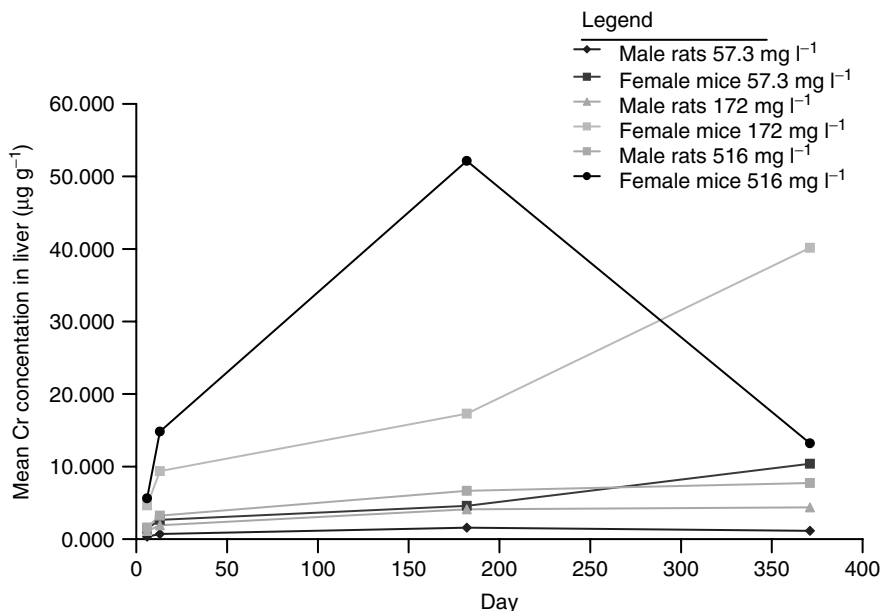


Figure 4 Mean chromium concentrations in liver of NTP female mice and male rats by dose group

approaches, Crump *et al.* [52] calculated an environmental unit risk value of $0.00978 \mu\text{g}^{-1} \text{m}^{-3}$ using the epidemiology data of Luippold *et al.* [60]. Crump and colleagues also calculated an occupational unit risk value (for exposure 8 h day^{-1} for 45 years) of $0.00205 \mu\text{g}^{-1} \text{m}^{-3}$, and Park *et al.* [58] calculated a corresponding value of $0.0049 \mu\text{g}^{-1} \text{m}^{-3}$, and were able to quantitatively account for the risk associated with smoking in the cohort. While it is a distinct advantage to be able to quantify risk using human data and to have multiple data sets from which to conduct the quantitative evaluation, the limitations of these estimates are that they are all based on workers exposed in a single industry, chromate production, and are all based on high concentration exposures that occurred over a relatively short duration (e.g., few months to few years) extrapolated to low level exposures on the basis of cumulative dose. Finally, it is important to consider that these cohorts all had a very high prevalence of smoking.

Allergic Contact Dermatitis. The dose–response relationship for ACD is characterized using the 10% minimum elicitation threshold ($\text{MET}_{10\%}$) for Cr(VI). A recent study by Proctor *et al.* [61] developed a $\text{MET}_{10\%}$ based on a repeated open application test (ROAT) for a population of individuals who were previously known to be allergic to Cr(VI). The study was performed to provide improved data for assessing the potential risk posed by Cr(VI) in environmental exposure situations (e.g., dermal contact with Cr(VI)-contaminated soil). The $\text{MET}_{10\%}$ from this study can be used to develop an RfD for dermal contact (skin-RfD, or s-RfD) for risk assessment of Cr(VI) in treated wood or in environmental media. Because the $\text{MET}_{10\%}$ is dependent on the sensitivity of the study population, Proctor *et al.* weighted $\text{MET}_{10\%}$ value to the patch-test reaction grades of the North American Allergic Contact Dermatitis Group. The resulting $\text{MET}_{10\%}$ values are provided in Table 1. Proctor *et al.* [61] also provided $\text{MET}_{10\%}$ values for two scenarios: (a) scenario 1 included only responses deemed to be allergic in nature by the principal investigator and (b) scenario 2 is based on all responses including allergic and those deemed to be irritant in nature.

No Uncertainty Factor (UF) factors are needed to convert the $\text{MET}_{10\%}$ values to s-RfDs because the tests were performed on sensitive humans, and ACD is an acute affect. For application of the s-RfD to the

Table 1 $\text{MET}_{10\%}$ calculated for scenarios 1 and 2 from the Proctor *et al.* [61] study and the patch-test-normalized populations for repeated open application test challenges to potassium dichromate^(a)

Population	Scenario ^(b)	Modeled $\text{MET}_{10\%}$ (ng of Cr(VI)/cm ²)
Proctor <i>et al.</i> study population	1	118
Patch-test-normalized population	1	364
Proctor <i>et al.</i> study population	2	86.6
Patch-test-normalized population	2	269

^(a) Reproduced from [61]. US Environmental Protection Agency, 2006

^(b) Scenario 1 includes allergic responses only; Scenario 2 includes all allergic and irritant responses

general population, the $\text{MET}_{10\%}$ value for the patch-test-normalized value is more appropriate. Thus, the s-RfD for the general population of Cr(VI) sensitized individuals is $269 \text{ ng of Cr(VI)/cm}^2$ for all responses (irritant and allergic) and $364 \text{ ng of Cr(VI)/cm}^2$ for allergy. Interestingly, the study population $\text{MET}_{10\%}$ of $118 \text{ ng of Cr(VI)/cm}^2$ from the Proctor *et al.* ROAT study is $\sim 30\%$ higher than that developed by Nethercott *et al.* of $89 \text{ ng of Cr(VI)/cm}^2$ using occlusive patch testing [28, 61]. The data necessary to weight the results by patch-test grade was not provided in the Nethercott *et al.* study; thus these data could not be normalized by patch-test grade.

Risk Characterization

Risk characterization is performed by combining the hazard identification, toxicity assessment, and exposure assessment to yield quantitative risk estimates and health-based guidelines for environmental media (*see Environmental Hazard; Environmental Health Risk*).

Cr(VI) Guidelines for Soil, Treated Wood, and other Surfaces Based on ACD

On the basis of the approach described in Nethercott *et al.* [28] and the patch-test-normalized $\text{MET}_{10\%}$ developed in Proctor *et al.* [61] ROAT study (364 ng cm^{-2}), the calculated soil standard that is

protective of ACD from contact with Cr(VI) in soil is 1820 mg kg^{-1} . While this value exceeds EPA's soil screening levels (SSLs) that are protective of Cr(VI) for soil ingestion and inhalation at residential sites [44], it is lower than that for soil ingestion for commercial and industrial sites and provides useful information for the risk assessment of Cr(VI) from short-term (acute) exposures to soil for other scenarios, such as soil contact for a construction worker. These data may also be used to assess the potential dermal exposures to Cr(VI) on surfaces such as building walls or equipment.

Cancer Risks Associated with Cr(VI) in Air

The theoretical excess cancer risk from inhalation exposure is quantified using measured or model predicted levels of Cr(VI) in air and cancer slope or unit risk factors for occupational and continuous lifetime exposures [52]. Although derived from human data, these factors are recognized to have several uncertainties and generally thought to overestimate risk at low levels of exposure. This occurs because the slope factors are based on high concentration exposures that occur over a relatively short duration which are used as a cumulative dose (e.g., 1 year of exposure to $100 \mu\text{g m}^{-3}$ is assumed to be equal to 40 years of exposure at $2.5 \mu\text{g m}^{-3}$). Using Occupational Safety and Health Administration (OSHA's) risk assessment, the theoretical excess lung cancer risk at the new limit of $5 \mu\text{g m}^{-3}$ for a 45-year occupational lifetime is 10–45 per 1000 [8]. Using the EPA [16] and Crump *et al.* [52] environmental risk assessments, the theoretical excess cancer risk associated with lifetime exposure to Cr(VI) at 1 ng m^{-3} – the approximate upper bound of background concentration of Cr(VI) in ambient air in urban areas, is 1 in 100 000.

Health-Based Soil Guidelines for Protection of Human Health and Wildlife

The environmental and human health risks posed by Cr(VI) in soil are characterized from many potential exposures including leaching to groundwater, non-cancer hazards due to dermal contact and incidental ingestion, and the potential cancer risk associated with inhalation of particulates suspended from vehicle traffic and wind erosion. Health-protective guidelines for Cr(VI) for industrial/commercial properties,

which support significant vehicle traffic, and construction scenarios involving significant dirt moving activities, can be more conservative (e.g., lower) than those protective of residential land use owing to the potential cancer risk associated with inhalation of suspended particulates [62]. Soil standards for residential land use protective of ingestion exposure among children are typically around 230 mg kg^{-1} [44], whereas the level protective of workers at sites with vehicle traffic can be in the range of $20\text{--}30 \text{ mg kg}^{-1}$ [44, 63]. Levels of Cr(VI) protective of drinking water sources are dependent on site-specific conditions, with default values that range from 2 to 20 mg kg^{-1} [44]. For all scenarios, particularly for leaching to groundwater and soil suspension owing to vehicle traffic, risk assessment and determination of cleanup levels should be performed on a site-specific basis, considering actual site conditions.

The US EPA has developed risk-based ecological soil screening levels (Eco-SSLs) for Cr(VI) and Cr(III). The Eco-SSLs are defined as chemical concentrations that are protective of ecological receptors that commonly come into contact with soil, or ingest biota that live in, or on, soil. If sufficient toxicological data are available for a chemical, Eco-SSLs are derived separately for four diverse groups of ecological receptors: plants, soil, invertebrates, and wildlife (i.e., birds and mammals). Sufficient toxicological information on chromium was available to derive Eco-SSLs only for wildlife receptors. For avian receptors, sufficient information was available to develop an Eco-SSL only for Cr(III), whereas for mammalian receptors sufficient information was available to develop Eco-SSLs for both Cr(III) and Cr(VI). The Eco-SSLs for Cr(VI) for mammalian receptors is 81 mg kg^{-1} [64].

Drinking Water Standard

The US federal drinking water standard of $100 \mu\text{g l}^{-1}$ (ppb) is for total chromium but designed to be protective of Cr(VI) for noncancer effects [51]. As the MOA and cancer risk assessment of Cr(VI) is developed from the recent NTP rodent bioassay, it is expected that these standards may be revised in the future and that separate standard for Cr(VI) and Cr(III) will be developed. Research to evaluate the MOA for these tumors, interspecies variability and extrapolation from rodents to humans, and low dose extrapolation of the dose–response relationship will

be necessary for use of the NTP data for human risk assessment of ingested Cr(VI).

References

- [1] USEPA (U.S. Environmental Protection Agency) (2007). *Framework for Metals Risk Assessment*, EPA 120/R-07/001, <http://www.epa.gov/osa/metalsframework/>.
- [2] NTP (National Toxicology Program) (2007). *NTP Technical Report on the Toxicological and Carcinogenesis Studies of Sodium Dichromate Dihydrate*, (CAS No. 7789-12-0) in F344/N Rats and B6C3F1 Mice (Drinking Water Studies) NTP TR 546, National Institute of Health, US Department of Health and Human Services, Peer Review draft, Research Triangle Park.
- [3] De Flora, S. (2000). Threshold mechanisms and site specificity in chromium(VI) carcinogenesis, *Carcinogenesis* **21**, 533–541.
- [4] ATSDR (Agency for Toxic Substances and Disease Control) (2000). *Toxicological Profile for Chromium (update)*, U.S. Public Health Service, Atlanta.
- [5] Oze, C., Bird, D.K. & Fendorf, S. (2007). Genesis of hexavalent chromium from natural sources in soil and groundwater, *Proceedings of the National Academy of Sciences* **104**(16), 6544–6549.
- [6] CDHS (California Department of Health Services) (2007). *Chromium-6 in Drinking Water: Regulatory Update*, <http://www.cdph.ca.gov/certlic/drinkingwater/Pages/Chromium6sampling.aspx>.
- [7] Barnhart, J. (1997). Occurrences, uses, and properties of chromium, *Regulatory Toxicology and Pharmacology* **26**, S3–S7.
- [8] OSHA (Occupational Safety and Health Administration) (2006). Occupational exposure to hexavalent chromium; final rule, 29 CFR parts 119, etc, *Federal Register* **71**(39), 10100–10385.
- [9] IARC (International Agency for Research on Cancer) (1990). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans Chromium, Nickel, and Welding*, World Health Organization, Lyon, Vol. 49.
- [10] Gibb, H.J., Lees, P.S., Pinsky, P.F. & Rooney, B.C. (2000). Clinical findings of irritation among chromium chemical production workers, *American Journal of Industrial Medicine* **38**, 127–131.
- [11] Frzek, J.P., Mumma, M.T., McLaughlin, J.K., Henderson, J.K. & Blot, W.J. (2001). Cancer mortality in relation to environmental chromium exposure, *Journal of Occupational and Environmental Medicine* **43**(7), 635–640.
- [12] Armienta-Hernandez, M. & Rodriguez-Castillo, R. (1995). Environmental exposure to chromium compounds in the Valley of Leon, Mexico, *Environmental Health Perspectives* **103**(Suppl. 1), 47–51.
- [13] GGHG (Greater Glasgow Health Board) (1991). *Assessment of the Risk to Human Health from Land Contaminated by Chromium Waste*, Department of Public Health, Glasgow.
- [14] Zhang, J.D. & Li, X.L. (1987). Chromium pollution of soil and water in Jinzhou, *Chung Hua Yu Fang I Shueh Tsa Chih* **21**, 262–264.
- [15] Proctor, D.M., Otani, J.M., Finley, B.L., Paustenbach, D.J., Bland, J.A., Speizer, N. & Sargent, E.V. (2002). Is hexavalent chromium carcinogenic via ingestion? a weight-of-evidence review, *Journal of Toxicology and Environmental Health, Part A* **65**, 701–746.
- [16] USEPA (U.S. Environmental Protection Agency), Integrated Risk Information System (IRIS) (1998). *Toxicological Profile Update on Hexavalent Chromium*, Washington, DC.
- [17] Kerger, B.D., Finley, B.L., Corbett, G.E., Dodge, D.G. & Paustenbach, D.J. (1997). Ingestion of chromium (VI) in drinking water by human volunteers: absorption, distribution, and excretion of single and repeated doses, *Journal of Toxicology and Environment Health* **50**, 67–95.
- [18] NTP (National Toxicology Program) (1996). *Final Report on the Reproductive Toxicity of Potassium Dichromate (CAS No. 7778-50-9) Administered in Diet to SD Rats*, National Institute of Environmental Health Sciences, Research Triangle Park.
- [19] NTP (National Toxicology Program) (1997). *Final Report on the Reproductive Toxicity of Potassium Dichromate (CAS No. 7778-50-9) Administered in Diet to BALB/c Mice*, National Institute of Environmental Health Sciences, Research Triangle Park.
- [20] Marks, J.G. & DeLeo, V.A. (1992). *Contact and Occupational Dermatology*, Mosbey Year Book, St. Louis.
- [21] Levin, H.N., Brunner, N.J. & Ratner, H. (1959). Lithographer's dermatitis, *Journal of the American Medical Association* **169**, 566–569.
- [22] Morris, G.E. (1958). Chrome dermatitis: a study of the chemistry of shoe leather with particular reference to basic chromic sulfate, *American Medical Association Archives of Dermatology* **78**(5), 612–618.
- [23] Paustenbach, D.J., Sheehan, P.J., Paull, J.M., Wisser, L.M. & Finley, B.L. (1992). Review of the allergic contact dermatitis hazard posed by chromium-contaminated soil: identifying a "safe" concentration, *Journal of Toxicology and Environment Health* **37**(1), 177–207.
- [24] Hansen, M.B., Johansen, J.D. & Menne, T. (2003). Chromium allergy: significance of both Cr(III) and Cr(VI), *Contact Dermatitis* **49**(4), 206–212.
- [25] Robinson, M.K., Gerberick, G.F., Ryan, C.A., McNamee, P., White, I.R. & Basketter, D.A. (2000). The importance of exposure estimation in the assessment of skin sensitization risk, *Contact Dermatitis* **42**(5), 251–259.
- [26] Friedmann, P.S. (2000). The immunology of allergic contact dermatitis: the DNCB story, *Advances in Dermatology* **5**, 175–196.
- [27] Jackson, E. & Goldmen, R. (eds) (2000). *Irritant Contact Dermatitis*, Marcel Dekker, New York.
- [28] Nethercott, J., Paustenbach, D., Adams, R., Fowler, J., Marks, J., Morton, C., Taylor, J., Horowitz, S. & Finley, B. (1994). A study of chromium induced allergic contact dermatitis with 54 volunteers: implications for

- environmental risk assessment, *Occupational and Environmental Medicine* **51**(6), 371–380.
- [29] Proctor, D.M., Fredrick, M.M., Scott, P.K., Paustenbach, D.J. & Finley, B.L. (1998). Prevalence of chromium allergy in the United States and its implications for setting soil cleanup levels: a cost-effectiveness case study, *Regulatory Toxicology and Pharmacology* **28**, 27–37.
- [30] Pratt, M.D., Belsito, D.V., DeLeo, V.A., Fowler, J.F., Jr., Fransway, A.F., Maibach, H.I., Marks, J.G., Mathias, C.G., Rietschel, R.L., Sasseville, D., Sherertz, E.F., Storrs, F.J., Taylor, J.S. & Zug, K. (2004). North American contact dermatitis group patch test results, 2001–2002 study period, *Dermatitis* **15**, 176–183.
- [31] Proctor, D.M., Panko, J.P., Liebig, E.W., Scott, P.K., Mundt, K.A., Buczynski, M.A., Barnhart, R.J., Harris, M.A., Morgan, R.J. & Paustenbach, D.J. (2003). Workplace airborne hexavalent chromium concentrations for the Painesville, Ohio, chromate production plant (1943–1971), *Applied Occupational and Environmental Hygiene* **18**, 430–449.
- [32] Glaser, U., Hochrainer, D., Kloppel, H. & Oldies, H. (1986). Carcinogenicity of sodium dichromate and chromium (VI/III) oxide aerosols, *Toxicology* **42**, 219–232.
- [33] Steinhoff, D., Hatfield, G., Gad, S. & Mohr, U. (1986). Carcinogenicity study with sodium dichromate in rats, *Experimental Pathology* **30**, 129–141.
- [34] Axelsson, G., Rylander, R. & Schmidt, A. (1980). Mortality and incidence of tumours among ferrochromium workers, *British Journal of Industrial Medicine* **37**, 121–127.
- [35] Montanaro, F., Ceppi, M., Demers, P.A., Puntoni, R. & Bonassi, S. (1997). Mortality in a cohort of tannery workers, *Occupational and Environmental Medicine* **54**, 588–591.
- [36] Derelenko, M.J., Rinehart, W.E., Hilaski, R.J., Thompson, R.B. & Loser, E. (1999). Thirteen-week subchronic rat inhalation toxicity study with a recovery phase of trivalent chromium compounds, chromic oxide, and basic chromium sulfate, *Toxicological Sciences* **52**, 278–288.
- [37] Johansson, A., Wiemik, A., Jasstrand, P. & Camner, P. (1986). Rabbit alveolar macrophages after inhalation of hexa- and trivalent chromium, *Environmental Research* **39**, 372–385.
- [38] Cole, P. & Rodu, B. (2005). Epidemiologic studies of chrome and cancer mortality: a series of meta-analyses, *Regulatory Toxicology and Pharmacology* **43**(3), 225–231.
- [39] Levy, L.S., Martin, P.A. & Bidstrup, P.L. (1986). Investigation of the potential carcinogenicity of a range of chromium containing materials on rat lung, *British Journal of Industrial Medicine* **43**(4), 243–256.
- [40] Sabty-Daily, R.A., Harris, P.A., Hinds, W.C. & Froines, J.R. (2005). Size distribution and speciation of chromium in paint spray aerosol at an aerospace facility, *Annals of Occupational Hygiene* **49**(1), 47–59.
- [41] Boice, J.D., Marano, D.E., Fryzek, J.P., Sadler, C.J. & McLaughlin, J.K. (1999). Mortality among aircraft manufacturing workers, *Occupational and Environmental Medicine* **56**, 581–597.
- [42] CARB (California Air Resources Board) (2004). *Monitoring Sites with Ambient Toxics Summaries: Chromium*, <http://www.arb.ca.gov/aqd/toxics/sitelists/crsites.html>.
- [43] Scott, P.K. & Proctor, D.M. (2007). Soil suspension/dispersion modeling methods for estimating health-based soil cleanup levels of hexavalent chromium at chromite ore processing residue sites, *Journal of Air Waste Management Association* (in press).
- [44] USEPA (U.S. Environmental Protection Agency) (2002). *Supplemental Guidance for Developing Soil Screening Levels for Superfund Sites*, Washington, DC.
- [45] Berry, W.J., Boothman, W.S., Serbs, J.R. & Edwards, P.A. (2004). Predicting the toxicity of chromium in sediments, *Environmental Toxicology and Chemistry* **23**, 2981–2992.
- [46] Becker, D.S., Long, E.R., Proctor, D.M. & Ginn, T.C. (2006). Toxicity and bioavailability of chromium in sediments associated with chromite ore processing residue, *Environmental Toxicology and Chemistry* **25**(10), 2576–2583.
- [47] Besser, J.M., Brumbaugh, W.G., Kemble, N.E., May, T.W. & Ingersoll, C.G. (2004). Effects of sediment characteristics on the toxicity of chromium(III) and chromium(VI) to the amphipod, *Hyalella azteca*, *Environmental Science and Technology* **38**, 6210–6216.
- [48] James, B.R., Petura, J.C., Vitale, R.J. & Mussoline, G.R. (1997). Oxidation-reduction chemistry of chromium: relevance to the regulation and remediation of chromate-contaminated soils, *Journal of Soil Contamination* **6**, 569–580.
- [49] Reynolds, J.F., Kemp, P.R., Ogle, K. & Fernandez, R.J. (2004). Modifying the ‘pulse-reserve’ paradigm for deserts of North America: precipitation pulses, soil water, and plant responses, *Oecologia* **141**(2), 194–210.
- [50] ChemRisk (2007). *Database of Total Chromium and Hexavalent Chromium Groundwater Samples*, Reported to or Collected by the United States Geological Survey (USGS) and Environmental Protection Agency has been Compiled by PRISMEDICAL Corporation (PMC) Prepared for Tierra Solutions, Inc, Houston.
- [51] USEPA (U.S. Environmental Protection Agency) (1991). National primary drinking water regulations-synthetic organic chemicals and inorganic chemicals; monitoring for unregulated contaminants; national primary drinking water regulations implementation; national secondary drinking water regulations (final rule), *Federal Register* **56**, 3526–3597.
- [52] Crump, C., Crump, K.S., Hack, E., Luippold, R.S., Mundt, K.A., Panko, J.P., Liebig, E.W., Paustenbach, D.J. & Proctor, D.M. (2003). Dose-response and risk assessment of airborne hexavalent chromium and lung cancer mortality, *Risk Analysis* **23**(6), 1155–1171.
- [53] Park, R.M. & Stayner, L.T. (2006). A search for thresholds and other nonlinearities in the relationship

- between hexavalent chromium and lung cancer, *Risk Analysis* **26**(1), 79–88.
- [54] MacKenzie, R.D., Anwar, R.A., Byerrum, R.U. & Hoppert, C.A. (1958). Chronic toxicity studies: hexavalent chromium and trivalent chromium administered in drinking water, *Archives of Industrial Health* **18**, 232–234.
- [55] Malsch, P.A., Proctor, D.M. & Finley, B.L. (1994). Estimation of a chromium inhalation reference concentration using the benchmark dose method: a case study, *Regulatory Toxicology and Pharmacology* **20**, 58–82.
- [56] Kuykendall, J.R., Kerger, B.D., Jarvi, E.J., Corbett, G.E. & Paustenbach, D.J. (1996). Measurement of DNA-protein cross-links in human leukocytes following acute ingestion of chromium in drinking water, *Carcinogenesis* **17**, 1971–1977.
- [57] de Zwart, L.L., Rempelberg, C.J.M., Sips, A.M., Welink, J. & van Engelen, J.G.M. (1999). Anatomical and physiological differences between various species used in studies on the pharmacokinetics and toxicology of xenobiotics, RIVM Report 623860 010, The National Institute of Public Health and the Environment, Bilthoven.
- [58] Park, R.M., Bena, J.F., Stayner, L.T., Smith, R.J., Gibb, H.J. & Lees, P.S. (2004). Hexavalent chromium and lung cancer in the chromate industry: a quantitative risk assessment, *Risk Analysis* **24**(5), 1099–1108.
- [59] Mancuso, T.F. (1975). Consideration of chromium as an industrial carcinogen, Paper presented at *The International Conference on Heavy Metals in the Environment*, Ontario, October 27–31, 1975.
- [60] Luippold, R.S., Mundt, K.A., Austin, R.P., Liebig, E., Panko, J.P., Crump, C., Crump, K. & Proctor, D.M. (2003). Lung cancer mortality among chromate workers, *Occupational and Environmental Medicine* **60**, 451–457.
- [61] Proctor, D.M., Gujral, J., Su, S. & Fowler, J.F. (2006). *Final Report Repeated Open Application Test for Allergic Contact Dermatitis due to Hexavalent Chromium [Cr(VI)] as Potassium Dichromate: Risk Assessment for Dermal Contact with Cr(VI)*, Prepared for USEPA Office of Pesticide Programs FPRL# 012406, Exponent, Irvine.
- [62] Proctor, D.M., Shay, E.C. & Scott, P.K. (1997). Health-based soil action levels for trivalent and hexavalent chromium: a comparison to state and federal standards. Chromium in soil: perspectives in chemistry, health and environmental regulation, *Special Issue Journal of Soil Contamination* **6**(6), 595–648.
- [63] NJDEP (New Jersey Department of Environmental Protection) (1998). *Soil Cleanup Criteria for Trivalent and Hexavalent Chromium*, Department of Environmental Protection, Trenton, http://64.233.169.104/search?q=cache:iDi8EGG00iAJ:www.state.nj.us/dep/srp/siteinfo/chrome/b_b_sum_3.htm+NJDEP+SCC+chromium&hl=en&ct=clnk&cd=1&gl=us NJ.
- [64] USEPA (U.S. Environmental Protection Agency) (2005). *Ecological Soil Screening Levels for Chromium, Interim Final*, OSWER Directive 9285.7-66, Office of Solid Waste and Emergency Response, Washington, DC.

Related Articles

Air Pollution Risk

Environmental Exposure Monitoring

Hazardous Waste Site(s)

Occupational Cohort Studies

Soil Contamination Risk

Statistics for Environmental Toxicity

Water Pollution Risk

DEBORAH M. PROCTOR

History and Examples of Environmental Justice

Environmental Justice Defined and its Relationship to Risk

Definition of Environmental Justice

Environmental justice refers to the spatial distribution of environmental risks, and costs and benefits across different population groups associated with siting, development, regulatory, and policy decisions that involve both the public and private sectors. Population groups included in environmental justice analyses are typically categorized by factors such as race, ethnicity, income and age. The United States Environmental Protection Agency (USEPA) defines environmental justice as follows [1]:

“Environmental Justice is the fair treatment and meaningful involvement of all people regardless of race, color, national origin, culture, education, or income with respect to the development, implementation, and enforcement of environmental laws, regulations, and policies. Fair Treatment means that no group of people, including racial, ethnic, or socioeconomic groups, should bear a disproportionate share of the negative environmental consequences resulting from industrial, municipal, and commercial operations or the execution of federal, state, local, and tribal environmental programs and policies. Meaningful Involvement means that: (1) potentially affected community residents have an appropriate opportunity to participate in decisions about a proposed activity that will affect their environment and/or health; (2) the public’s contribution can influence the regulatory agency’s decision; (3) the concerns of all participants involved will be considered in the decision-making process; and (4) the decision-makers seek out and facilitate the involvement of those potentially affected.”

Relationship to Risk Assessment and Risk Management: Some Examples

When risk-assessment and risk-management strategies and techniques were first developed, population as the target of potential exposures was viewed in the aggregate; that is, population was usually not

subdivided into population sectors. The individual was usually considered the unit of exposure. As the civil rights and environmental movements emerged, justice issues by population sector emerged. Some examples are noteworthy. First, in the *environmental health risk assessment* (see **Environmental Health Risk**) area, the exposure component of risk assessment methodologies involves estimates of the amount of a contaminant that is absorbed. Fish is consumed in different ways and in different quantities by population sector. Some low-income populations, for example, consume fish from waters that are more contaminated than others. Some populations consume the entire fish rather than just small portions of it. Finally, fish constitutes a greater proportion of the diets of some populations than others. All of these situations imply the need to adapt exposure values by population sector. Secondly, asthma is now considered to be on the rise. Hospitalizations due to asthma tend to be concentrated in low-income and minority areas, for example, in New York City [2]. This suggests a need to divide risk assessments based on air quality and inhalation according to population sector (see **Air Pollution Risk**).

The History and Origin of Environmental Justice

The Evolution of the US Experience

Environmental justice is a lens through which to view risks to human life, health, and well-being. It has evolved in the United States over decades and in many contexts. The common theme, as defined above, is the existence of disparities based on social and economic characteristics of individuals and groups. Social and economic characteristics of communities as a foundation for portraying disparities originally emphasized race and class. These categories have been expanded to include many other often overlapping characteristics considered to potentially reflect human vulnerabilities. For example, susceptibility to added burdens of pollution are potentially associated with ethnicity, since certain lifestyles are associated with exposure to different levels of pollution; certain potentially vulnerable age groups, in particular, the very old and the very young; various life stages, such as pregnancy; and health status – those with disabilities or disease. Disparities pertain broadly to exposure to adverse conditions and

access to resources to reduce these exposures and their adverse effects.

Several concepts have emerged to characterize these concerns, namely: equity, justice, and racism. Environmental equity and environmental justice are distinctly different concepts: equity being a more neutral term referring to distribution of assets or opportunities, whereas justice is normative encompassing the fairness of a distribution or opportunities [3, p. 633]. Racism further encompasses a deliberate attempt at shifting the distribution of burden and discrimination.

The origins of environmental equity, justice, and racism in the United States are generally attributed to the civil rights movement of the mid-twentieth century ([4, p. 650], and codified in Title VI of the Civil Rights Act of 1964 (Public Law 88–352, Title VI, Sec. 601, July 2, 1964, 78 Stat. 252)) that required that “No person in the United States shall, on the ground of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving Federal financial assistance.” During the late 1960s and the decade following, most of the major environmental laws were passed, many involving substantial federal funding, thus potentially invoking the requirements of Title VI. Implementing this order in the context of environmental law occurred initially with respect to abandoned hazardous waste disposal sites. In 1980, the Comprehensive Environmental Response, Compensation and Liability Act (CERCLA, known as the *Superfund Act*) was passed to regulate these sites. The United States General Accounting Office (US GAO, now called the *US Government Accountability Office*) identified one of the early cases of waste siting in a predominantly minority community in North Carolina [3, 5]. Following that, during the 1980s, waste sites were identified in communities of color and low-income communities throughout the country. Case studies moved to the realm of statistical analysis that captured environmental justice as a more widespread phenomenon [4, 6]. In 1987, the United Church of Christ published the report of a study of the racial and socioeconomic characteristics of the communities in which hazardous waste sites were located [7], published in a more extensive form a few years later [8]. In spite of the national attention drawn to this issue and the increasing sophistication of the methods of analysis, early environmental

justice cases involving waste sites were not won by environmental justice advocates. For example, three landfill cases were brought and lost by environmental justice supporters in the late 1980s and early 1990s – *Bean v Southwestern Waste Management Corp.*, *East Bibb Twiggs Neighborhood Association v Macon-Bibb County Planning & Zoning Commission*, and *R.I.S.E., Inc. v Kay*. The cases were brought on the basis of disproportionate shares of minorities in some of the areas surrounding landfills, but the cases were lost due in part to the lack of specific benchmarks to interpret those shares and the focus of the courts on intentionality or discrimination, which they felt were not supported [3, 4].

Attention to environmental justice spread from a focus on waste disposal sites (both active and inactive or abandoned) to other forms of pollution, for example, *air pollution* (see **Air Pollution Risk**) [9]. Two decades after the formal emergence of the environmental movement in the United States, the president signed an executive order, E.O. 12989 in 1994, which underscored procedures to incorporate environmental justice throughout federal actions [10].

Paralleling environmental justice issues in connection with waste and other environmental exposures was the emergence of injustices in the area of infrastructure, many of which were connected with environmental contamination. The adverse environmental aspects of transportation, water, and energy have emerged as key justice issues [11]. In 2001, transportation accounted for 82% of carbon monoxide emissions, 56% of nitrogen oxide emissions, and 42% of volatile organic compounds; energy accounted for 86% of sulfur dioxide emissions and 39% of nitrogen oxides from fuel combustion [12: 580]. These pollutants are the key regulated air pollutants under the Clean Air Act. Transportation poses multifaceted justice issues, for example, in connection with lack of access to public transit forcing reliance of lower income people on older automobiles that expose them to higher emissions. A key justice issue related to transportation pertains to the proximity of lower income groups to roadways and traffic increasing their exposure to transportation-related air pollutants and noise [13]. Risk-related quantitative approaches have been applied to transportation justice issues [14]. In the area of energy, similarly, the poor, racial and ethnic minorities may seek lower cost housing near power plants. The siting of natural gas fired turbines in low-income New York City neighborhoods, in

2001, to alleviate a summer power shortage was very controversial in that the decision bypassed normal environmental reviews [15]. Water rights and water quality issues have been a continual subject of environmental justice both nationally and internationally [16] and the latest USEPA strategic plan also prioritizes this area [17]. Injustice issues have arisen in connection with other areas of infrastructure, not directly related to environmental conditions, such as access to computers and the internet.

Linkages between environmental conditions and injustices were largely drawn on the basis of proximity. Though proximity is still commonly used, in the late 1990s and early 2000s, more direct and sophisticated measures of environmental health risk emerged through the availability of risk models. The USEPA working group in its 1992 report underscored the integration of environmental justice into risk-based approaches for environmental protection [18]. Models continued to be developed that often included a geographic component through the use of geographic information systems (GIS) [19, p. 246]. In that way, these models not only provided direct measures of risk, but also enabled proximity measures to be linked to them thus strengthening their use as a surrogate measure for environmental risk.

The 1990s saw some of the most devastating and catastrophic natural hazards ever experienced in the United States. The World Bank reported that 3.4 billion people worldwide were exposed to at least one natural hazard, and that the likelihood of experiencing some of the greatest losses and highest costs was higher in poorer countries [20]. Once again there were indications that exposure to and effects of these hazards were in many cases felt disproportionately by certain sectors of society. Studies of Hurricane Andrew, one of the most devastating and costly weather-related catastrophes, identified low-income and minority communities as those that were not able to seek shelter or recover [21]. The Gulf Coast hurricanes of August–September 2005 dealt a final blow. Low-income groups and communities of color experienced the worst of the devastation, since much of their housing was in low-lying areas. These communities experienced the largest burden in terms of the ability to escape and then recover and rebuild. Environmental concerns among racial minorities in the Gulf Coast area were recognized almost 10 years prior to the flooding [22]. Environmental justice continues to evolve as environmental issues have

changed. For example, it has now emerged as a dimension of sustainability [23, 24]. Five principles of equity that can be related to environmental justice are defined in the field of sustainable development: intergenerational equity, intragenerational equity, geographical equity, procedural equity, and interspecies equity [25].

International Issues in Environmental Justice

Environmental justice and fairness issues are also prominent in discussions about developing countries and global issues. Two areas where environmental justice is often invoked in international policy discussions are global climate change and the transboundary movement of toxic and hazardous substances. In the case of global climate change, the discussions center around past greenhouse gas emissions as countries expand their economies, and about projected emissions and their expected impact. This divide between developed and developing countries was prominent in discussions about the *Kyoto Protocol to the United Nations Framework Convention on Climate Change* [26]. The United States argued that it would not ratify binding emission reduction targets if developing countries such as China and India did not commit to reductions as well. Developing countries argued that it would be unfair for developed countries to have benefited from emitting large amounts of greenhouse gases as they developed and to now impose restrictions on poor countries that are attempting to raise the standard of living of their citizens [27, 28]. Measures of greenhouse gas emissions, such as carbon dioxide (CO₂), methane (CH₄), and nitrous oxides (NO_x) are presented as total emissions, emissions per capita, and emissions per unit of gross domestic product (GDP).

The other area where environmental justice plays a prominent role is in the regulation of transboundary movements of hazardous and toxic substances (*see Fate and Transport Models*). As developed countries implemented stricter environmental regulations since the 1960s, the cost of disposal of toxic and hazardous wastes increased substantially. A growing number of cases where these wastes ended up being dumped in developing countries without adequate disposal capabilities resulted in the growing concern about fairness and environmental justice (*see Hazardous Waste Site(s)*). Such cases are still being reported. In 2006, a large toxic sludge that originated in Europe was dumped untreated near a poor

neighborhood in Abidjan, Ivory Coast. The sludge was blamed for at least eight deaths and hundreds of cases of nausea, headaches, skin sores, nose bleeds, and stomach aches [29]. Concern about cases of toxic and hazardous waste dumping on developing countries with inadequate institutions and facilities to accommodate such wastes prompted members of the United Nations to ratify the Basel Convention, which was adopted in 1989. The main goals of the Basel Convention are to minimize the generation of hazardous waste, to safely dispose of these wastes as close as possible to their origin, and to regulate the transboundary movement of these wastes [30]. An additional concern is the international trade in e-waste, which refers to the disposal and the international trade of used computers and other electronic equipment. These products which are disposed of in developed countries often make their way to poor countries in Asia where they are taken apart and their constituent parts and materials are separated and reused. These activities take place under conditions that are often considered harmful and expose workers to high environmental risks [31].

Environmental justice and the distribution of environmental hazards are not limited to domestic discussions in developed countries. In the context of developing countries, lack of access to basic infrastructure services, such as potable water, wastewater disposal and treatment, and solid waste collection by underprivileged groups often leads to large human health burdens for these populations [32].

Methodologies to Discern the Evidence

A recent literature review of environmental justice studies concludes that people of color and lower socioeconomic status (SES) are exposed to higher levels of environmental risks according to outcome measures, such as proximity to hazardous waste facilities, exposures to air, noise, and water pollution, quality of housing and residential crowding, quality of schools and conditions of the work environment [33]. A well-documented example is lead levels in blood. Several studies have shown that children of minority communities and low-income groups have significantly higher levels of lead in the blood relative to the overall population [34]. An assessment of the environmental justice literature based on a meta-analysis of 49 studies concludes that there is strong

evidence of environmental inequities based on race but little evidence of environmental inequities based on economic class [19]. This section discusses some of the techniques used to assess environmental justice issues and the research challenges associated with this field of inquiry.

Environmental justice analyses have used a number of analytical tools including computer modeling, statistical modeling, and GIS to gain a better understanding of differential environmental risks to different populations. In the case of proximity analyses to sources of environmental risks, such as incinerators, waste disposal facilities, highways, and others, census data can be used to describe the neighborhood or community surrounding the source of risk. Among the socioeconomic characteristics included in such analyses in the United States are percentages of African-Americans, Hispanics, native Americans, households and individuals living below the poverty line, elderly people, and others. In examining the potential effects of proximity to facilities associated with air pollution emissions, additional measures that characterize vulnerable populations may also be included, such as percentage of children under the age of 17 who have had an asthma attack. Instead of using a single measure, more recent studies increasingly use multiple indicators that may also include rate of unemployment, types of employment, education level, and housing market characteristics such as price of housing and percentage of owners and renters [35].

Defining the neighborhood or community around a facility often represents a challenge to researchers since defining a unit of analysis that allows for a comparison with other units further away from the source of environmental risks is not always easy. Spatial units around the source of environmental risk and the census units typically used to obtain the characteristics of the geographical units are often different from each other. Moreover, the spatial distribution and density of people living within a unit of analysis may vary within each unit and among units. An early review of the issues involved in spatial definitions for environmental justice analysis was provided by Zimmerman [3]. In the United States, units of analysis used to obtain information about the area around a source of environmental risk include census block groups [36, 37] census tracts [36], zip codes [8], cities and metropolitan areas [4], and states [19]. Defining the unit of analysis can have important



Figure 1 Point buffers around a waste transfer station in Bronx County, New York [Reproduced from [36]. © New York University, 2002.]

consequences, and changing the scale of the unit of analysis may change the results of an environmental justice analysis [19, 37, 38].

Pioneering environmental justice studies used statistical modeling to predict the location of a source of environmental risk based on the characteristics of a neighborhood. Evidence of environmental injustice was reported if a positive statistical association between the likelihood of finding such a facility increased as the percentage of minorities and poor people increased in a neighborhood [7, 39].

GIS have become one of the most important analytical tools for environmental justice studies. Spatial coincidence studies map sources of environmental hazards and look for clustering of these sources. GIS maps of these types may also include socioeconomic characteristics of the populations in the area to see if

any clustering occurs in areas with high percentages of minority or disadvantaged groups. Another method used in GIS analyses is to create buffers around a source of environmental risk. The shape of the buffer depends on the distribution of the pollutant discharges around a source of risk. A line buffer looks at the characteristics within a specified distance along a line. Applications of line buffers include looking at the characteristics of populations around rail lines, highways or truck routes. Point buffers use circles around a point source of environmental risk, such as a hazardous waste handling facility. Figure 1 illustrates the use of point buffers and shows four buffers around a waste transfer station in Bronx County, New York. The radii of the buffers shown are 1/4, 1/2, 3/4, and 1 mile [36]. GIS is then used to extract information from sources such as the US census to understand the

characteristics of the population living in the specified buffer. These characteristics can then be compared with other areas, such as randomly generated buffers of equal size in other parts of the city or larger units, such as the city or the county.

Using buffers such as the ones shown in Figure 1 presents some important challenges when using data sources such as a census. The units of analysis of a census, such as census block groups, census tracts, and zip codes do not usually match the geometry of a point buffer of a specified radius. GIS software is used to extract the information by including only those units that are within the buffer or by making some assumptions about the units that are only partially within the buffer. Three methods are typically used to decide what type of information should be used in the estimates of population characteristics within a buffer [36]. In the polygon intersection method any unit that falls within the buffer, even if only partially, is included for estimates of population characteristics. This method tends to overestimate the results if significant portions of a unit of census data fall outside the buffer. The polygon containment method includes only those units that fall completely within the buffer. This method tends to underestimate population characteristics if a significant number of the census units are only partially within the buffer and thus not included in the estimates. The aerial interpolation method includes all units that are totally or partially within the buffer. However, for those units that are only partially within the buffer this method estimates the population characteristics by making the assumption that the population is evenly distributed within the unit or by using an area weighted technique and apportioning that data to the estimate. Depending on the available data and the characteristics of the populations in the units of analysis, this method can theoretically provide the most accurate results [40, 41].

A study in Bronx County, New York, used GIS to estimate the odds ratios of asthma hospitalizations based on proximity to point sources of environmental risks such as Toxic Release Inventory (TRI) sites and mobile sources of environmental risks such as proximity to highways and major truck routes. The risk of asthma hospitalization near sources of pollution was estimated by constructing buffers around these sources and comparing hospitalization rates within the buffers to those outside the buffers. The buffers around TRI facilities had a radius of one-half of a

mile and other stationary sources of pollution had a radius of one-quarter of a mile. Buffers around highways and major truck routes had a radius of 150 m. Overall, about 66% of the study area was included in the buffers. The study, which used aerial interpolation to estimate the risks of asthma hospitalizations based on proximity, found differential risks depending on the source of pollution. Living within the combined buffers is associated with 30% higher chance of being hospitalized for asthma relative to being outside the buffers. Living within the buffers around TRI sites and stationary sources of pollution is associated with 60–66% higher chance of being hospitalized for asthma relative to living outside the buffers. For highways and major truck routes the risk differential is not nearly as high but the author suggests this may be an artifact of the assumption made that the population around highways and truck routes are evenly distributed within the buffers [37].

An alternative type of proximity analysis is to examine the proximity of sources of environmental risk to point locations associated with sensitive populations of particular interest. Sensitive populations may include asthmatic children or the elderly and the point locations associated with these groups are schools and nursing homes, respectively. This type of analysis was also conducted in the South Bronx, New York, an area with a higher percentage of Hispanic residents (60%) and African-American residents (39%) than New York City as a whole [42]. The study estimated the number and percentage of prekindergarten to eighth grade students who attend schools located within 150 m of a major highway. This is the distance within which concentrations of air pollutants from vehicle traffic are believed to be significantly higher than background concentrations [43], thus justifying its use as a criterion in the analysis. The South Bronx has one of the highest rates of asthmatic children in the country and children are thus considered a sensitive population relative to sources of air pollution [44]. The study estimates that about 18% of students in the prekindergarten to eighth grade category attend schools within 150 m of a highway. The figure for New York city as a whole is only about 8%. This suggests that children attending school in the South Bronx are more likely to be exposed to environmental health risks associated with the high traffic densities associated with highways than children in other parts of New York City [45].

In understanding environmental justice around siting decisions it is important to distinguish between the present socioeconomic characteristics around a source of environmental risk and those characteristics at the time the facility was sited. In some cases, siting decisions were made in neighborhoods with minority or disadvantaged populations, whereas in other cases the neighborhoods change over time after the siting process [35]. Although proximity analyses provide important information and inputs to siting and land use decisions, recent environmental justice literature reviews suggest that the focus of environmental justice analyses should move beyond the kinds of studies described above to include individual environmental hazard exposures and multiple and cumulative exposures to hazards [33, 46].

Beyond proximity analysis, more sophisticated techniques to estimate exposures to environmental risks include distance-decay functions, wind direction and dispersion modeling, and weighted emissions and toxicity [35]. Liu presented an early approach to air quality by using air pollutant dispersion from sources [9]. In the area of air pollution and public health an increasingly popular method to measure exposures is to provide individuals with air pollution monitoring equipment that is carried at all times for a specified period of time. Exposures over time can then be compared to diaries of health symptoms such as asthma attacks, wheezing, and coughing. A study in Bronx County, New York suggests that increased individual exposures to elemental carbon, a pollutant closely associated with vehicle exhaust, is associated with increased wheezing and coughing symptoms among asthmatic children [47]. Exposure analyses of this kind carried out for individuals in different areas can be used to compare exposures across populations with different socioeconomic characteristics. Such methods can also help researchers understand the complex relationships between factors, such as poor nutrition, lack of access to health care, social stress, exposures to multiple environmental hazards, and health outcomes [35], thus enriching the basis for quantitative environmental health risk assessment.

Policy Implications for Risk Assessment and Risk Management

Environmental justice issues have shaped virtually every area of environmental policy and decision

making. The field of environmental impact assessment, which arose with the passage of the National Environmental Policy Act (NEPA) of 1969, began with environmental triggers as initiators of the process with social and economic impacts gradually added once an environmental trigger existed. These social and economic impacts started out as descriptive portrayals of populations and economies, and did not necessarily emphasize imbalances or disparities. In the past decade or so, however, environmental justice issues have emerged not only as major components of an environmental impact statement (EIS) but as triggers as well. Environmental reviews under NEPA and its state versions are critical platforms for environmental justice. The siting of natural gas turbines, described above, is one where environmental reviews became a focal point of the conflict. Not all justice arguments are antidevelopment; some are about development. For example, in 2007, arguments by African-American proponents of a bridge linking two African-American communities in South Carolina, have met with opposition from nonminority groups based on potential environmental impacts from the portion of the bridge crossing a water body [48].

On a broader level, the incorporation of environmental justice within public policy has been sporadic, and has emerged slowly since E.O 12898 was passed. In 2005, the USEPA produced a working draft providing a framework for integrating environmental justice into the strategic plan of 2006–2011 [49]. The strategic planning process encompasses goals and objectives and metrics to achieve them, in accordance with the Government Performance and Results Act (GPRA), submitted annually to the Congress [49: 2]. The working draft on environmental justice indicated that the plan preceding the 2006–2011 strategic plan contained environmental justice commitments only in one place – single goal, Goal 4 for “Healthy Communities and Ecosystems” and a single objective and subobjective within that goal. On November 5, 2005, the EPA Administrator issued a memorandum expanding the number of goal areas in the form of priorities for environmental justice to eight categories: “(1) reduce asthma attacks, (2) fish and shellfish safe to eat, (3) reduce exposure to air toxics, (4) water safe to drink, (5) reduced incidence of elevated blood lead levels, (6) collaborative problem solving, (7) ensure compliance, and (8) revitalization of brownfields and contaminated sites [50].” These

priorities were incorporated into the EPA Strategic Plan produced September 30, 2006 [17: 125, footnote 107].

Process issues are as critical as the substance of environmental justice. The environmental justice movement is marked by leadership of both well-known national environmental organizations and grass roots community and environmental organizations. The relationship between these two levels has been a complex dynamic. The community level as the driver has become a popular theme in the environmental justice movement, and one that was possibly adopted from the broader environmental movement that preceded it. The generation of knowledge and its use for decision making at the local level were identified in the field of public health and termed *popular epidemiology* by Brown [51] and continued for over a decade later as *citizen epidemiology* and *barefoot epidemiology* [52] and *street science* by Coburn [53].

The environmental justice movement has drawn the attention of both national and grass roots groups and organizations. Its affect on policy has gradually evolved in the decades following both the emergence of the environmental movement in the early 1970s and the environmental justice movement that followed. As the environmental area has taken new directions, environmental justice has moved with it. Global problems of deforestation, global climate change, ozone, radiation, and other transboundary problems have justice dimensions. Questions about process will persist with each new issue. What is clear, however, is that environmental justice is now an important component of risk assessment and risk management.

Acknowledgment

This research was supported by a grant from the USEPA for the South Bronx Environmental Health and Policy project, numbers 9821520-03, 982152-05, and 982152-06.

References

- [1] U.S. Environmental Protection Agency (EPA) (2007). *Environmental Justice*, at <http://www.epa.gov/compliance/resources/faqs/ej/index.html#faq2> (accessed Feb 2007).
- [2] Restrepo, C. (2006). *Asthma Hospital Admissions and Ambient Air Pollution in Four New York Counties: Bronx, Kings, New York and Queens*, Dissertation, New York University, Wagner Graduate School, New York.
- [3] Zimmerman, R. (1994). Issues of classification in environmental equity: how we manage is how we measure, *Fordham Urban Law Journal* **XXI**, 633–669.
- [4] Zimmerman, R. (1993). Social equity and environmental risk, *Risk Analysis* **13**, 649–666.
- [5] U.S. General Accounting Office (GAO) (1983). *Siting of Hazardous Waste Landfills and Their Correlations with Racial and Economic Status of Surrounding Communities*, GAO/RCED-83-168, Washington, DC.
- [6] Hamilton, J. & Viscusi, W.K. (1999). *Calculating the Risks: The Spatial and Political Dimensions of Hazardous Waste Policy*, MIT Press, Cambridge.
- [7] United Church of Christ, Commission for Racial Justice (1987). *Toxic Wastes and Race in the United States*, New York.
- [8] Goldman, B.A. (1991). *The Truth about Where You Live: An Atlas for Action on Toxins and Mortality*, Random House, New York.
- [9] Liu, T. (1996). Urban ozone plumes and population distribution by income and race: a case study of New York and Philadelphia, *Journal of the Air and Waste Management Association* **46**, 207–215.
- [10] U.S. Government, Executive Office of the President (1994). *Federal Actions to Address Environmental Justice in Minority Populations and Low-Income Populations*, Executive Order 12898, U.S. EPA, Washington, DC.
- [11] Rubin, V. (2006). *Safety, Growth, and Equity: Infrastructure Policies that Promote Opportunity and Inclusion*, PolicyLink, at <http://www.policylink.org/pdfs/Safety-GrowthEquity.pdf> (accessed Feb 2007).
- [12] Wright, R.T. (2005). *Environmental Science*, 9th Edition, Prentice Hall, Upper Saddle River.
- [13] Forkenbrock, D.J. & Schweitzer, L.A. (1999). Environmental justice in transportation planning, *Journal of the American Planning Association* **65**, 96–111.
- [14] Mills, G.S. & Neuhauser, K.S. (2000). Quantitative methods for environmental justice assessment of transportation, *Risk Analysis* **20**, 377–384.
- [15] Perez-Pena, R. (2001). State admits plants headed to poor areas, *The New York Times*, B1.
- [16] Gleick, P.H. (2000). The human right to water, in *The World's Water*, P.H. Gleick, ed, Island Press, Washington, DC.
- [17] U.S. Environmental Protection Agency (2006). *2006–2011 EPA Strategic Plan: Charting Our Course*, Washington, DC.
- [18] U.S. Environmental Protection Agency, Office of Policy, Planning and Evaluation, Environmental Equity Workgroup (1992). *Environmental Equity: Reducing Risk for All Communities*, Washington, DC.
- [19] Ringquist, E.J. (2006). Environmental justice: normative concerns, empirical evidence, and government action, in *Environmental Policy*, N.J. Vig & M.E. Kraft, eds, CQ Press, Washington, DC, pp. 239–263.
- [20] The World Bank (2005). *Natural Disaster Hotspots: A Global Risk Analysis*, Washington, DC.

- [21] Peacock, W.G., Morrow, B.H. & Gladwin, H. (1997). *Hurricane Andrew: Ethnicity, Gender and the Sociology of Disasters*, Routledge, London.
- [22] Burby, R.J. & Strong, D.E. (1997). Coping with chemicals: blacks, whites, planners, and industrial pollution, *Journal of the American Planning Association* **63**, 469–480.
- [23] Agyeman, J. (2005). *Sustainable Communities and the Challenge of Environmental Justice*, New York University Press, New York.
- [24] Agyeman, J., Bullard, R.D. & Evans, B. (2003). *Just Sustainabilities. Development in an Unequal World*, Earthscan, London.
- [25] Houghton, G. (1999). Environmental justice and the sustainable city, *Journal of Planning Education and Research* **18**, 233–243.
- [26] United Nations Framework Convention on Climate Change (UNFCCC) (2007). *Kyoto Protocol to the United Nations Framework Convention on Climate Change*, at <http://www.unfccc.int/resource/docs/convkp/kpeng.html> (accessed Feb 2007).
- [27] Jacoby, H., Prinn, R. & Schmalensee, R. (1998). Kyoto's unfinished business, *Foreign Affairs* **77**(4), 64.
- [28] Paarlberg, R. (1999). Lapsed leadership: U.S. international environmental policy since Rio, in *The Global Environment: Institutions, Law, and Policy*, N. Vig & R. Axelrod, eds, CQ Press, Washington, DC.
- [29] Polgreen, L. & Simons, M. (2006). Global sludge ends in tragedy for Ivory Coast, *The New York Times*, A1.
- [30] Secretariat of the Basel Convention (2007). *Introduction*, at <http://www.basel.int/pub/basics.html> (accessed Feb 2007).
- [31] The Basel Network Convention and Silicon Valley Toxics Coalition (2002). *Exporting Harm: The High-Tech Trashing of Asia*, at <http://www.ban.org/E-waste/techno/trashfinalcomp.pdf> (accessed Feb 2007).
- [32] Satterthwaite, D. (2003). The links between poverty and the environment in urban areas of Africa, Asia, and Latin America, *The Annals of the American Academy of Political and Social Science* **590**(1), 73–92.
- [33] Brulle, R.J. & Pellow, D.N. (2006). Environmental justice: human health and environmental inequalities, *Annual Review of Public Health* **27**, 103–124.
- [34] Evans, G. & Kantrowitz, E. (2002). Socioeconomic status and health: the potential role of environmental risk exposure, *Annual Review of Public Health* **23**, 303–331.
- [35] Szasz, A. & Meuser, M. (1997). Environmental inequalities: literature review and proposals for new directions in research and theory, *Current Sociology* **45**(3), 99–120.
- [36] Naphtali, Z. (2004). Environmental equity issues associated with the location of waste transfer stations in the South Bronx, in *South Bronx Environmental Health and Policy Study: Transportation and Traffic Modeling, Air Quality, Waste Transfer Stations, and Environmental Justice Analyses in the South Bronx*: C. Restrepo & R. Zimmerman, eds, *Final Report for Phase II & III*, at <http://www.icisnyu.org/admin/files/ICISPhaseIIandIIIreport.pdf> (accessed Feb 2007).
- [37] Maantay, J. (2007). Asthma and air pollution in the Bronx: methodological and data considerations in using GIS for environmental justice and health research, *Health and Place* **13**(1), 32–56.
- [38] Sheppard, E., Leitner, H., McMaster, R. & Tian, H. (1999). GIS-based measures of environmental equity: exploring their sensitivity and significance, *Journal of Exposure Analysis and Environmental Epidemiology* **9**(1), 18–28.
- [39] Goldman, B. & Fitton, L. (1994). *Toxic Wastes and Race Revisited*, Center for Policy Alternatives, Washington, DC.
- [40] Liu, F. (2000). *Environmental Justice Analysis Theories, Methods and Practice*, Lewis Publishers, New York.
- [41] Forkenbrock, D. & Sheeley, J. (2004). *Effective methods for environmental justice assessment*, National Cooperative Highway Research Program (NCHRP) Report 532, Transportation Research Board, at http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_532.pdf (accessed Feb 2007).
- [42] Zimmerman, R., Restrepo, C., Hirschstein, C., Holguín-Veras, J., Lara, J. & Klebenov, D. (2002). *South Bronx Environmental Studies, Public Health and Environmental Policy Analysis: Final Report for Phase I*, at <http://www.icisnyu.org/assets/documents/SouthBronxPhaseIReport.pdf> (accessed Feb 2007).
- [43] Hitchins, J., Morawska, L., Wolff, R. & Gilbert, D. (2000). Concentrations of submicrometer particles from vehicle emissions near a major road, *Atmospheric Environment* **34**, 51–59.
- [44] Garg, R., Karpati, A., Leighton, J., Perrin, M. & Shah, M. (2003). *Asthma Facts, 2nd Edition*, New York City Department of Health and Mental Hygiene.
- [45] Restrepo, C., Naphtali, Z. & Zimmerman, R. (2006). *Land Use, Transportation and Industrial Facilities: GIS to Support Improved Land Use Initiatives and Policies, the South Bronx, NYC*. Institute for Civil Infrastructure Systems (ICIS), at http://www.icisnyu.org/south_bronx/admin/files/HandoutWagnerOct162006.pdf (accessed Feb 2007).
- [46] Northridge, M.E., Stover, G.N., Rosenthal, J.E. & Sherard, D. (2003). Environmental equity and health: understanding complexity and moving forward, *American Journal of Public Health* **93**(2), 209–214.
- [47] Spira-Cohen, A., Chen, L.C., Clemente, J., Blaustein, M., Gorczynski, J. & Thurston, G.D. (2006). A health effects analysis of traffic-related PM pollution among children with asthma in the South Bronx, NY, Abstract presented at *The International Conference on Environmental Epidemiology and Exposure*, Paris, September 2–6 2006.
- [48] Nossiter, A. (2007). Race, politics and a bridge in South Carolina, *The New York Times*, at <http://www.nytimes.com/2007/02/25/us/25bridge.html> (accessed Feb 2007).

- [49] U.S. Environmental Protection Agency (2005). *Framework for Integrating Environmental Justice*, Washington, DC.
- [50] Johnson, S.L. (2005). *Reaffirming the U.S. Environmental Protection Agency's Commitment to Environmental Justice*, Memorandum, U.S. EPA, Washington, DC, at <http://www.epa.gov/compliance/resources/policies/ej/admin-ej-commit-letter-110305.pdf> (accessed Feb 2007).
- [51] Brown, P. (1993). When the public knows better: popular epidemiology, *Environment* **35**, 16–40.
- [52] Nuclear Energy Information Service (NEIS) (2005). News citizen epidemiology—the next step, *Citizen Epidemiology: The Next Step* **24**(1), 1–8.
- [53] Coburn, J. (2005). *Street Science*, MIT Press, Cambridge.

Related Articles

Risk and the Media

Scientific Uncertainty in Social Debates Around Risk

Stakeholder Participation in Risk Management Decision Making

Statistics for Environmental Justice

CARLOS E. RESTREPO AND RAE ZIMMERMAN

History of Epidemiologic Studies

Epidemiology can be defined as the scientific method used to assess health and its determinants in populations. Central to the method are measures of disease occurrence and study designs for identifying causes of disease and quantifying their strength as causal agents. Because epidemiological studies directly address threats to human health, their findings, when available, have received emphasis in qualitative and quantitative risk assessments.

The utility of epidemiologic data for risk assessment has been widely discussed [1, 2]. Pundits argue that epidemiologic data are not available for many agents and that they are too often flawed by poor quality and biases that cannot be fully controlled [3]. Epidemiologic studies have also been deemed as uninformative for the “weak associations” anticipated for typical levels of exposure to many environmental agents. Proponents of use of epidemiologic evidence, while acknowledging the limitations of observational studies, advance its strengths: the investigation of the effects of actual exposures as received by the population; the characterization of effect across the full range of susceptibility in the population; and, above all, the direct relevance of epidemiologic evidence to public health [4, 5]. Guidelines for the conduct of epidemiologic research relevant to risk assessment have been offered as one solution to strengthening the evidence base [4, 6, 7].

Epidemiologic data can be relevant to each of the four components of risk assessment, as set out in the 1983 “Red Book” of the National Research Council [8]. In particular, the findings of epidemiological studies have proved critical in establishing causation for some of the most important environmental causes of human disease: for example, cigarette smoking, radiation, and asbestos. Principles for evaluating epidemiological evidence within the broader context of other lines of scientific evidence have been developed and applied for decades [9]. Establishing causation is one potential outcome of hazard identification, the starting point in risk assessment (*see Causality/Causation*). The principles originally developed and applied by epidemiologists in evaluating evidence for causation have now been adapted into “weight-of-evidence” schemes. In the absence of

epidemiological findings, hazard identification rests on animal toxicology and consideration of mode of action.

Epidemiologic data indicative of an adverse effect of an agent on health, when available, are strongly weighted in the evaluation of the weight of evidence to determine if an agent presents a hazard. Human data provide direct evidence of a hazard without the need to extrapolate from knowledge of toxicity of analogous agents or from another species. In fact, as we have gained a deeper understanding of the complexity of cross-species extrapolation from animal to man, such extrapolations are viewed with less certainty, unless based in an understanding of human and animal pathways of absorption and metabolism, and of mechanisms of action. Further, epidemiologic studies evaluate the impact of exposures received by the population, including complex mixtures which may not be readily replicable in the laboratory. Epidemiologic research captures the consequences of interactions among agents, and investigations in populations may capture the full range of susceptibility.

If the data collected in an epidemiological study capture dose or a surrogate for dose (generally exposure), then the study may be informative on the nature of the dose–response relationship (*see Dose–Response Analysis*). We consider limitations of the epidemiological data for this purpose and approaches for informatively analyzing epidemiological data for dose–response assessment. Some epidemiological studies, particularly those based in populations, may also inform exposure assessment. Finally, the epidemiological measure, the attributable risk, estimates the burden of disease resulting from a particular exposure. Estimates can be made for those exposed or for the full population comprising both exposed and nonexposed.

History of Epidemiology in Risk Assessment

Even before the formalization of risk assessment, epidemiological studies were generating data relevant to hazard identification, dose–response assessment, and risk characterization. A century ago, the principal focus of epidemiological inquiry was infectious diseases; at the start of the twentieth century in the United States and other industrialized countries, infectious diseases were the leading causes of

morbidity and mortality. As the frequency of these diseases declined and lifespan lengthened, a new group of diseases – so-called “chronic diseases” – emerged as a predominant cause of morbidity and mortality. These diseases included cancer, coronary heart disease, and chronic bronchitis and emphysema, now referred to as *chronic obstructive pulmonary disease* (COPD). In parallel with the epidemic of chronic diseases, the discipline of “chronic disease epidemiology” evolved from the traditions of infectious disease epidemiology, with continued evolution and the development of new approaches.

The causes of cancer, coronary heart disease, and chronic lung disease were the initial targets for chronic disease epidemiologists. By the middle of the century, mortality from lung cancer and heart disease were rising, and chronic lung diseases were increasingly prevalent. The goal of the first wave of studies was identification of causal risk factors, equivalent to hazard identification in risk assessment (see **Causality/Causation**). Many of these studies and their findings are well known: *case-control studies* (see **Case-Control Studies**) and *cohort studies* (see **Cohort Studies**) of smoking and lung cancer, and the Framingham and other cohort studies of the causes of coronary heart disease. The data from these studies were largely analyzed to determine if the putative causal factors were associated statistically with risk for the health outcome. Beginning with the decade of the 1950s, a lengthening list of causes of chronic diseases was identified: cigarette smoking, asbestos exposure (see **Asbestos**), hypertension, and alcohol consumption, for example. These same methods are still used and continue to be informative for identifying new etiologic factors.

As the extent of the epidemiological evidence on particular factors deepened, the need was recognized for an approach for evidence synthesis with the goal of causal inference. Criteria were proposed for this purpose, most notably by Sir Austin Bradford Hill in the United Kingdom and the members of the Advisory Committee to the US Surgeon General, authors of the landmark 1964 Report on smoking and health [10–12]. These criteria were intended to guide an integrated assessment of observational evidence with findings from other lines of research (Table 1). The criteria of consistency and strength of association have proved useful in evaluating the potential role of bias in producing an association. The criterion of coherence addresses the plausibility

Table 1 Sir Austin Bradford Hill’s causal criteria^(a)

Aspects of association to be considered before deciding on causation

Strength
Consistency
Specificity
Temporality
Biological gradient
Plausibility
Coherence
Experiment
Analogy

^(a) See Reference [10]

of a causal association. The concepts embodied in these criteria are evident in schemes for evidence evaluation used by diverse agencies. The 2004 Report of the US Surgeon General reviews the evolution of these criteria and categories for classifying the level of certainty of causal inference [12].

Early in the development of chronic disease epidemiology, the need for risk characterization was identified. In 1953, Morton Levin published a key paper that described an epidemiological parameter, now generally referred to as the population attributable risk (PAR). By the early 1950s, there was already substantial evidence that smoking was strongly associated with lung cancer risk. Levin reasoned that for a causal factor, it would be useful to know the risk both for an exposed individual and for an exposed population. The PAR estimator proposed by Levin is:

$$PAR = \frac{P_E(RR - 1)}{1 + P_E(RR - 1)} \quad (1)$$

where P_E is the population prevalence of exposure and RR is the relative risk associated with the exposure. The PAR ranges from 0 to 1.0 (0–100%), and its value increases with both P_E and RR . The PAR offers a risk characterization that reflects the exposure distribution and the excess risk from the exposure. It can be generalized to incorporate multiple levels of exposure, each having an associated RR value [13].

The epidemiological data on risk factors often covered multiple levels of exposure; assessment of the relationship between increasing exposure, dose, or a surrogate was an objective analytic strategy in many investigations. Evidence of a dose–response

relationship is supportive of a causal association, not only because a dose–response relationship may be consistent with understanding of the mechanism of action, but because sources of bias become less tenable as an explanation for the association. The approaches used initially involved categorization of exposure into ordered categories. Stratification was used to control for confounding, i.e., the unwanted influence of other factors, and methods were developed for dose–response assessment from stratified data [14].

For many chronic diseases, there are multiple risk factors, and exploration of the dose–response relationship for one may require the simultaneous control of several others. Such control is most conveniently and informatively accomplished using regression models. Beginning in the 1960s, regression models were used for this purpose and subsequently a wide range of models has been developed to accommodate the variety of outcomes and data structures spanned by epidemiological research. For the purpose of risk assessment, some of the most commonly used models are multiple logistic regression, Poisson regression, and the Cox proportional hazards models, used for binary outcomes, incidence, and time-to-event, respectively. More recently developed models handle repeated and correlated events and exposures.

One of the earliest examples of analysis of epidemiological data for risk assessment can be found in the estimation of radiation risks using the cohort study of Japanese atomic bomb survivors. The initial cohort was established within five years of the 1945 bombings and follow-up of blast survivors continues to now. The first excess occurrence of malignancy came with a wave of acute leukemia in both children and adults. By the mid-1950s, there was an evident excess risk, which was dose-dependent.

In a landmark 1957 paper, Lewis assessed the relationship between excess leukemia risk and radiation dose from the blast and concluded that a linear representation was reasonable [15]. Lewis graphically examined accruing data from the atomic bomb survivors that described how risk increased in relation to estimated radiation dose from the blast. He found the relationship to be consistent with linearity and, based on biological considerations, he selected a model without a threshold. As follow-up of the cohort continued, excess risk for solid tumors was also evident. Dose–response relationships have now been repeatedly assessed for cancers [16].

The present Lifespan Study Cohort emerged from the population assembled in the early 1950s. The data have proved to be an extraordinary resource for risk modeling, offering a large population with well-characterized exposure to a single agent and careful follow-up for cancer incidence and mortality from cancer and other diseases. The data have proved to be a rich resource for risk modeling of epidemiological data, as follow-up lengthened and excess solid tumor occurrence was manifest. Increasingly sophisticated statistical models have been used to examine how cancer risk varies with dose and modifying factors such as age at exposure and attained age. Over time, the close estimates have been refined and statistical approaches developed to account for measurement error in these estimates.

Other radiation-exposed groups have also been included in epidemiological studies and the resulting data used for risk models. One notable example is the cohort studies of underground uranium miners who received continuous, but time-varying exposure to radon underground (*see Radon Risk*). The data from 11 cohorts were pooled during the 1990s by the investigators and also by the Biological Effects of Ionizing Radiation (BEIR) VI Committee. Models were developed that allowed radon-related risk to vary by a set of time-related factors: age at first exposure, time since exposure, attained age, cumulative exposure, and exposure rate. Because of the size of the data set (>68 000 participants) and the number of lung cancer deaths (2700), the analyses proved informative, and a relative risk model was developed that estimated risk by time since exposure, with modification of risk by attained age and exposure rate. Data from other radiation-exposed groups have been combined: persons receiving therapeutic radiation and nuclear workers, for example [17].

The “lessons learned” in assessing radiation risk were extended to other exposures, typically involving worker groups exposed to carcinogens at doses sufficient to cause detectable excess cancer risk. Examples include asbestos and lung cancer and mesothelioma in diverse worker groups, and benzene and leukemia. For cancer, epidemiological data from general population studies have also been used in cancer risk assessments: e.g., arsenic and drinking water (*see Arsenic*) and secondhand smoke (SHS) and lung cancer (*see Environmental Tobacco Smoke*). The examples include general population samples with high exposures (arsenic), and typical exposures

Table 2 Recent application of epidemiologic data in risk assessment

Topic	Application of epidemiologic data
Airborne particulate matter (PM)	Background - Under the Clean Air Act, the EPA is required to review the National Ambient Air Quality Standards. Use of epidemiology - A risk assessment was an important component of the review, quantifying the number of premature deaths and hospitalizations due to PM and the number that could be avoided due to attainment of current or alternate standards [18]. - This risk assessment was based on the significant body of epidemiologic evidence relating particulate matter to increased mortality and morbidity risk, including daily time-series studies [19, 20] and long-term cohort studies [21, 22]. - The NMMAPS^(a) was a key piece of epidemiologic evidence, which quantified the relationship between PM and mortality and hospitalizations in the 90 largest cities in the United States.
Methyl mercury	Background - The establishment of a US EPA RfD^(b) for methylmercury was driven by epidemiologic evidence. This RfD^(b) would be used to guide the Agency's future risk management policies. Use of epidemiology - The National Research Council was asked to conduct a risk assessment to help guide the development of the RfD^(b) [23]. - Early data on health effects of methylmercury from accidental poisonings were documented in Minimata Bay, Japan; and Iraq. - Adverse neurodevelopmental effects from contaminated fish consumption were observed in studies from New Zealand and the Faroe Islands, which were key studies used in the development of the RfD^(b) [24, 25]. - The body of epidemiologic evidence regarding the health effects continues to grow, with recent research suggesting a role in cardiovascular disease.
Arsenic in drinking water	Background - In 2001, the US EPA adopted a standard of 10 ppb for arsenic in drinking water under the Safe Water Drinking Act. The standard represented a reduction from the 50 ppb interim standard that had been in place. Use of epidemiology - The US EPA characterized the risk of chronic arsenic ingestion in a 1998 report [26]. - The adoption of a more protective drinking water standard followed a 1999 National Resource Council report evaluating EPA's characterization of risk. - This report and EPA's risk characterization relied on epidemiologic evidence from Taiwan, Chile, and Argentina [27].
Lead	Background - In the 1920s, tetraethyl lead was introduced into automobile gasoline to reduce engine knocking and improve performance. Lead was also commonly used in household paints and other products. Use of epidemiology - The 1970 Clean Air Act required air pollution standards for lead and a gradual phase out of lead in gasoline. This was accelerated through the introduction and widespread use of catalytic converters, which were not compatible with leaded gasoline. In 1978, lead-based paint was banned for use in housing. - Toxic effects of high doses of lead had been known for some time. Regulatory actions led to dramatically reduced blood lead levels among children in the United States.

(continued overleaf)

Table 2 (continued)

Topic	Application of epidemiologic data
	• Epidemiologic research in the late 1970s first identified an association between blood lead levels previously thought to be safe and decrements in neurocognitive function (measured by IQ tests and other scales) [28]. This evidence led the CDC to develop guidelines for treatment of lead poisoning and set a “level of concern” in children of $10 \mu\text{g dl}^{-1}$. • More recent epidemiologic evidence has suggested that neurocognitive effects may occur in children below this concentration, and that there may be no threshold for these effects [29].
Active smoking	Background • In the early part of the twentieth century, cigarette consumption increased dramatically in the United States, including a surge during and after World War II. Use of epidemiology • In the 1950s, a series of case–control studies suggested an association between smoking and lung cancer [30–32] and the first major cohort study, “the British Physicians’ study” was begun [33]. • The accumulated body of evidence was summarized in the landmark 1964 report of the US Surgeon General [11], in which epidemiologic data were considered a key line of evidence. On the basis of this evidence, the report concluded that cigarette smoking was a cause of lung cancer. • Subsequent Surgeon General’s reports on tobacco and health have been released to update the weight of epidemiologic and other evidence and highlight specific diseases (e.g., cardiovascular disease (CVD), COPD), subpopulations (e.g., youth, women, racial/ethnic minorities), and topics such as smokeless tobacco use, smoking cessation, and involuntary smoking. • Epidemiologic data has been used to estimate the attributable burden of disease due to smoking, both in the United States [34] and globally [35].
Secondhand smoke (SHS)	Background • Following the increasing awareness of the harms of tobacco smoke, attention turned to the harms of exposure to SHS. Use of epidemiology • The 1986 Surgeon General’s report identified SHS as a cause of lung cancer and other disease in nonsmokers. This assessment was based on epidemiologic and other lines of evidence [36, 37]. • In a 1992 risk assessment, the US EPA listed SHS as a Group A (known human) carcinogen, making it subject to federal regulation and workplace restrictions [38]. • The risk assessment relied on over 30 epidemiologic studies for cancer alone, along with other lines of evidence to draw its conclusions and quantify the burden of disease due to SHS exposure. According to the report: • “The availability of abundant and consistent human data, especially human data at actual environmental levels of exposure to the specific agent (mixture) of concern, allows a hazard identification to be made with a high degree of certainty.” • Using a pooled relative risk from epidemiologic studies, the report concluded that SHS is responsible for 3000 lung cancer deaths in nonsmokers and 150 000 to 300 000 lower respiratory tract infections in infants, annually. • Since then, updated reports summarizing the epidemiologic evidence and policy implications have been prepared by International Agency for Research on Cancer [39] and the US Surgeon General’s office [40].
Trans fats	Background • Trans fats, produced through the partial hydrogenation of vegetable oils, are used in many baked goods. Use of epidemiology • Intake of trans fats causes inflammation and adverse effects on serum lipids and endothelial cell function [41].

Table 2 (continued)

Topic	Application of epidemiologic data
Global burden of disease	• Evidence from cohort and case–control studies have shown that intake of trans fats increases the risk of coronary heart disease (CHD) “more than any other macronutrient” on a per-calorie basis [41]. • Attributable risk calculation suggests that near elimination of trans fats could result in 72 000–228 000 fewer CHD events annually in the United States [41]. • On the basis of this evidence, the US FDA has required product labeling and some jurisdictions (Denmark, New York) have eliminated the use of these man-made fats.
	Background
	• The World Health Organization’s Global Burden of Disease project attempts to quantify the overall burden of morbidity and mortality and the proportion due to modifiable risk factors in different regions of the world.
	Use of epidemiology

^(a) NMMAPS: National Morbidity Mortality and Air Pollution Study

^(b) RfD: reference dose

(SHS). The range of epidemiological data underlying recent risk assessment is broad (Table 2) and relates to both cancer and noncancer outcomes.

Uses of Epidemiology in Risk Assessment

Hazard Identification

As described above, epidemiologic data are typically favored over animal toxicology in assessing the existence of a hazard. In using epidemiologic data for the hazard identification step, interpretation of the evidence is fully parallel to the assessment of the causality of an association between an exposure and an adverse health effect. There are no specific guidelines for interpretation of epidemiologic data in risk assessments that go beyond the conventionally applied criteria for causality. As with the widely applied criteria for causality, there may be disagreement on the proper classification of epidemiologic evidence for the purpose of hazard identification.

The interpretation of negative epidemiologic findings in the hazard identification process is a major

source of disagreement. Studies construed as “negative” typically have findings that are not statistically significant; i.e., the data are statistically consistent with no association between agent and adverse outcome. Such “absence of association”, however, is not equivalent to there being no effect. There are few instances in which epidemiological data have been strongly supportive of no association.

Dose–Response Assessment

For the purpose of risk assessment, characterization of the exposure–response relationship in the range of human exposures is needed. These data may be available from epidemiological studies. For some agents, e.g., SHS and indoor radon data on risks are available in the exposure range of interest, and risk to the general population can be estimated directly; for others, data from exposures above those received by the population may be available from worker groups, e.g., radon exposures of uranium miners, or from persons who have been accidentally exposed, e.g., radiation exposures of the survivors of the atomic

bomb blasts in Hiroshima and Nagasaki. For exposures observed above the range of usual environmental levels, exposure–response relationships estimated at the higher exposures are extended downwards.

Exposure (or dose) may be related to disease risk in a pattern that varies with the temporal profile of exposure and with interactions with other factors, such as age, genetic susceptibility, and environmental factors. Epidemiologic studies rarely include comprehensive data on exposure or dose during the biologically relevant interval, and exposures are often estimated from incomplete data. Consequently, the analyses may not reflect the underlying biological process, and the model applied for data analysis may not give a correct picture of the dose–response relationship.

A further issue is the possibility of error in the exposure indicators, resulting in misclassification of individuals as to their exposure status or level of exposure. Consequently, misclassification of exposure may bias the description of the dose–response relationship. Simple generalizations concerning the consequences of measurement error cannot be made [43]. Random error or nondifferential misclassification tends to blunt the dose–response relationship but other effects may also occur [44]. Nonrandom errors or differential misclassification may increase or decrease the gradient of the dose–response relationship.

In analyzing epidemiologic data to characterize the dose–response relationship, there is *a priori* interest in determining if the dose–response relationship is statistically significant – i.e., whether the null hypothesis of a flat dose–response relationship can be rejected – and in characterizing the shape of the relationship. For the former, the statistical significance of the trend in relation to exposure is of interest. Initially, the shape of the dose–response relationship is explored descriptively, but inevitably statistical models are used to quantify the change in response with increasing exposure. To date, models have been used primarily to characterize the effects of carcinogenic agents; less effort has been directed at noncancer effects, because human data have been less abundant.

For cancer, a number of alternative models have been used to characterize the association between exposure and risk (Figure 1) [8]. Arguments for the biologic plausibility of some of the key alternatives can be made, and use of the alternative models may lead to results with substantially different public

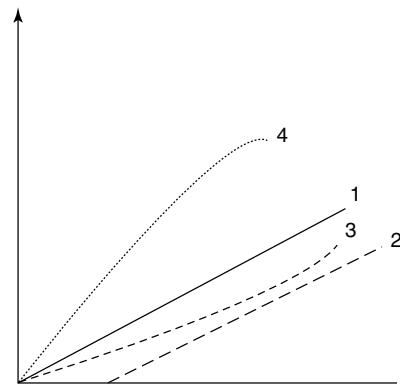


Figure 1 Examples of dose–response models used for carcinogens: (1) linear nonthreshold mode; (2) linear threshold model; (3) sublinear nonthreshold model; and (4) supralinear nonthreshold model [Reproduced from [8]. © National Academy Press, 1983.]

health and regulatory implications. The models differ in the presence of a threshold of exposure that must be exceeded before cancer occurs and in the shape of the relationship between exposure and response. In addition to a linear relationship between exposure and response, sub and supralinear alternatives may also be plausible.

A model of the exposure–response relationship might be estimated from epidemiologic data to extrapolate from higher exposures, where observations have been made, to the lower levels at which general population exposures occur, or to describe the exposure–response relationship quantitatively at typical exposures. Steenland and Deddens [45] supply some practical guidance on modeling exposure–response relationships for risk assessment. They favor simple parametric models. In modeling epidemiologic data, *a priori* biologic considerations should determine the choice of the model to the extent possible. For example, a linear relationship without a threshold would be consistent with an exposure that caused irreparable genetic damage with a single “hit” from the agent. Genetic injury from α particles released by deposited radionuclides, “internal emitters”, is an example [46]. For some agents, modeling approaches have been established that have a biologic rationale and historical precedent. For ionizing radiation, a linear no-threshold relationship has long been assumed [17]. The assumption that there is no threshold is also protective of public health as it implies that no level of exposure is safe.

Epidemiologic data may also be fit with alternative models of the exposure–response relationship if there is uncertainty about the most appropriate shape. Model fit may be used to guide the selection of the “best” model. However, epidemiologic data are rarely sufficiently abundant to provide powerful discrimination among alternative models, and sample size requirements for comparing fit of alternative models having different public health implications may be very high [47, 48]. For example, Lubin and Gail [48] calculated sample size requirements for testing if the exposure–response relationship for radon and lung cancer differed by 100 percent from that observed in underground miners. The calculations indicated that approximately 10 000 participants would be needed in a case–control study if sufficient power were to be achieved.

Exposure Assessment

For the purpose of risk assessment, information is needed on the full distribution of exposures in the population. Measures of central tendency may be appropriate for estimating overall risk to the population, but reliance on central measures alone may hide the existence of more highly exposed persons with unacceptable levels of risk. For example, the average concentration of indoor radon in US homes is about 1 picocurie per liter (pCi l^{-1}) of air (a measure of radioactivity) [46]. About 5% of homes are above the EPA guideline for remedial action of 4 pCi l^{-1} and some homes have concentrations of 1000 pCi l^{-1} or more. Focusing solely on control measures to lower the average exposure would obscure the need to find the homes in the tail of the distribution which have unacceptable levels. The EPA has recognized the need to characterize the upper end of the exposure distribution in its exposure assessment guidelines [49].

To date, epidemiologic data have not played prominent roles in exposure assessment. However, epidemiologic-based surveillance and screening approaches have the potential to provide important insights into population exposures. For example, blood lead surveillance has been a critical part of understanding the risks of environmental lead. Population-based survey data such as the National Health and Nutrition Examination Survey (NHANES) may be available on exposures, often describing contact with sources, e.g., exposure to

smokers as an indicator of exposure to SHS [50], or levels of biomarkers, e.g., cotinine [40].

Risk Characterization

Risk characterization represents the final step in risk assessment. This step puts the risk in perspective for risk managers and the public. During this phase, the data on exposure are combined with the exposure–response relationship to estimate the potential risk posed to exposed populations. The risk characterization becomes the fulcrum for decision-making and the basis for communication with stakeholders [51]. Epidemiologic studies may provide a direct risk characterization if the findings can be linked to a specific population and the exposure assessment component of the study is sufficient.

The PAR and years of life lost, both measures of the burden of imposed morbidity or mortality, are appropriate parameters for risk characterization. The US Centers for Disease Control and the World Health Organization (WHO) have used such measures to estimate risk from the major modifiable causes of death and disease in the United States and globally [34, 52]. In the United States for example, Mokdad *et al.* estimates that about 800 000 deaths annually are attributable to tobacco use and poor diet/physical inactivity [34]. Similarly, WHO’s Global Burden of Disease project estimates the number of avoidable deaths, years of life lost, and disability adjusted life years lost. The latter two indicators take into account the age of death and onset of disease [52].

Strengthening the Use of Epidemiology

The continued strengthening of the quality of epidemiologic studies should lead to its larger role in risk assessment. Approaches are available to quantify and adjust for measurement error, and flexible statistical modes may be incorporated to minimize potential confounding. The continued development of analytical laboratory methods to measure low-level dose biomarkers in human tissues, along with the use of flexible statistical models and meta- or pooled analyses, will allow for improved precision of risk estimates and a more precise exploration of dose–response relationships. Such improvements would minimize the need for the host of default uncertainty factors typically incorporated into human health risk assessment.

Minimizing Bias and Confounding in Epidemiologic Studies

A principal challenge in the conduct and interpretation of epidemiological findings is to minimize the presence of bias and *confounding* (see **Confounding**). *Confounding* may result in a distortion of the true exposure–disease relationship arising from the presence of a third factor, which is both associated with the exposure and an independent cause of the disease. Failing to address confounding through the study design or in the data analysis may result in biased estimates of the association between exposure and disease. *Bias* refers to “systematic error in the design or conduct of a study” [53]. Such an error can arise during sample selection (selection bias) or data collection (information bias), and will result in systematic departures from the true relationship. This is in contrast to *random* error, which results from the random selection of a sample of the population, and which can be characterized quantitatively through the use of confidence intervals.

Several methods exist to assess the extent and implications of bias or uncertainty. Random sampling error is characterized through the use of confidence intervals, which describe a range of estimates that captures the true value with some degree of certainty. Confounding in epidemiological studies is addressed through either restricting the sample population to only one level of a potential confounder (e.g. non-smokers) or through statistical methods for deriving adjusted estimates. It is considerably more difficult to ensure that bias (or systematic error) is minimized in epidemiological studies. Some approaches include the use of validation studies or sensitivity analyses, to examine both the potential for bias and direction of impact on the relationship under study.

Validation studies can be conducted to quantify and adjust for measurement error, assuming the presence of a gold-standard measure [54]. For exposure measurement error, the gold-standard is often a biological marker of dose. Such studies can be conducted on a subset of the study population to adjust exposure assessment for the full sample. Sensitivity analyses are useful approaches to quantify the impact and direction of other potential biases. Flexible statistical modeling approaches, including generalized additive models and nonparametric regression, can be used to reduce residual confounding and allow for fewer assumptions when estimating the shape

of dose–response relationships, specifically when dealing with large sample sizes [54].

Combining Evidence

Meta-Analysis. Epidemiologic studies may have limited power to address the effects of exposures in a range that would be of concern for the general population. Meta-analysis (see **Meta-Analysis in Nonclinical Risk Assessment**) has been used to summarize data for the steps of hazard identification and dose–response assessment. It has been used increasingly as a tool for summarizing experimental and observational studies [55] and methods for epidemiologic data have been extensively reviewed [56, 57]. Meta-analysis provides a potentially more informative interpretation of the evidence when sampling variation and small studies have obscured the status of the evidence. Thus, statistical power for detecting an effect is gained, and dose–response relationships can be described with greater precision. Meta-analysis has now been applied to a number of environmental agents including, for example, SHS [38], electromagnetic radiation [51], nitrogen dioxide [58], and indoor radon [59].

Use of meta-analysis to characterize the relationship between indoor air pollution and lung cancer risk is instructive. A number of case–control studies on indoor radon and lung cancer were initiated over the last decade to estimate the risk of indoor radon directly and to avoid the uncertainty arising from extrapolation of risks from underground miners to the general population. Sample size needs of these studies were not fully appreciated as they were designed [48], and a number of studies have had null findings but wide confidence intervals on effect estimates. Lubin and Boice [60] carried out a meta-analysis of eight case–control studies that were completed as of mid-1996. Although five of the individual studies were considered to be “negative”, the overall effect of indoor radon on lung cancer was statistically significant and the dose–response relationship was fully compatible with observations in miners.

Pooled Analysis. Pooled analysis refers to the analysis of individual-level data from multiple studies, unlike meta-analysis which aggregates data at the study level. This approach has been informative in characterizing dose–response relationships for radiation and cancer and for evaluating modifiers of

these relationships. Pooled analyses on radiation risks have been published for nuclear workers [61], underground miners [46, 62], and for women who received repeated fluoroscopy in the management of tuberculosis [63].

The analyses of data from cohort studies of radon-exposed underground miners are illustrative. Data on dose–response relationships can be obtained from individual studies, but variation exists among the studies in the magnitude of the estimated excess risk. To characterize the dose–response relationship between indoor radon and lung cancer, the BEIR IV Committee obtained data from four epidemiologic studies, including about 22 000 subjects who had experienced 350 lung cancer deaths during follow-up. The data were analyzed with regression methods to characterize the dose–response relationship. A linear model was used, and the risks were found to decline with increasing time since exposure and increasing age at observation. With the completion of additional studies, the data available for pooling expanded to 68 000 men and 2700 lung cancer deaths drawn from 11 cohorts [62]. In the resulting model, lung cancer risk was further found to increase with lower exposure rates; i.e., at a given level of cumulative exposure those receiving exposures at lower rates have higher risk. Because the data base included smoking information for a substantial number of participants, risks were also estimated separately for smokers and for nonsmokers.

Ranking Quality/Utility of Epidemiologic Evidence for Risk Assessment

In using epidemiologic data in risk assessment, interpretation of the evidence is fully parallel to the assessment of the causality of an association between an exposure and an adverse health effect. This assessment of causality is a judgment based on an evaluation of all available evidence. The conventional set of criteria most often referred to in the context of assessing causality was proposed by Hill [10]. These criteria, which include the strength of association, consistency between studies, and biological plausibility, were intended as guidelines, not hard and fast rules. There are no specific guidelines for interpretation of epidemiologic data in risk assessments that go beyond these conventionally applied criteria. However, several frameworks for classifying epidemiologic evidence for use in risk assessment

have been presented. A classification for ranking epidemiologic studies for use in hazard identification and dose–response assessment has been proposed by Hertz-Picciotto [4] and modified by van den Brandt *et al.* [54]. This classification scheme is based on both the internal validity (i.e., extent of potential bias and confounding) of the study and the characterization of exposure and dose–response. Swaen [64] describes a classification scheme which also focuses on the internal validity of the study, including the use of an appropriate study design, minimizing bias and confounding, and reducing measurement error in the exposure and outcome.

Summary

Epidemiological evidence has a central role in risk assessment, when available. It can contribute to all elements of a quantitative risk assessment. Epidemiological data have the notable strength of being based on the “right” species, “real-world” exposures, and often the full range of susceptibility. Observational data often have inherent limitations, but strategies are available to handle them. The development of radiation risk models from epidemiological studies and the use of the models in quantitative risk assessment have set a risk assessment paradigm grounded in epidemiological research.

References

- [1] Graham, J.D. (1995). In *The Role of Epidemiology in Regulatory Risk Assessment*, J.D. Graham, ed, Elsevier Publishing, New York.
- [2] Samet, J.M., Schnatter, R. & Gibb, H. (1998). Epidemiology and risk assessment, *American Journal of Epidemiology* **148**, 929–936.
- [3] Graham, J.D., Paustenbach, D.J. & Butler, W.J. (1995). In *The Role of Epidemiology in Regulatory Risk Assessment*, J.D. Graham, ed, Elsevier Publishers, New York, pp. 1–14.
- [4] Hertz-Picciotto, I. (1995). Epidemiology and quantitative risk assessment: a bridge from science to policy, *American Journal of Public Health* **85**, 484–491.
- [5] Burke, T.A. (1995). In *The Role of Epidemiology in Regulatory Risk Assessment*, J.D. Graham, ed, Elsevier Publishers, New York, pp. 149–155.
- [6] Federal Focus (1996). *Principles for Evaluating Epidemiologic Data in Regulatory Risk Assessment*, 0-9654148-0-9.

- [7] Aucther, T.G. (1995). In *The Role of Epidemiology in Regulatory Risk Assessment*, J.D. Graham, ed, Elsevier Publishers, New York, pp. 143–148.
- [8] National Research Council (NRC), Committee on the Institutional Means for Assessment of Risks to Public Health (1983). *Risk Assessment in the Federal Government: Managing the Process*, National Academy Press, Washington, DC.
- [9] Goodman, S.N. & Samet, J.M. (2006). In *Cancer Epidemiology and Prevention*, D. Schottenfeld & J.F. Fraumeni, eds, Oxford University Press, New York, pp. 3–9.
- [10] Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**, 295–300.
- [11] US Department of Health Education and Welfare (DHEW) (1964). *Smoking and Health: Report of the Advisory Committee to the Surgeon General*, DHEW Publication No. [PHS] 1103, U.S. Government Printing Office, Washington, DC.
- [12] US Department of Health and Human Services (USDHHS) (2004). *The Health Effects of Active Smoking: A Report of the Surgeon General*, U.S. Government Printing Office, Washington, DC.
- [13] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, Lippincott Williams & Wilkins, Philadelphia.
- [14] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.
- [15] Lewis, E.B. (1957). Leukemia and ionizing radiation, *Science* **125**, 963–972.
- [16] Radiation Effects Research Foundation (RERF) (2007). http://www.rerf.or.jp/index_e.html.
- [17] National Research Council (NRC), Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation (2005). *Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2*, National Academy of Sciences, Washington, DC.
- [18] US Environmental Protection Agency (EPA) (2004). *Air Quality Criteria for Particulate Matter*, EPA/600/p-99/022aD and bD, National Center for Environmental Assessment, Research Triangle Park.
- [19] Samet, J.M., Zeger, S., Dominici, F., Currier, F., Coursac, I., Dockery, D., Schwartz, J. & Zanobetti, A. (2000). *The National Morbidity, Mortality, and Air Pollution Study (NMMAPS), Part II: Morbidity and Mortality from Air Pollution in the United States*, Health Effects Institute.
- [20] Samet, J.M., Zeger, S., Dominici, F., Dockery, D. & Schwartz, J. (2000). *The National Morbidity, Mortality, and Air Pollution Study (NMMAPS), Part I: Methods and Methodological Issues*, Health Effects Institute, Boston.
- [21] Dockery, D.W., Pope III, C.A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris Jr, B.G. & Speizer, F.E. (1993). An association between air pollution and mortality in six U.S. cities, *New England Journal of Medicine* **329**, 1753–1759.
- [22] Pope III, C.A., Thun, M.J., Namboodiri, M.M., Dockery, D.W., Evans, J.S., Speizer, F.E. & Heath Jr, C.W. (1995). Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults, *American Journal of Respiratory and Critical Care Medicine* **151**, 669–674.
- [23] National Research Council (NRC) & Commission on Life Sciences (2000). *Toxicological Effects of Methylmercury*, National Academy Press, Washington, DC.
- [24] Grandjean, P., Weihe, P., White, R.F., Debes, F., Araki, S., Yokoyama, K., Murata, K., Sorensen, N., Dahl, R. & Jorgensen, P.J. (1997). Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury, *Neurotoxicology and Teratology* **19**, 417–428.
- [25] Rice, D.C. (2004). The US EPA reference dose for methylmercury: sources of uncertainty, *Environmental Research* **95**, 406–413.
- [26] US Environmental Protection Agency (EPA) (1998). *Research Plan for Arsenic in Drinking Water*, EPA/600/R-98/042, Office of Research and Development, National Center For Environmental Assessment, Research Triangle Park.
- [27] National Research Council (NRC) (2001). *Arsenic in Drinking Water: 2001 Update*, National Academy Press, Washington, DC.
- [28] Needleman, H.L., Davidson, I., Sewell, E.M. & Shapiro, I.M. (1974). Subclinical lead exposure in Philadelphia schoolchildren. Identification by dentine lead analysis, *New England Journal of Medicine* **74**, 245–248.
- [29] Canfield, R.L., Henderson Jr, C.R., Cory-Slechta, D.A., Cox, C., Jusko, T.A. & Lanphear, B.P. (2003). Intellectual impairment in children with blood lead concentrations below 10 microg per deciliter, *New England Journal of Medicine* **348**, 1517–1526.
- [30] Levin, M.L., Goldstein, H. & Gerhardt, P.R. (1950). Cancer and tobacco smoking: a preliminary report, *Journal of the American Medical Association* **143**, 336–338.
- [31] Wynder, E.L. & Graham, E.A. (1950). Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma: a study of six hundred and eighty-four proved cases, *Journal of the American Medical Association* **143**, 329–336.
- [32] Doll, R. & Hill, A.B. (1950). Smoking and carcinoma of the lung, *British Medical Journal* **2**, 739–748.
- [33] Doll, R. & Hill, A.B. (1954). The mortality of doctors in relation to their smoking habits: a preliminary report, *British Medical Journal* **1**, 1451–1455.
- [34] Mokdad, A.H., Marks, J.S., Stroup, D.F. & Gerberding, J.L. (2004). Actual causes of death in the United States, 2000, *Journal of the American Medical Association* **291**, 1238–1245.
- [35] Murray, C.J. & Lopez, A.D. (1997). Global mortality, disability, and the contribution of risk factors: global burden of disease study, *Lancet* **349**, 1436–1442.
- [36] US Department of Health and Human Services (USDHHS) (1986). *The Health Consequences of Involuntary Smoking: A Report of the Surgeon General*, DHHS

- Publication No. (CDC) 87-8398, U.S. Government Printing Office, Washington, DC.
- [37] National Research Council (NRC) & Committee on Passive Smoking (1986). *Environmental Tobacco Smoke: Measuring Exposures and Assessing Health Effects*, National Academy Press, Washington, DC.
 - [38] US Environmental Protection Agency (EPA) (1992). *Respiratory Health Effects of Passive Smoking: Lung Cancer and Other Disorders*, EPA/600/006F, U.S. Government Printing Office, Washington, DC.
 - [39] International Agency for Research on Cancer (IARC) (2004). *Tobacco Smoke and Involuntary Smoking*, IARC monograph 83, Lyon.
 - [40] US Department of Health and Human Services (USDHHS) (2006). *The Health Effects of Involuntary Exposure to Tobacco Smoke*, Centers for Disease Control and Prevention (CDC), Rockville.
 - [41] Mozaffarian, D., Katan, M.B., Ascherio, A., Stampfer, M.J. & Willett, W.C. (2006). Trans fatty acids and cardiovascular disease, *New England Journal of Medicine* **354**(15), 1601–1613.
 - [42] Cohen, A.J., Anderson, H.R., Ostro, B., Pandey, K.D., Krzyzanowski, M., Kunzli, N., Gutschmidt, K., Pope III, C.A., Romieu, I., Samet, J.M., Smith, K.R. (2004). In *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors*, M. Ezzati, A.D. Lopez, A. Rodgers & C.J. Murray, eds, World Health Organization, Geneva, pp. 1353–1434.
 - [43] Armstrong, B.K., Saracci, R. & White, E. (1992). *Principles of Exposure Measurement in Epidemiology*, Oxford University Press, New York.
 - [44] Dosemeci, M., Washolder, S. & Lubin, J.H. (1990). Does nondifferential misclassification of exposure always bias a true effect toward the null value? *American Journal of Epidemiology* **132**, 746–748.
 - [45] Steenland, K. & Deddens, J.A. (2004). A practical guide to dose-response analyses and risk assessment in occupational epidemiology, *Epidemiology* **15**, 63–70.
 - [46] National Research Council (NRC), Committee on Health Risks of Exposure to Radon, Board on Radiation Effects Research and Commission on Life Sciences (1999). *Health Effects of Exposure to Radon (BEIR VI)*, National Academy Press, Washington, DC.
 - [47] Land, C.E. (1980). Estimating cancer risks from low doses of ionizing radiation, *Science* **80**, 1197–1203.
 - [48] Lubin, J.H. & Gail, M.H. (1990). On power and sample size for studying features of the relative odds of disease, *American Journal of Epidemiology* **90**, 552–566.
 - [49] US Environmental Protection Agency (EPA) (2005). *A Citizen's Guide to Radon*, 402-K02-006, US Government Printing Office Washington, DC.
 - [50] Jenkins, R.A., Moody, R.L., Higgins, C.E. & Moneyhun, J.H. (1991). Nicotine in environmental tobacco smoke: comparison of mobile personal and stationary area sampling, *Proceedings of the 1991 EPA/APCA Symposium on Measurement of Toxic and Related Air Pollutants*, Raleigh, pp. 437–442.
 - [51] National Research Council (NRC), Committee on Risk Characterization and Commission on Behavioral and Social Sciences and Education (1996). *Understanding Risk. Informing Decisions in a Democratic Society*, National Academy Press, Washington, DC.
 - [52] World Health Organization (2007). *Causes of Death-Estimates*, Geneva.
 - [53] Szklo, M. & Nieto, F.J. (2000). *Epidemiology: Beyond the Basics*, Aspen, Gaithersburg.
 - [54] Van den Brandt, P., Voorrips, L., Hertz-Picciotto, I., Shuker, D., Boeing, H., Speijers, G., Guittard, C., Kleiner, J., Knowles, M., Wolk, A., Goldbohm, A. (2002). The contribution of epidemiology, *Food and Chemical Toxicology* **40**, 387–424.
 - [55] Petitti, D.B. (1994). *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*, Oxford University Press, New York.
 - [56] Greenland, S. (1987). Quantitative methods in the review of epidemiologic literature, *Epidemiologic Reviews* **87**(9), 1–30.
 - [57] Dickersin, K. & Berlin, J.A. (1992). Meta-analysis: state-of-the-science, *Epidemiologic Reviews* **14**, 154–176.
 - [58] Hasselblad, V., Kotchmar, D.J. & Eddy, D.M. (1992). Synthesis of environmental evidence: nitrogen dioxide epidemiology studies, *Journal of the Air and Waste Management Association* **42**, 662–671.
 - [59] Boice, J.D. & Lubin, J.H. (1996). Lung cancer risks: comparing radiation with tobacco, *Radiation Research* **146**, 356–357.
 - [60] Lubin, J.H. & Boice, J.D., Jr. (1997). Lung cancer risk from residential radon: meta-analysis of eight epidemiologic studies, *Journal of the National Cancer Institute* **89**, 49–57.
 - [61] Cardis, E., Gilbert, E.S., Carpenter, L., Howe, G., Kato, I., Armstrong, B.K., Beral, V., Cowper, G., Douglas, A., Fix, J., Fry, S.A., Kaldor, J., Lave, C., Salmon, L., Smith, P.G., Voelz, G.L., Wiggs, L.D., (1995). Effects of low doses and low dose rates of external ionizing radiation: cancer mortality among nuclear industry workers in three countries, *Radiation Research* **142**, 117–132.
 - [62] Lubin, J.H., Boice Jr, J.D., Edling, C., Hornung, R.W., Howe, G., Kunz, E., Kusiak, R.A., Morrison, H.I., Radford, E.P., Samet, J.M., Tirmarche, M., Woodward, A., Xiang, Y.S., Pierce, D.A. (1994). *Radon and Lung Cancer Risk: A Joint Analysis of 11 Underground Miners Studies*, U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, Bethesda.
 - [63] Howe, G.R. & McLaughlin, J. (1996). Breast cancer mortality between 1950 and 1987 after exposure to fractionated moderate-dose-rate ionizing radiation in the Canadian fluoroscopy cohort study and a comparison with breast cancer mortality in the atomic bomb survivors study, *Radiation Research* **145**, 694–707.

- [64] Swaen, G.M. (2006). A framework for using epidemiological data for risk assessment, *Human and Experimental Toxicology* **25**, 147–155.

**Environmental Health Risk Assessment
Statistics for Environmental Toxicity**

Related Articles

JONATHAN M. SAMET AND BENJAMIN J.
APELBERG

Environmental Exposure Monitoring

Environmental Health Risk

Homeland Security and Transportation Risk

Transportation infrastructures (*see* **Managing Infrastructure Reliability, Safety, and Security**) make attractive targets of intentional harmful attacks because of their visibility, accessibility, and capacity to carry large numbers of commuters in a relatively confined space. Historical data on terrorist incidents worldwide show that approximately 60% of attacks have targeted the transportation sector [1]. Terrorism can be generally defined as “premeditated, politically motivated violence or deliberate threats of violence against noncombatant targets by subnational groups or clandestine agents, usually intended to influence an audience” [2].

The coordinated hijackings and deliberate crashes of airplanes into the World Trade Center in New York and the Pentagon in Washington, DC, on September 11, 2001, dramatically heightened both the perception and the reality of the threat terrorism poses to public transportation systems. This event, unprecedented in the annals of terrorism, focused national attention on aviation security in the United States (*see* **Near-Miss Management: A Participative Approach to Improving System Reliability**). However, maritime and surface transportation systems also continue to be vulnerable to attacks by terrorists who seek to attract publicity, inflict high numbers of civilian casualties, and cause political and economic disruption. The London subway bombings on July 7, 2005, provided more recent and prominent evidence of this continuing vulnerability.

Methodologies for Assessing and Managing Transportation Security Risks

The US government responded to the September 11, 2001, attacks by elevating terrorism to the top of the country’s national security and law enforcement agendas. Existing and newly created federal agencies now face the daunting challenge of determining how to allocate finite resources to manage risks while addressing threats and enhancing security across all key transportation modes: aviation, maritime, rail, pipelines, highways, trucking, and busing, and public mass transit.

The Transportation Security Administration, created in November 2001 as part of the Aviation and Transportation Security Act, has pledged to follow a systems-based risk-management approach that incorporates the three key elements of all risk assessment methodologies: threat, vulnerability (including criticality), and consequence (*see* Figure 1).

A threat assessment (*see* **Use of Decision Support Techniques for Information System Risk Management; Simulation in Risk Management; Reliability of Consumer Goods with “Fast Turn Around”**) identifies and evaluates potential threats on the basis of factors such as capabilities, intentions, and past activities. A vulnerability assessment recognizes weaknesses that may be exploited by identified threats and suggests countermeasures to address those weaknesses. A criticality assessment evaluates and ranks assets and functions in terms of specific criteria, such as their importance to public safety and the economy, as a basis for determining which structures or processes are relatively more important to protect from attack [3]. The characterization of risk is completed through analysis of the consequences associated with the partial or total loss of critical assets because of intentional harmful attacks.

Threat Assessment

Terrorism presents an “asymmetric” threat in which terrorists employ surprise and relatively low-cost weaponry to inflict or threaten catastrophic damage on large populations and property, causing fear, panic, and disruption. Assessing such threats is more difficult and often derives from assumptions about the capability and intent of terrorist organizations. These assumptions may include the value the threat places on certain assets for political or morale purposes, the threat’s ability to gain access to critical assets, and the disruptive or destructive capability of the weapons the threat may employ (including threatening statements, hoaxes, and real weapons of mass destruction) [4].

In March 2002, the US government established the Homeland Security Advisory System to provide a national framework for communicating the nature and degree of terrorist threats between government officials and citizens. The system uses color codes to convey whether there is a severe (red), high (orange), elevated (yellow), guarded (blue), or low (green)

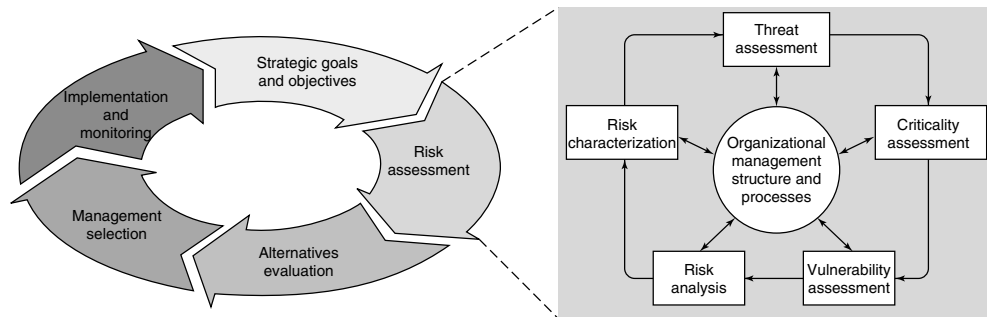


Figure 1 Risk-management cycle

risk of terrorist attacks. In assessing a threat, system managers consider a variety of questions:

- Is the threat credible?
- Is the threat corroborated?
- Is the threat specific or imminent?
- How grave is the threat?

With regard to specific transportation assets, public agencies responsible for assessing and managing security risks may apply related criteria to make an informed judgment [4]:

- **Existence** – Is there a group or an individual that is known to or potentially could be operating within the jurisdiction with the capability to create or use a particular weapon or tactic?
- **History** – Has the jurisdiction experienced any past terrorist activity?
- **Intent** – Are there credible advocacies or threats of force or violence, or acts, or preparations to act, evidencing the intent to create or use a weapon or tactic?
- **Capability** – Is there credible information that a specific group or individual possesses the needed training, skills, finances, and access to resources to develop or acquire a particular weapon or tactic?
- **Target** – Is there credible information indicative of preparations for specific terrorist operations against a critical asset?

Vulnerability Assessment

A vulnerability assessment involves two key steps: (a) the identification of critical transportation assets and (b) evaluation of the vulnerability of critical

assets to one or more presumed threats. Identification of critical assets typically begins by organizing a team of experienced staff most familiar with the transportation assets under assessment. For highway assets, for example, such a team might include operations and maintenance staff, design and construction engineers, traffic engineers, and field personnel. Transportation assets can be organized into four major categories: infrastructure, facilities, equipment, and personnel. By considering the mission of the transportation system and of the responsible government agency, the study team can develop a list of key assets that enable achievement of the mission. An assessment of highway transportation assets might yield a list similar to Table 1.

Judgment of the relative importance of assets proceeds on the basis of critical factors such as casualty risk, potential effects on government continuity and emergency response, the military importance of the asset, economic impact, and the availability of alternative resources to perform an asset's primary function.

At least three factors influence the vulnerability of a particular transportation asset: (a) visibility and attendance, (b) access, and (c) site-specific hazards. Vulnerability increases with greater awareness of the existence of an asset and the number of people typically present, as well as with greater availability of an asset to ingress and egress by a potential threat element. Vulnerability is further increased by the presence of materials on site that have biological, nuclear, incendiary, chemical, or explosive properties in quantities that would expend initial response capabilities if compromised [4].

Numerical scales have been developed to assess both the criticality and vulnerability of transportation

assets, but simpler ordinal scales are also used. For example, the Federal Transit Administration grades the accessibility of assets on a five-letter scale as depicted in Table 2.

The vulnerability of a transportation asset also depends on the specific nature of the threat. Attacks with explosives or incendiaries have characteristics that differ from attacks involving acts of force or the use of chemical, biological, or radiological agents. A recent report on security measures for ferry systems [5] identified 11 different security locations within ferry systems and considered the vulnerability of each area to different types of threats. For example, Table 3 presents hypothetical relative vulnerabilities of security areas to delivery of an improvised explosive device (IED) by various modes.

Consequence Assessment

The specific consequences of attacks on transportation assets depend on the nature of the attack and the impact of the loss of the asset. Consequences can include loss of life and property associated with the attack, as well as loss of infrastructure needed to support economic activity, military deployment, or the ability to respond effective to other emergencies [4].

Results from the criticality and vulnerability assessments assist risk analysts in focusing the consequence assessment on scenarios expected to pose the greatest risks to transportation security. The most severe consequences usually derive from attacks on the most critical and vulnerable assets – the so-called Quadrant I assets in the criticality and vulnerability matrix (see Figure 2).

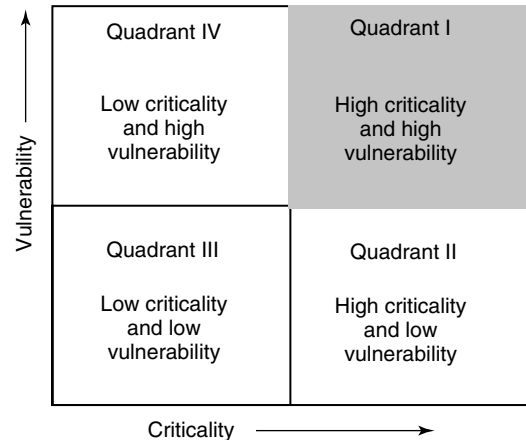


Figure 2 Criticality and vulnerability matrix

Table 1 Examples of critical highway transportation assets^(a)

Infrastructure	Facilities	Equipment	Personnel
Arterial roads	Headquarters buildings	Hazardous materials	Contractors
Interstate roads	Traffic operations centers	Roadway monitoring	Employees
Bridges	Regional complexes	Signal/control systems	Vendors
Overpasses	Maintenance stations	Messaging systems	Visitors
Barriers	Ports of entry	Vehicles	
Roads on dams	Chemical storage areas	Communication systems	
Tunnels	Fueling stations		

^(a) Reproduced from [4] with permission from Science Applications International Corporation. © 2002

Table 2 Attack vulnerability levels based on accessibility to threat element^(a)

Description	Level	Ease of access
Very easy	A	Very easy to access area or affect function undetected
Easy	B	Relatively easy to access area or function, no significant barriers to prevent
Difficult	C	Difficult to access area or function, various barriers in place
Very difficult	D	Very difficult to access area or function, barriers very difficult to overcome
Too much effort	E	Extremely difficult and cumbersome to access area or function, no history of incursion or attempted incursion

^(a) Reproduced from [4] with permission from Science Applications International Corporation. © 2002

Table 3 Hypothetical relative vulnerabilities of security areas to IEDs

Location	IED delivery mode					
	Person	Vehicle	Vessel	Artillery	Mine	Overhead
Beyond site boundary	M	M	N/A	L	N/A	**
Facility perimeter	M	H	N/A	L	N/A	**
Vehicle parking	H	H	N/A	M	N/A	**
Vehicle holding	M	H	N/A	L	N/A	**
Passenger waiting area	H	M	N/A	L	N/A	**
Terminal operations	M	L	N/A	H	N/A	**
Adjacent to ferry (shore)	H	L	N/A	H	N/A	**
Adjacent to ferry (water)	M	L	H	H	H	**
Onboard (unrestricted)	H	L*	N/A	L	N/A	**
Onboard (restricted)	M	H*	N/A	L	N/A	**
In transit	L	L	H	H	M	**

H: high; M: medium; L: low; and N/A: not applicable for this security area

*Assumes that onboard cargo area, including vehicle storage area, is restricted

**Assumes similar vulnerability among security areas without specification of a particular mode

Risk is informally defined as probability times consequence. Thus, a complete risk analysis weighs the estimated frequency of a threat, as well as its estimated consequences. Such estimates can be made using either qualitative or quantitative methods. In either case, the goal is to identify the attack scenarios posing the greatest transportation security risks. Subsequently, the key task for risk managers is to determine how overall security risk should be reduced through the deployment of countermeasures that either reduce the chance of an attack or mitigate its consequences.

Qualitative versus Quantitative Risk Analyses

Because of the difficulties in quantifying key components of threat, vulnerability, and consequence assessments, analyses of transportation security risk typically employ qualitative methods in making judgments about the relative magnitudes of various risk scenarios. Although it is often impractical to apply quantitative risk analysis to a large number of risk scenarios [1], such methods may be more readily used after an initial screening to identify the scenarios of greatest concern.

Some elements of a transportation security risk analysis are relatively amenable to quantitative treatment. Databases on terrorist attacks are maintained by the US government, nonprofit research organizations such as the Mineta Transportation Institute and the RAND Corporation, and private companies [2, 6].

These databases support research on the targets of terrorist attacks and the weapons and tactics used by the perpetrators. For example, Figure 3 summarizes the distribution of targets in attacks on public surface transportation systems during the 80-year period from 1920 to 2000. Buses, subways, and trains are the most popular targets, collectively involved in more than half of the attacks in this period. Similarly, Figure 4 shows that bombings predominate as the tactic of choice in attacks on public transportation systems, accounting for 60% of all such incidents from 1920 to 2000 [2].

Historical data on events involving the use of various types of weapons and tactics can also be used in quantitative assessments of the consequences of future attacks. Table 4 lists the typical number or range of casualties incurred in attacks involving threats available or potentially available to terrorists. Ongoing concerns about the risks of terrorist attacks involving chemical, biological, radiological, or nuclear (CBRN) agents stem from the great destructive force of these weapons.

Transportation Security Risk Management

Risk management involves two key steps: (a) determining which risks identified in the risk assessment process require mitigation and (b) developing and implementing plans or actions that are required to ensure that the selected risks are reduced to more acceptable levels. Protective security measures are

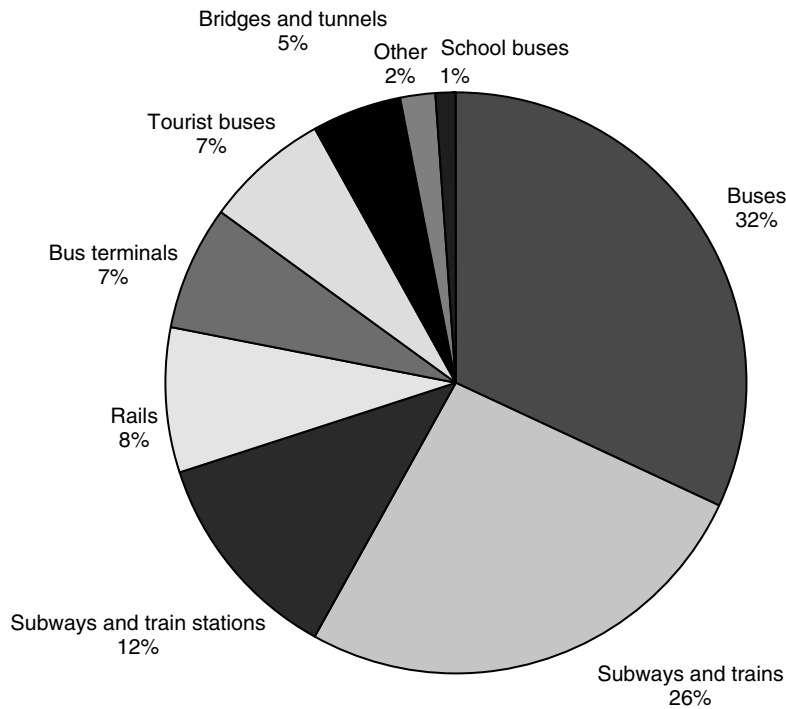


Figure 3 Targets of attacks on public surface transportation systems (1920–2000)

actions intended to reduce vulnerability, deny an adversary opportunity, or increase response capacity during a period of heightened alert [7].

Countermeasures to protect critical assets such as transportation infrastructure commonly involve site-work, building/structure, detection, and procedural elements [4]:

- *Site-work* elements address the area surrounding a facility or an asset and may include perimeter barriers, landforms, and standoff distances.
- *Building and structure* elements are directly associated with facilities and structures and may include walls, doors, windows, and roofs.
- *Detection* elements include closed-circuit television, motion detectors, alarms, and weapon detectors, as well as any guards needed to operate this equipment or to perform similar functions.
- *Procedural* elements refer to the protective measures required by regulations, policies, or plans that provide the basis for developing the other elements.

Security countermeasures are intended to provide target systems with additional capabilities in terms of deterrence, detection, and defense [4]. Deterrence countermeasures are effective when a potential aggressor is dissuaded from attacking an asset because of a perceived increase in the risk that the attack will fail or that the aggressor will be captured. The success of deterrence will vary depending on the asset's attractiveness, as well as the sophistication and objectives of the aggressor. An effective detection measure provides added capabilities to sense an act of aggression, assess the validity of the detection, and communicate the appropriate information to a response force. Defensive measures protect an asset by preventing or delaying an aggressor's movement toward the asset or by shielding the asset from the effects of tools, weapons, and explosives.

For example, in compiling a list of possible countermeasures for critical bridges, one state department of transportation identified the following items [4]:

- increased patrol by law enforcement;
- increased patrol by coast guard;

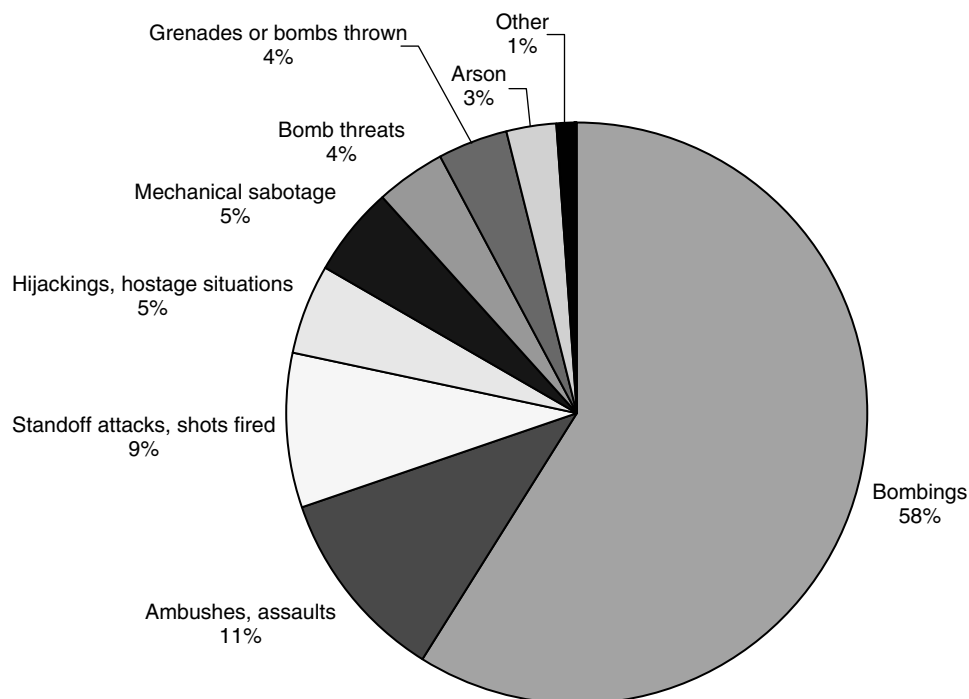


Figure 4 Tactics used against public transportation systems (1920–2000)

Table 4 Specific threat types and historical casualty rates^(a)

Specific threat type	Historical casualties per incident (number of people affected)
Nuclear	50 000
Chemical	5000
Biological	1500–5000
Radiological	1500–5000
Explosives/incendiary	50–3000
Fire arms/armed assault	10–100
Hijack/hostage	10–100
Cyber/information security	Unknown

^(a) Reproduced from [7]. Federal Transit Administration, 2006

- rapid removal of abandoned vehicles;
- barriers around bridge piers;
- construction of barriers around cable anchors of suspension bridges;
- higher level of identification for personnel working on or around affected bridges;
- security cameras, monitors, and related software to monitor sensitive areas;

- removal of vegetation to provide clear lines of sight.

Making rational decisions among options for managing transportation security risks requires estimation of the capital, operating, and maintenance costs of candidate countermeasures, as well as a comparative evaluation of their effectiveness in reducing either the potential for or consequences of attacks on assets given specific threats and vulnerabilities [4].

One recently published cost–benefit analysis (*see Environmental Performance Index; Environmental Risk Regulation; Global Warming; Operational Risk Modeling*) [8] compared several alternative device configurations for aviation-checked baggage security screening:

- explosive detection systems;
- explosive trace detection;
- dual energy X ray;
- backscatter machines.

The study considered single-device configurations, as well as two-device scenarios in which a second

device is used to screen all bags signaling alarms when passed through the first device. Configurations were evaluated with respect to the expected direct cost per checked bag and the expected number of successful threats per one billion checked bags. Results indicated that, for the current security environment, single-device use of backscatter machines is superior to other configurations in terms of expected cost and number of successful threats.

Such cost–benefit analyses can assist risk managers in allocating resources most effectively to enhance transportation security.

Current Perspective on Transportation Security

National transportation systems in developed countries, such as the United States, Japan, and member nations of the European Union, are vast interconnected networks of diverse modes. Key modes include aviation, maritime traffic, and surface transportation (e.g., highways, rail, and transit). For example, the US transportation system moves over 30 million tons of freight and supports approximately 1.1 billion passenger trips each day [3]. Because of its diversity and size, a national transportation system is vital to a country's economy and national security.

Transport vehicles and facilities have frequently been targets of terrorist attacks, hijackings, and sabotage around the world. Transportation systems possess characteristics that make them especially vulnerable to terrorism: the capacity to carry millions of people daily in enclosed spaces; elements with symbolic, as well as strategic, value; both military and key civilian functions; and an open and distributed infrastructure that complicates efforts to deploy effective security measures [9].

Transportation assets are not only targets of terrorist attacks, but can serve as a weapon or delivery mechanism, as demonstrated dramatically in the attacks in the United States on the World Trade Center and the Pentagon. Regardless of the number of casualties involved, attacks that disrupt public movements can seriously impact a region's economy and can erode the public's faith in the government's ability to provide basic protections to its citizens [2].

The specter of international terrorists with more ambitious goals, as well as the resources to plan and launch increasingly destructive attacks, has established a new context for transportation security and

planning. Governments of countries that perceive an elevated level of threat are investing considerable resources to implement comprehensive programs to enhance security in all key modes of transportation. These efforts include formal risk assessment and management efforts. Recent initiatives in the United States serve as an illustrative case study.

Aviation

Airplanes were the weapons of choice in the devastating terrorist attack of September 11, 2001. Consequently, recent efforts to enhance transportation security – particularly, such efforts in the United States – have focused on the aviation industry. A new federal agency, the Transportation Security Administration, was created to provide federal control of passenger and baggage screening operations. All checked baggage is now screened using explosive detection systems. Access restrictions to planes and airport facilities have been implemented by adding security personnel, access controls, and screening measures. Expanded security operations include more random searches throughout airports, verification of employee identities, and more comprehensive background checks on passengers and employees [9]. Air marshals provide in-air security on high-risk flight routes, and measures such as cockpit doors with locking mechanisms have been introduced to improve protection of the flight officers.

Compared with other modes of transportation, air travel possesses some characteristics that facilitate implementation of security measures. Airplanes are housed in fairly closed and reasonably controlled locations. Also, airport terminal access to airplanes is controlled by relatively few entry points [2]. Recent initiatives have significantly improved aviation security and lowered the risk of hijackings or sabotage. However, concerns remain about vulnerabilities in the handling of air cargo and in procedures at smaller airports serving general aviation. Research is ongoing to develop new aviation security technologies, such as passenger profiling systems, bombproof cargo containers, and better screening tools for explosives and other materials [9].

Maritime Traffic

Shipping and ports are vital elements in the economic and military activities of most countries in

the world. The extensive size, open accessibility, and metropolitan location of most ports ensure a free flow of trade but present great challenges to effective monitoring and control of traffic through the ports. Large ships, such as oil tankers, warships, cargo ships, or cruise liners, are vulnerable because of their strategic value, visibility, slow speed, and limited maneuverability. Critical infrastructures located along the coast or on inland waterways – including bridges, power plants, and oil refineries – have few protections against approaching ships carrying explosives or other harmful materials [9].

Because of the urgent need to address aviation vulnerabilities after the September 11, 2001, attacks, the United States has placed secondary emphasis on strengthening maritime security. However, initial risk assessments were conducted for all ports to identify critical infrastructure and key assets, assess vulnerabilities, develop mitigation strategies, and highlight best practices. Protection has been improved by deploying fences, surveillance cameras, and additional personnel to guard facilities, as well as by limiting vehicular traffic around seaports. The coast guard has increased patrols of waterways, coastline, ports, and coastal facilities. Intelligence agencies have created large databases to track cargo, ships, and seamen in a search for anomalies that could be indicators of terrorist activities [9].

Perhaps the most serious terrorist threat to maritime security involves the smuggling of explosives or CBRN weapons into a country using cargo containers. To address this threat, the American government has launched a container security initiative with international cooperation. The initiative strategy is to use automated information to identify and target high-risk containers, detection technology to quickly screen high-risk containers, and smarter, tamper-proof containers to reduce susceptibility to misuse [9].

Surface Transportation

Surface transportation comprises a diverse array of long-distance passenger and freight vehicles and facilities, mass transit vehicles and facilities, and related infrastructures. Elements include trains, trams, subways, buses, trucks, railway lines, highways, stations, depots, control facilities, bridges, and tunnels. In virtually any country of the world, this complex network of systems is absolutely essential to support daily life and economic activity [9].

Terrorists have a history of attacking such surface transportation assets as cars and buses, typically with explosives or firearms. Indeed, about one-third of all terrorist attacks worldwide target surface transportation systems [9]. However, recent terrorist activities reflect a shift from isolated bombings, hijackings, and hostage-taking toward inflicting higher numbers of civilian casualties, more extensive property damage, and increasingly devastating effects on economies [10].

From this perspective, surface transportation assets generally have low attractiveness to terrorists, relative to other targets, because the potential for casualties is fairly modest. Transit and rail systems are regarded by law enforcement agencies to be among the most likely targets. The principal threats against highway physical assets are explosive attacks on such key links as bridges, interchanges, and tunnels. The most vulnerable highway facilities are those that play important regional and strategic roles, are difficult to replace, and would cause maximal disruption in case of attack [10].

In the United States, efforts to enhance surface transportation security have generally involved improvements to the physical security of the most critical assets and the control of access to these assets. Perimeter security measures include additional fences and barriers, surveillance cameras, intrusion detection systems, and better lighting. Many organizations have employed additional security staff and launched public awareness campaigns to deter and prevent terrorist attacks [9].

Because of the need for openness and accessibility at thousands of entry points, the wide geographic distribution of assets, and the static nature of some routes, complete protection of surface transportation infrastructure is simply not feasible. Better physical security and public vigilance, as well as an increased security presence, will not prevent all possible attacks. Effective security strategies must also seek to minimize loss of life and disruption through planning, coordination, and training involving transportation operators, law enforcement authorities, and emergency response organizations [9].

Conclusions

Any particular terrorist attack is essentially a low-probability, high-consequence event (*see Near-Miss Management: A Participative Approach to*

Improving System Reliability). Therefore, strategies to mitigate risks to transportation assets should differentiate among priorities largely on the basis of target criticality [10]. Both the scale and nature of transportation systems necessitate that a reasonable degree of risk must be accepted, even for critical assets, because complete mitigation is not feasible. No security system can stop determined terrorists from deploying explosive, biological, or chemical weapons in public places [2]. However, effective countermeasures increase the difficulty of executing a terrorist attack, as well as the chances of detecting and identifying the perpetrators, minimize casualties and disruptions, reduce panic, and reassure an alarmed citizenry. Formal methodologies for assessing and managing risks to transportation security provide a valuable conceptual structure and practical tools for allocating resources in cost-effective ways to improve public safety.

References

- [1] Leung, M., Lambert, J.H. & Mosenthal, A. (2004). A risk-based approach to setting priorities in protecting bridges against terrorist attacks, *Risk Analysis* **24**, 963–984.
- [2] Jenkins, B.M. & Gersten, L.N. (2001). *Protecting Public Surface Transportation against Terrorism and Serious Crime: Continuing Research on Best Security Practices*, Mineta Transportation Institute, San Jose.
- [3] Government Accountability Office (2005). *Transportation Security: Systematic Planning Needed to Optimize Resources*, GAO-05-357T, Washington, DC.
- [4] Science Applications International Corporation (2002). *A Guide to Highway Vulnerability Assessment for Critical Asset Identification and Protection, Prepared for American Association of State Highway and Transportation Officials*, Vienna.
- [5] Science Applications International Corporation (2006). *Public Transportation Security: Security Measures for Ferry Systems*, Transportation Research Board, Washington, DC, Vol. 11.
- [6] Bogen, K.T. & Jones, E.D. (2006). Risks of mortality and morbidity from worldwide terrorism: 1968–2004, *Risk Analysis* **26**, 45–59.
- [7] Battelle (2006). *Transit Agency Security and Emergency Management Protective Measures*, Federal Transit Administration.
- [8] Jacobson, S.H., Karnani, T., Kobza, J.E. & Ritchie, L. (2006). A cost-benefit analysis of alternative device configurations for aviation-checked baggage security screening, *Risk Analysis* **26**, 297–310.
- [9] Institute for Security Technology Studies (2003). *On the Road to Transportation Security*, Dartmouth College, Hanover.
- [10] Ham, D.B. & Lockwood, S. (2002). *National Needs Assessment for Ensuring Transportation Infrastructure Security, Prepared for American Association of State Highway and Transportation Officials*, Herndon.

DUANE L. STEFFEY

Hormesis

A central biological concept is the dose–response relationship (*see* **Dose–Response Analysis**) upon which the fields of toxicology and risk assessment are built. Assumptions concerning the nature of the dose–response affect the questions that scientists ask, how studies are designed, including the number and the spacing of doses, risk assessment procedures that are employed by agencies such as EPA and FDA, as well as cost–benefit assessments (*see* **Cost–Effectiveness Analysis**) that often guide the plethora of decisions upon which health systems and environmental regulations are based. In short, society is enormously dependent on the qualitative and quantitative accuracy of dose–response models, particularly when the likely exposures are outside (i.e., usually lower) the range at which the predictive biological models were tested. This is certainly the case for the vast majority of environmental agents.

Historical Foundations of the Dose–Response

Society has looked to the fields of toxicology and risk assessment for guidance on what is the nature of the dose–response, especially in the low-dose zone. For nearly eight decades, the international scientific community has offered a consistent and unwavering judgment that the most fundamental dose–response is that of the threshold model, with below-threshold responses being assumed to be indistinguishable from unexposed control-group values. This basic conclusion of the scientific community has led to the establishment of the threshold dose–response model (*see* **Threshold Models**) as the conceptual centerpiece for hazard assessment testing and bioassay designs, for uncertainty factor application in assessing risks from chemicals and drugs and for how to design preclinical investigations and clinical trials (*see* **Randomized Controlled Trials**) for the pharmaceutical industry. As discussed further below, the basis of this “firmly believed” conclusion has recently been undermined by several studies that have surprisingly provided the first ever large scale direct testing of the validity of the threshold dose–response model. The threshold dose–response model poorly predicted responses in

the low-dose zone (i.e., below threshold) for chemical and pharmaceutical agents [1–3].

Despite the current erosion of its credibility to accurately predict responses below the toxicological threshold, the threshold dose–response model was useful in the early part of the twentieth century in predicting toxicity at high doses and such dose–responses often seemed to display thresholds below which no toxicity occurred [4]. This threshold dose–response relationship had enormous practical value, since below-threshold exposures suggested that toxicity would not be a concern. Guided by the threshold dose–response model, the goal of governmental hazard assessment activities became one of defining the toxic dose–response and estimating the highest safe dose or the toxic threshold. On the basis of this frame of reference, toxicology quickly became a high dose–few dose discipline. The threshold model immediately became valuable in assessing toxic doses in the occupational domain and useful in establishing exposure standards; it was also useful in estimating doses that would eliminate pests, such as disease-transmitting insects. Regulatory agencies, therefore, focused on reliably estimating toxicological thresholds since below that dose, safety was assumed.

During the middle and latter decades of the twentieth century, agencies such as the FDA and EPA veered away from the threshold model when addressing low-dose cancer risks; such agencies decided to adopt a linear at low-dose modeling strategy, mostly to assuage a strong societal preoccupation with fear of cancer rather than a belief, or a data-based conclusion, that carcinogens did not act via a threshold scheme as well. This decision to accept linear at low-dose modeling for cancer risk assessment has been problematic ever since. The most important limitation of this modeling strategy is that it simply cannot be practically tested, that is, validated, at risks of relevance to society (i.e., less than one person adversely affected in a population of 1000). Since this model cannot be practically tested at even moderately low risks to society, its predictions were required to be either believed or at least accepted as accurate. However, since the implementation of these beliefs within society was enormously costly, the adoption of the linear-no-threshold (LNT) model for carcinogens has been a source of great controversy and public policy assessment.

Substantial debate over the past 30 years in the toxicology and risk assessment communities has

therefore focused on whether the threshold or the LNT model was most likely correct. While debate has continued, there has been no practical way to resolve this toxicological question since epidemiological methods are too insensitive, while toxicological methods would require enormous resources. For example, in the famous FDA megamouse study of the late 1970s even 24 000 mice were too few to answer the question since risks could only be resolved to the 1/100 risk level, thereby yielding the name *ED01 study* [5].

While the toxicological community thought the principal dose–response debate was over threshold *versus* LNT models, this proved to be the wrong question. Over the past decade, sufficient data have emerged to suggest that the most fundamental model in the biological and biomedical sciences is neither the threshold or LNT model, but rather the hormesis dose–response model, which is characterized by a low-dose stimulation and a high-dose inhibition as seen in either inverted U- or

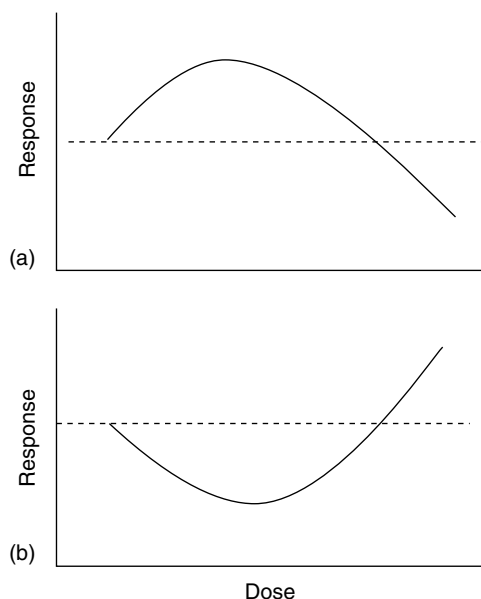


Figure 1 (a) The most common form of the hormetic dose–response curve depicting low-dose stimulatory and high-dose inhibitory response, the β - or inverted U-shaped curve. Endpoints displaying this curve include growth, fecundity, and longevity. (b) The hormetic dose–response curve depicting low-dose reduction and high-dose enhancement of adverse effects. Endpoints displaying this curve include carcinogenesis, mutagenesis, and disease incidence

J-shaped dose–responses (Figure 1). The documentation to support the conclusion of the primacy of the hormetic model is widely published and substantial [6, 7]. The hormetic dose–response has been demonstrated to be highly generalizable across biological models (e.g., plant, microbial, invertebrate, vertebrate), endpoints, and chemical classes. Using *a priori* entry and evaluative criteria, the hormetic model has also been found to be far more common than the revered threshold model in multiple head-to-head comparisons [1–3] involving data from toxicological and pharmaceutical studies. Using rigorous *a priori* entry and evaluative criteria, these investigations indicated the hormetic dose–response occurred unequivocally in approximately 40% of the dose–responses. When directly compared with the threshold model, hormesis was found to be 2.5 times more common than dose–responses satisfying the threshold model.

Historical Blunders

Given these substantial findings, the question is how could the entire fields of pharmacology, toxicology, other life science disciplines, and risk assessment make an error on the most fundamental concept upon which their disciplines were built, that is, the dose–response. The critical mistake, which involved concluding that the threshold model was the most fundamental dose–response model, grew out of a long, hostile and economically driven dispute between what is now called *traditional medicine* and the medical practice of homeopathy. At the center of the dose–response controversy was Hugo Schulz, a physician and academic pharmacologist at the University of Grieswald in northern Germany, who reported that low doses of numerous disinfectants reliably caused a low-dose stimulation and a high-dose inhibition of yeast metabolism, that is, the hormetic dose–response [8, 9]. Schulz quickly came to believe that these findings provided the explanatory principle of homeopathy, a belief based on earlier research with homeopathic medicinal preparations that seemed to effectively cure patients with gastroenteritis, yet did so without killing the causative bacteria [10]. Schulz hypothesized that this homeopathic medicine, and others in homeopaths’ arsenal of treatments, may enhance the adaptive capacity of ill individuals, thereby affecting disease resistance.

Data demonstrating low-dose stimulatory responses within yeasts led him to believe this hypothesis was validated. While Schulz became widely known within the homeopathic field, he and his work became the object of considerable attacks and ridicule by leaders in the field of traditional medicine, especially those with high visibility in the field of experimental pharmacology [11–13]. The criticism of Schulz was more political than scientific as Clark tried to associate Schulz with extremist (i.e., high dilutionist) positions within homeopathy. The work of Schulz was disparaged in the leading pharmacological textbooks and the nascent hormetic dose–response concept (i.e., then called the *Arndt–Schulz law*, later to be named *hormesis* in 1943 by forestry researchers at the University of Idaho based on the Greek word *to excite*) [14] became largely excluded from the mainstream of biomedical research [15]. In addition to trying to prevent the theories of Schulz from gaining acceptance, Clark and colleagues promoted the use of the threshold model and its widespread application, including numerous publications by Bliss [16, 17] that were targeted to different biological subdisciplines (e.g., microbiology, entomology, food sciences, etc.). In addition, colleagues of Clark also developed the probit model for dose–response modeling and constrained low-dose modeling to asymptotically approach the origin, denying it the flexibility to model J-shaped dose–responses [18, 19]. This had the effect of negating the possible biological reality of hormetic effects, suggesting that such responses were background variation. These constraining biostatistical approaches became integrated into toxicology, adopted by regulatory agencies and are still used to guide low-dose risk assessment modeling for carcinogens. This perspective was easily viewed as correct since the hormetic dose–response was hard to prove because of the modest nature of the low-dose stimulatory response (see discussion below). This being the case, it was easier to simply consider such responses as background variability and not to test such below-threshold response hypotheses in a rigorous fashion, since it would require more doses, larger number of subjects for improved statistical power calculations along with the greater need for study replication. For a number of reasons, therefore, the hormetic model was marginalized while the threshold dose–response model became relied upon and promoted.

Resurgence of Hormesis

Despite numerous obstacles, the hormesis concept has witnessed a major resurgence in interest during the past decade. This “rediscovery” of the hormesis concept was ironically stimulated by regulatory agencies such as the US EPA that had implemented highly conservative linearity at low-dose modeling into risk assessment practices for carcinogens. This practice resulted in such prohibitively expensive exposure standards and clean-up activities that alternative dose–response models were sought to challenge these regulatory activities by major private-sector entities. It was within that context that hormesis generated much interest; it was suggested that carcinogens would display not only a dose–response threshold thereby challenging the linearity concept, but that doses below the threshold may reduce risks below background (i.e., display hormesis). In fact, substantial studies dealing with chemical carcinogens [20] and radiation [21] have reported J-shaped hormetic dose–responses for cancer endpoints. For example, the ED01 study, the largest cancer bioassay ever conducted with rodents, indicated a clear J-shaped dose–response for the genotoxic carcinogen, 2-acetoaminofluene (2-AAF) induced bladder cancer [5] (Figure 2). Very reliable J-shaped dose–responses have also been reported for several liver carcinogens, such as DDT (Figure 3) [22] along with considerable mechanistic follow-up research.

While many toxicologists now acknowledge the existence and reproducibility of hormetic dose–responses, there is the demand to better understand the underlying mechanisms to account for hormetic dose–responses. In fact, numerous cases of specific mechanisms have been clarified, and reliably account for the biphasic nature of the dose–response in several dozen receptor-based systems (Table 1) affecting a broad range of tissues and endpoints from numerous animal models.

Despite the broad diversity of hormetic-biphasic dose–responses along with the uniqueness of the tissue specific receptor-based mechanisms, the quantitative features of this vast array of dose–response relationships are remarkably similar with respect to the amplitude and width of the stimulatory response (Figure 4). The magnitude of the hormetic dose–response is characteristically modest, with the maximum stimulation usually only 30–60% greater than

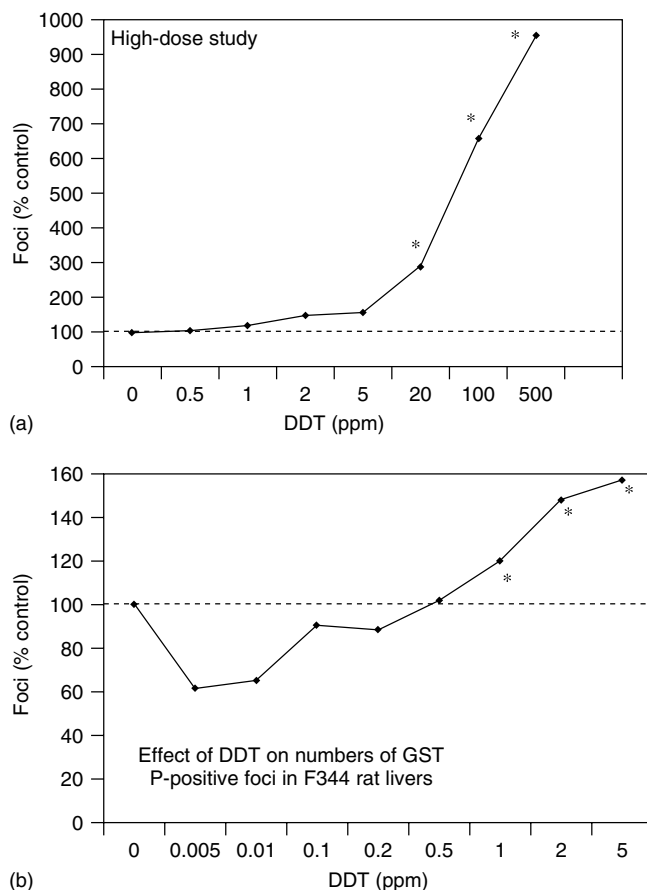


Figure 2 Effect of DDT on number of GST P-positive foci in F344 rat livers in two bioassays assessing different but slightly overlapping doses of carcinogen. Note: as the dose decreases the J-shaped dose-response becomes evident. Also note difference in scale between the two graphs [Reproduced from [23]. © John Wiley & Sons, Inc., 2002.]

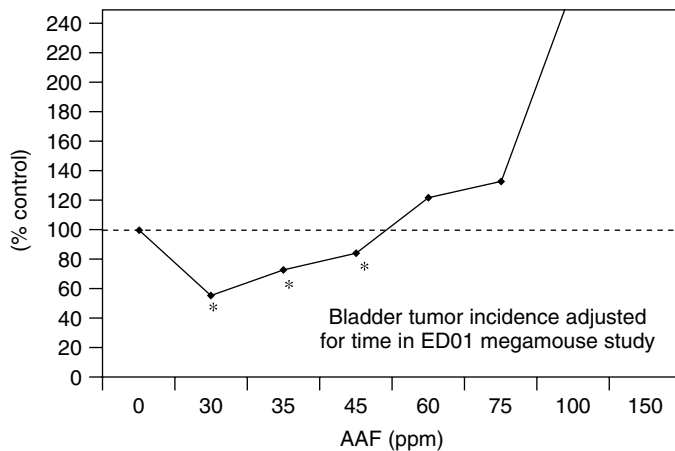


Figure 3 Bladder tumor incidence adjusted for time in ED01 megamouse study [Reproduced from [5]. © Academic Press, 1981.]

Table 1 A partial listing of receptor systems displaying biphasic dose–response relationships

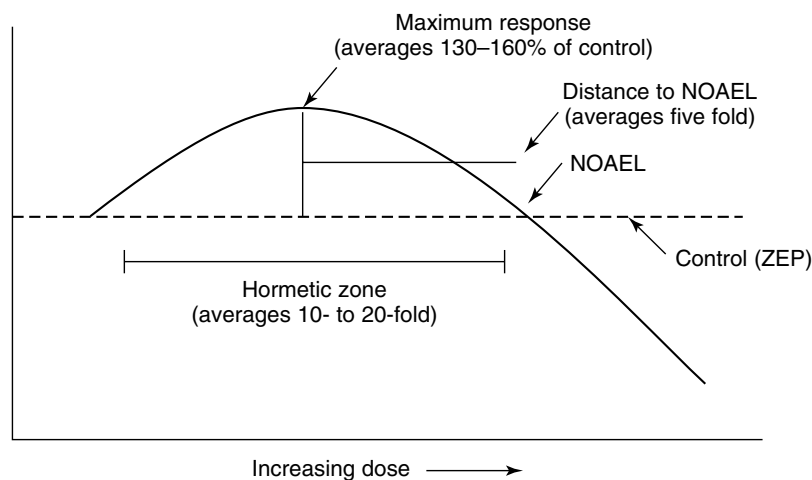
Adenosine	Neuropeptides
Adrenoceptor	Nitric oxide
Bradykinin	NMDA
CCK	Opioid
Corticosterone	Platelet-derived growth factor
Dopamine	Prolactin
Endothelin	Prostaglandin
Epidermal growth factor	Somatostatin
Estrogen	Spermine
5-HT (serotonin)	Testosterone
Human chorionic gonadotrophin	Transforming growth factor β
Muscarinic	Tumor necrosis factor α

control values. This is the case regardless of toxic potency as well as with respect to normal and high-risk segments of the population [24]. That is, that individuals displaying greater susceptibility to toxic agents display the same quantitative features of the dose–response as the more resistant members of the population. The width of the stimulatory response is typically within a range of approximately 1/5–1/20th of the toxic threshold as estimated by the no observed adverse effect level (NOAEL) or the zero equivalent point (ZEP). In a small minority (i.e., 2%) of the hormetic dose–responses, the width of the stimulatory responses has been reported to exceed a factor of 1000 [2]. The reasons for

this wider stimulatory zone is unknown, but may be related, at least in part, to population sample heterogeneity.

Hormesis in Perspective

Of particular importance is that the hormetic dose–response is a basic component to the traditional toxicological dose–response, with the hormetic response being immediately below and contiguous with the toxic threshold. This is of considerable value since it permits researchers and risk assessors to place the hormetic response within a dosage zone that can be tested and either validated or disproven, a feature that linearity at low-dose modeling cannot offer; the hormesis response can also be placed within a risk assessment framework that can be integrated into biostatistical modeling as well as fit into uncertainty factor evaluation for application in noncarcinogen risk assessment. This generalized dose–response feature of hormesis, independent of endpoint, has important implications as a feature that can facilitate the harmonization of carcinogen and noncarcinogen risk assessment methods, a long-standing issue for current risk assessment practices. Consequently, it is expected that the hormetic dose–response will have a progressively more influential role in toxicology and risk assessment. This is especially the case, as low-dose concerns become even more significant to society.

**Figure 4** Dose–response curve depicting the quantitative features of hormesis

References

- [1] Calabrese, E.J. & Baldwin, L.A. (2001). The frequency of U-shaped dose-responses in the toxicological literature, *Toxicological Sciences* **62**, 330–338.
- [2] Calabrese, E.J. & Baldwin, L.A. (2003). The hormetic dose response model is more common than the threshold model in toxicology, *Toxicological Sciences* **71**, 246–250.
- [3] Calabrese, E.J., Staudenmayer, J.W., Stanek, E.J. & Hoffmann, G.R. (2006). Hormesis outperforms threshold model in NCI anti-tumor drug screening data, *Toxicological Sciences* **94**, 368–378.
- [4] Calabrese, E.J. (2007). Threshold-dose-response model – RIP: 1911 to 2006, *BioEssays* **29**, 686–688.
- [5] Society of Toxicology (SOT) (1981). Re-examination of the ED01 study. Adjusting for time on study, *Fundamental and Applied Toxicology* **1**, 67–80.
- [6] Calabrese, E.J. (2005a). Paradigm lost, paradigm found: the re-emergence of hormesis as a fundamental dose response model in the toxicological sciences, *Environmental Pollution* **138**, 378–411.
- [7] Calabrese, E.J. & Blain, R. (2005). The occurrence of hormetic dose responses in the toxicological literature, the hormesis database: an overview, *Toxicology and Applied Pharmacology* **202**, 289–301.
- [8] Schulz, H. (1887). Zur Lehre von der Arzneiwirkung, *Virchows Archiv fur Pathologische Anatomie und Physiologie fur Klinische Medizin* **108**, 423–445.
- [9] Schulz, H. (1888). Über Hefegifte, *Pflugers Archiv fur die Gesamte Physiologie des Menschen und der Tiere* **42**, 517–541.
- [10] Schulz, H. & Crump, T. (2003). NIH Library Translation (NIH-98-134). Contemporary medicine as presented by its practitioners themselves, Leipzig, 1923:217–250, *Nonlinearity in Biology, Toxicology, and Medicine* **1**, 295–318.
- [11] Clark, A.J. (1937). *Handbook of Experimental Pharmacology*, Verlag Von Julius Springer, Berlin.
- [12] Clark, A.J. (1933). *Mode of Action of Drugs on Cells*, Arnold, London.
- [13] Clark, A.J. (1927). The historical aspects of quackery, Part 2, *British Medical Journal* **July–Dec**, 960–960.
- [14] Southam, C.M. & Erhlich, J. (1943). Effects of extracts of western red-cedar heartwood on certain wood-decaying fungi in culture, *Phytopathology* **33**, 517–524.
- [15] Calabrese, E.J. (2005b). Historical blunders: How toxicology got the dose-response relationship half right, *Cellular and Molecular Biology* **51**, 643–654.
- [16] Bliss, C.I. (1940). The relation between exposure time, concentration and toxicity in experiments on insecticides, *Annals of the Entomological Society of America* **33**, 721–766.
- [17] Bliss, C.I. (1956). The calculation of microbial assays, *Bacteriological Reviews* **20**, 243–258.
- [18] Bliss, C.I. (1935a). Estimating the dosage-mortality curve, *Journal of Economic Entomology* **25**, 646–647.
- [19] Bliss, C.I. (1935b). The comparison of dosage-mortality data, *Annals of Applied Biology* **22**, 307–333.
- [20] Calabrese, E.J. & Baldwin, L.A. (1998). Can the concept of hormesis be generalized to carcinogenesis? *Regulatory Toxicology and Pharmacology* **28**, 230–241.
- [21] Calabrese, E.J. & Baldwin, L.A. (2002a). Radiation hormesis and cancer, *Human and Ecological Risk Assessment* **8**, 327–353.
- [22] Kinoshita, A., Wanibuchi, H., Wei, M. & Fukushima, S. (2006). Hormesis in carcinogenicity of non-genotoxic carcinogens, *Toxicologic Pathology* **19**, 111–122.
- [23] Sukata, T., Uwagawa, S., Ozaki, K., Ogawa, M., Nishikawa, T., Iwai, S., Kinoshita, A., Wanibuchi, H., Imaoka, S., Funae, Y., Okuno, Y. & Fukushima, S. (2002). Detailed low-dose study of 1,1-B IS(*p*-chlorophenyl)-2,2,2-trichloroethane carcinogenesis suggests the possibility of a hormetic effect, *International Journal of Cancer* **99**, 112–118.
- [24] Calabrese, E.J. & Baldwin, L.A. (2002b). Hormesis and high-risk groups, *Regulatory Toxicology and Pharmacology* **35**, 414–428.

Related Articles

Detection Limits

Low-Dose Extrapolation

Potency Estimation

Vulnerability Analysis for Environmental Hazards

What are Hazardous Materials?

EDWARD J. CALABRESE

Hotspot Geoinformatics

Introduction and Background

Quantitative environmental and *ecological risk assessment* (see **Ecological Risk Assessment**) has been a relatively new and rapidly developing area during the past quarter century. Numerous important issues have been discussed in the two recent NAS/NRC reports entitled *science and judgment in risk assessment*, and *issues in risk assessment*. On the methodological side, a special issue of the international journal, *Environmental and Ecological Statistics*, is devoted to statistical issues and approaches to some current topics in environmental risk assessment [1].

We now live in the age of geospatial technologies. The age-old geographic issues of societal importance are now thinkable and analyzable. The geography of disease is now just as doable as genetics of disease, for example (see **Spatial Risk Assessment; Spatiotemporal Risk Analysis**), and it is possible to pursue it in an intelligent manner with the availability of the rapidly advancing geospatial information technology.

Government agencies continue to require meaningful summaries of georeferenced data to support policies and decisions involving risk maps, geographic targets, and resource allocations. So also the public with initiatives of digital governance, enabling transparency, efficiency, and efficacy in the risk assessment and management.

This article briefly introduces hotspot geoinformatics for hotspot detection and prioritization, and provides examples of societal importance involving environmental risk.

Ecological Diversity as a Motivating Example

Quantification for Ecological Diversity

Conservation biology, landscape ecology, and ecosystem-oriented natural resources management lend considerable urgency to issues and approaches concerning biodiversity assessment. Most of the traditional approaches and statistical tools are plot-based

with a goal of definitive characterization. Diversity, however, is relative to a spatial scale, temporal scale, and taxocene spectrum. Patterns may be more informative than absolutes in this regard.

The issues are fundamental in that, explaining the effects of environment on the distribution and abundance of species is the essence of much ecological work. The controversies arise from the intrinsic scientific importance of diversity theories, as well as from the broad economic and social ramifications of considering biodiversity in land use decisions. At the heart of the scientific and social controversies regarding diversity are problems of quantification, interpretation, and analysis.

The classical view of diversity remains important for intensive studies of particular ecological communities and forest stands [2]. However, the emerging sciences of landscape ecology and conservation biology have made evident the logistical and economical impracticality of such intensive observational coverage for regions in the order of square kilometers and larger [3]. These spatial scales are necessarily encompassed by contemporary ecosystem-oriented resource management and design of regional/national networks of biodiversity reserves. Furthermore, species/area and minimum viable population issues become fundamental in these matters.

The multidimensional character of diversity can be revealed by establishing an intrinsic, and index free, diversity ordering. In effect, diversity may appear to have decreased when viewed from one vantage point (i.e., index), and increased when viewed from a different perspective.

In view of the inadequacy of a single index, Patil and Taillie [4, 5] quantify diversity by means of diversity profiles. A diversity profile is a curve depicting the simultaneous values of a large collection of diversity indices. Thus, the profile portrays the views of diversity from many different vantage points simultaneously and in a single picture.

Differences in community diversity are studied by comparing profiles. If the two communities are intrinsically comparable, then one profile will lie uniformly above the other. Conversely, when the communities are not intrinsically comparable, their profiles may intersect. But even here, the profiles can reveal which portions of the community have undergone opposing diversity changes.

Indicators for Ecological Diversity

We proceed to consider ways of coping with complexity and confounding that embrace multiple indicators rather than agonizing over choices and conflicts of diversity measures. We contemplate enlarging the orders of indicators to encompass some interactions in a formal manner that accommodates both parameterization and visualization. We conclude by noting the convergence of biodiversity and ecological community concepts at meter scales, but not for broader landscape, regional, and global scales of ecological organization.

The indefiniteness regarding biodiversity that can give rise to frustration is well expressed by Taylor [6] in the following quote:

Diversity so pervades every aspect of biology that each author may safely interpret the word as he wishes and there is consequently no central theme to the subject. We cannot be sure if this flexibility is healthy or due to lack of discipline, but it can be traced back to the beginnings of interest in biological diversity ...

The recent programs of the US National Science Foundation probing biocomplexity in many contexts serve to provide evidence that the flexibility addressed by Taylor is both healthy and indicative of need for strengthening discipline with regard to scientific constructs and means by which they are made operational.

Indicators/expressions of this nature are appropriate, and therefore of value, if they convey the desired information within the budget and delivery delays that are acceptable. One method is more efficient than another if it conveys the required information either more rapidly or at less cost. Conveying more information at the same cost and timing is not necessarily desirable if unwanted information has to be processed or filtered.

Increasingly sophisticated management, intervention, remediation, and regulation require a continuing flow of multiple indicators for various aspects of ecosystems.

What ecosystem managers and regulators seek is a complementary set of indicators that captures aspects of interest. We have thus entered the exciting age of indicators; for ecological diversity and biodiversity.

And so also for environmental risk exactly in a similar manner.

Hotspot Geoinformatics of Detection and Prioritization

In geospatial and spatiotemporal surveillance, it is important to determine whether any variation observed may reasonably be due to chance or not. This can be done using tests for spatial randomness, adjusting for the uneven geographical population density as well as for age and other known risk factors. One such test is the spatial scan statistic, which is used for the detection and evaluation of local clusters or hotspot areas. This method is now in common use by various governmental health agencies, including the National Institutes of Health, the Centers for Disease Control and Prevention, and the state health departments in New York, Connecticut, Texas, Washington, Maryland, California, and New Jersey. Other test statistics are more global in nature, evaluating whether there is clustering in general throughout the map, without pinpointing the specific location of high or low incidence or mortality areas.

Scan Statistic Approach for Geospatial Hotspot Detection

Three central problems arise in geographical surveillance for a spatially distributed response variable indicator, generating a cell-wise constant cellular surface (see Figure 1). These are (a) identification of areas having exceptionally high (or low) response, (b) determination of whether the elevated response can be attributed to chance variation (false alarm) or is statistically significant, and (c) assessment of explanatory factors that may account for the elevated response. The circle-based spatial scan statistic

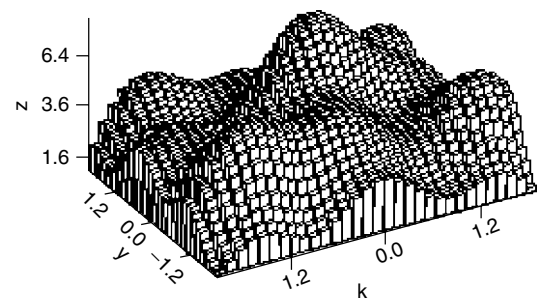


Figure 1 Cellular surface

[7, 8] has become a popular method for detection and evaluation of disease clusters. In space–time, the scan statistic can provide early warning of disease outbreaks and can monitor their spatial spread. The response variable/indicator in this case is disease rate in the case–control situation for binomial distribution and in the case count situation for Poisson distribution.

ULS Scan Statistic. A new version of the spatial scan statistic is designed for detection of hotspots of arbitrary shapes and for data defined either on a tessellation or a network. This version looks for hotspots from among all connected components of upper level sets (ULSs) of the response rate and is therefore called the *upper level set* scan statistic [9]. The method is adaptive with respect to hotspot shape since candidate hotspots have their shapes determined by the data rather than by some *a priori* prescription like circles or ellipses.

Continuous Response Situations. The circle-based scan statistic methodology can be extended to include continuous response distributions, such as, three parametric families of distributions: gamma distribution, lognormal distribution, and scaled beta distribution. The first two families apply to responses that can range from zero to infinity, while the third is for bounded responses. The overall approach is to model the mean and relative variance in terms of the size variable. These moments are functions of the parameters of the response distribution, so that a likelihood function can be written down and parameters estimated by maximum likelihood.

Our strategy for handling continuous responses is thus to model the mean and variance of each cellular response distribution in terms of the size variable Aa for cell a ; modeling is guided by the principle that the mean response should be proportional to Aa and the relative variability should decrease with Aa . Just as with the Poisson and binomial models, we take the Ya to be independent. The approach is best illustrated for the gamma family of distributions, as follows:

We parameterize the gamma distribution by (k, β) , where k is the index parameter and β is the scale parameter. Thus, if Y is a gamma distributed variate,

$$E[Y] = k\beta \text{ and } \text{Var}[Y] = k\beta^2 \quad (1)$$

Both k and β can vary from cell to cell but additivity with respect to the index parameter suggests that we take k proportional to the size variable:

$$ka = \frac{Aa}{c} \quad (2)$$

where c is an unknown parameter but whose value is the same for all a . This gives the following mean and squared coefficient of variation:

$$E[Ya] = \beta Aa/c \text{ and } \text{CV}^2[Ya] = c/Aa \quad (3)$$

Multivariate Hotspotting. When we have multiple indicators representing multidimensional landscape level cellular environmental risk, multivariate hotspotting has been conceptualized [10].

Hotspot Prioritization with Multicriteria Indicators

At times, several hotspots are discovered and in response to several stakeholders resulting in their criteria, scores, and/or indicators, the hotspots need to be prioritized and ranked without crunching the indicators into an index. This gives rise to a data matrix of rows for hotspots and columns for indicator scores. And the problem becomes a partial order theory problem and has been addressed in [11]. Broadly speaking, this paper is concerned with the question of ranking a finite collection of objects when a suite of indicator values is available for each member of the collection. The objects can be represented as a cloud of points in indicator space, but the different indicators (coordinate axes) typically convey different comparative messages and there is no unique way to rank the objects while taking all indicators into account. A conventional solution is to assign a composite numerical score to each object by combining the indicator information in some fashion. Consciously or otherwise, every such composite involves judgments (often arbitrary or controversial) about trade-offs or substitutability among indicators.

Rather than trying to combine indicators, we take the view that the relative positions in indicator space determine only a partial ordering and that a given pair of objects may not be inherently comparable. Working with Hasse diagrams of the partial order, we study the collection of all rankings that are compatible with the partial order (linear extensions).

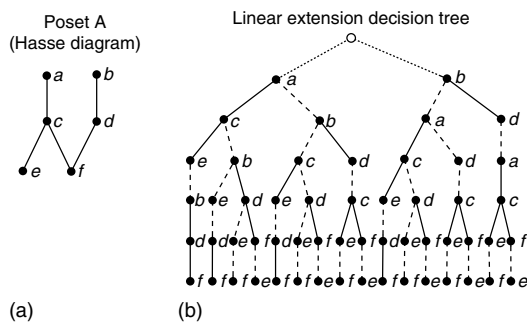


Figure 2 Hasse diagram of poset A (a) and a decision tree enumerating all possible linear extensions of the poset (b). Every downward path through the decision tree determines a linear extension. Dashed links in the decision tree are not implied by the partial order and are called *jumps*. If one tried to trace the linear extension in the original Hasse diagram, a “jump” would be required at each dashed link. Note that there is a pure-jump linear extension (path a, b, c, d, e, f) in which every link is a jump

See Figure 2. In this way, an interval of possible ranks is assigned to each object. The intervals can be very wide, however. Noting that ranks near the ends of each interval are usually infrequent under linear extensions, a probability distribution is obtained over the interval of possible ranks. This distribution, called the *rank-frequency distribution*, turns out to be unimodal (in fact, log-concave) and represents the degree of ambiguity involved in attempting to assign a rank to the corresponding object. See Table 1.

Stochastic ordering of probability distributions imposes a partial order on the collection of rank-frequency distributions. This collection of distributions is in one-to-one correspondence with the original collection of objects and the induced ordering

on these objects is called the *cumulative rank-frequency (CRF) ordering*; it extends the original partial order. See Figure 3. Although the CRF ordering need not be linear, it can be iterated to yield a fixed point of the CRF operator. We hypothesize that the fixed points of the CRF operator are exactly the linear orderings. The CRF operator treats each linear extension as an equal “voter” in determining the CRF ranking. It is possible to generalize to a weighted CRF operator by giving linear extensions differential weights either on mathematical grounds (e.g., number of jumps) or empirical grounds (e.g., indicator concordance). Explicit enumeration of all possible linear extensions is computationally impractical unless the number of objects is quite small. In such cases, the rank-frequencies can be estimated using discrete Markov chain Monte Carlo (MCMC) methods (see **Bayesian Statistics in Quantitative Risk Assessment**).

Some Examples

Drinking Water Quality and Water Utility Vulnerability

New York City has installed 892 drinking water sampling stations across the five boroughs. Each 4.5-ft high station is located outdoors and draws water from a nearby water main (see Figure 4). The purpose is to monitor general water quality, detect potential health threats, and thwart bioterror activity. Sampling frequency was increased after the 9/11 attacks and, currently, about 47 000 water samples are analyzed annually. Parameters analyzed include: bacteria, chlorine, pH, inorganic and organic pollutants, color, turbidity, odor, and many others. The

Table 1 Rank-frequency table for the poset of Figure 2. Each row gives the rank-frequency distribution for the corresponding element of the poset

Element	Rank						Totals
	1	2	3	4	5	6	
a	9	5	2	0	0	0	16
b	7	5	3	1	0	0	16
c	0	4	6	6	0	0	16
d	0	2	4	6	4	0	16
e	0	0	1	3	6	6	16
f	0	0	0	0	6	10	16
Totals	16	16	16	16	16	16	

network version of the ULS scan statistic can provide a real-time surveillance system for detecting and evaluating water quality hotspots within the distribution system on a parameter-by-parameter basis.

An overall assessment of water quality at each sampling station taking all parameters into account is achieved by employing recent progress on multi-criterion ranking using poset (partially ordered set) prioritization [11] (see **Water Pollution Risk**).

New York City Subway System Syndromic Surveillance

For certain problems, there is an underlying network structure on which we will want to perform the

cluster detection and evaluation. For example, the New York City Department of Health (DOH) is monitoring the New York subway system and water distribution networks for bioterrorism attacks. In such a scenario, a circular scan statistic is not useful as two individuals close to each other in Euclidian distance may be very far from each other along the network. However, the ULS methods will be employed for the detection and evaluation of clusters on a predefined network. The essentially linear structure of these networks, compared with tessellation-derived networks, is expected to have a major impact on the form of the null distributions and their parametric approximations.

The DOH and Metropolitan Transportation Authority (MTA) began monitoring subway worker absenteeism in October 2001 as one of several surveillance systems for the early detection of disease outbreaks. Each day the MTA transmits an electronic line list of workers absent the previous day, including work location and reason for absence. DOH epidemiologists currently monitor temporal trends in absences in key syndrome categories (e.g., fever-flu or gastrointestinal illness). Analytic techniques are needed for detecting hotspots within the subway network.

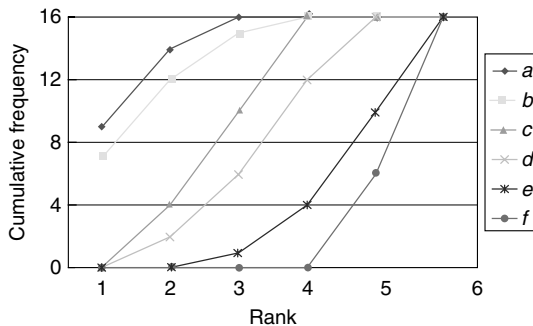


Figure 3 Cumulative rank-frequency distributions for poset A



Figure 4 New York city drinking water sampling station

Choice of Indicator Class for Watershed Biological Impairment

There has been considerable work on determining a suitable method to accomplish a satisfactory ordering of a group of objects, when there are multiple evaluation criteria. Data from 21 watersheds of the Atlantic slope consortium (ASC) has been examined with the goal of determining an accurate ranking by overall watershed condition. In particular, there are three levels of indicators that range from Level I to Level III, increasing in the quality and accuracy of the data as well as the cost and effort needed to obtain the data. Due to the high cost, Level III data is available only for six watersheds; however, the interest is in ranking all 21 watersheds [12]. We use elements of poset (partial order set) theory as a foundation for our analysis. The poset linear extension method can be used to find rankings without using an index, relying only on pairwise comparisons of the objects.

The data relates to a set of 21 watersheds from the ASC, which has the goal for determining an accurate ranking of the health of the watersheds [13] (see

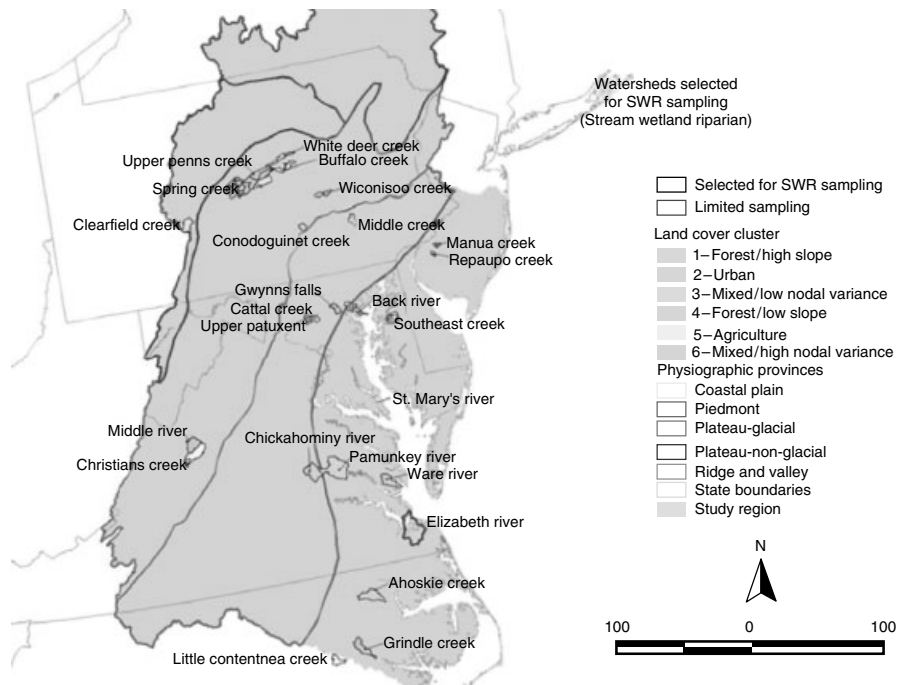


Figure 5 The 21 watersheds from the mid-Atlantic area

Figure 5). There are 15 indicators each of which measures a facet of the watershed and are grouped into Level I to Level III, increasing in quality and accuracy as well as amount of cost and effort. For all 21 watersheds, we have data for seven Level II indicators, and for five Level I indicators. The data matrix for Level II indicators has 21 rows and 7 columns, and the data matrix for the Level I indicators has 21 rows and 5 columns. For Level I and II, the data was scored between 0 and 1. The three levels are as follows:

- Level III – intensive field assessment
- Level II – rapid field assessment
- Level I – landscape assessment from geographic information systems (GISs).

The indicators are categorized by the level of accessibility and availability. Level III, intensive field assessment, is the most expensive among the three levels. We consider the indicators of Level III, the best quality of data we have regarding the watershed. Due to the money and effort in the procedure of obtaining this data, it is available only for six watersheds.

The three Level III indicators are collected from the Index of Biological Integrity (IBI) developed by the Environmental Protection Agency (EPA). They are benthic IBI, fish IBI, and NO_3 . The two IBI indicators are biological indicators, while the NO_3 is considered a chemical indicator.

The Level II rapid field assessment is obtained from on-site sampling. A certain level of expertise is involved in the field assessment. Generally, Level II data is relatively cheap compared to the Level III data. The Level I landscape assessment is the satellite data, and is the easiest to access and the least expensive. The poset prioritization and ranking analysis of the data recommends use of Level II or even Level I indicators to estimate the true ranking of the 21 watersheds by their biological condition.

More Examples and Applications

Geospatial and spatiotemporal environmental risk mapping commands vast literature. Some publications concentrate on geospatial cellular modeling of environmental risk, the probability of adverse event, and thus providing a cellular surface. Some provide

discrete or continuous profiles, allowing extraction of indicators, whereas still others, attempt to identify and provide surrogate indicators for the environmental risk. Applications abound. See, for example, publications and/or websites for:

1. Regional Vulnerability Assessment Indicators of the USEPA ReVA program.
2. Indicators of Ecological Value, Ecological Sensitivity, and Anthropogenic Pressure of the Map of Italian Nature Program.
3. Tsunami Inundation Risk Mapping and Management Program of NOAA.
4. Environmental Risks and Surrogate Indicators around the World for Forest Fires, Pests, Insects and Diseases, Infectious Disease Vectors, Floods, Landslides, Droughts, and others.
5. The Next Generation of Ecological Indicators of Wetland Condition, EcoHealth, the Journal.
6. Ecological Indicators, the Journal.
7. For several environmental and ecological landscape level indicators and related analyses, see Johnson and Patil [14] and Myers and Patil [15].

Future Directions

Surveillance geoinformatics of hotspot detection and prioritization for environmental risk is a critical need of the twenty-first century. Next generation decision support system within this context is crucial. It will be productive to build on the present effort in the directions of prototype and user-friendly methods, tools, and software, and also in the directions of thematic groups, working groups, and case studies important at various scales and levels. The authors have such a continuation effort in progress within the context of digital governance with the National Science Foundation (NSF) support and would welcome interested readers to join this collaborative initiative.

Acknowledgments

This material is based upon work supported by (a) the National Science Foundation under Grant No. 0307010, and (b) the United States Environmental Protection Agency under Grants No. CR-83059301 and No. RD-83244001-0. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the agencies.

References

- [1] West, W. & Piegorsch, W. (2009). Modern benchmark analysis for environmental risk assessment, *Environmental and Ecological Statistics* (in press).
- [2] Gove, J., Patil, G.P. & Taillie, C. (1994). A mathematical programming model for maintaining structural diversity in uneven-aged forest stands with implications to other formulations, *Ecological Modelling* **79**, 11–19.
- [3] Scott, J., Csuti, B., Estes, J. & Anderson, H. (1989). Status assessment of biodiversity protection, *Conservation Biology* **3**, 85–87.
- [4] Patil, G.P. & Taillie, C. (1979). A study of diversity profiles and orderings for a bird community in the vicinity of Colstrip, Montana, in *Contemporary Quantitative Ecology and Related Econometrics*, G.P. Patil & M. Rosenzweig, eds, International Cooperative Publishing House, Fairland, pp. 23–48.
- [5] Patil, G.P. & Taillie, C. (1982). Diversity as a concept and its measurement, *Journal of the American Statistical Association* **77**, 548–567 (invited discussion paper).
- [6] Taylor, L.R. (1978). Bates, Williams, Hutchinson – a variety of diversities, in *Diversity of Insect Faunas*, I.A. Mound & N. Waloff, eds, Blackwell Scientific Publications, Oxford, pp. 1–18.
- [7] Kulldorff, M. & Nagarwalla, N. (1995). Spatial disease clusters: detection and inference, *Statistics in Medicine* **14**, 799–810.
- [8] Kulldorff, M. (1997). A spatial scan statistic, *Communications in Statistics: Theory and Methods* **26**, 1481–1496.
- [9] Patil, G.P. & Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots, *Environmental and Ecological Statistics* **11**, 183–197.
- [10] Modarres, R. & Patil, G.P. (2008). Hotspot detection with bivariate data, *Journal of Statistical Planning and Inference* in press.
- [11] Patil, G.P. & Taillie, C. (2004). Multiple indicators, partially ordered sets, and linear extensions: multi-criterion ranking and prioritization, *Environmental and Ecological Statistics* **11**, 199–228.
- [12] Patil, G.P., Bhat, K.S., McKenney-Easterling, M. & Kase, M. (2007). A toolbox for the choice of indicator classes for ranking of watersheds, *Environmental and Ecological Statistics* (Manuscript).
- [13] Brooks, R., McKinney-Easterling, M., Brinson, M., Rheinhardt, R., Havens, K., O'Brien, D., Bishop, J., Rubbo, J., Armstrong, B. & Hite, J. (2006). *A Stream-Wetland-Riparian (SWR) Index for Assessing Condition of Aquatic Ecosystems in Small Watersheds Along the Atlantic Slope of the Eastern U.S.* (Manuscript).
- [14] Johnson, G.D. & Patil, G.P. (2006). *Environmental and Ecological Statistics Series: Volume 1: Landscape Pattern Analysis for Assessing Ecosystem Condition*, Springer, New York.

- [15] Myers, W. & Patil, G.P. (2006). *Environmental and Ecological Statistics Series: Volume 2: Pattern-Based Compression of Multi-Band Image Data for Landscape Analysis*, Springer, New York.

Environmental Health Risk

GANAPATHI P. PATIL, SHARADCHANDRA W.
JOSHI AND STEPHEN L. RATHBUN

Related Articles

Environmental Hazard

Human Reliability Assessment

Overview of Human Reliability Assessment

Human reliability assessment (HRA) is concerned with predicting the impacts of human performance and error on risk. HRA is a hybrid discipline, belonging simultaneously to the fields of risk assessment and reliability engineering, as well as human factors and ergonomics (the study of humans in work environments). It seeks to deal with the difficult issue of human error, which can cause or contribute to accidents, but also with human error recovery, which can save lives and prevent or reduce system losses. The human operator (a generic term used throughout this chapter) is the key ingredient in many industries and work systems, including nuclear power plants, chemical and oil and gas installations, aviation and air traffic management systems, rail networks, space missions, military systems, and hospitals. These are all areas where HRA is currently in use or being developed to help avoid accidents leading to fatalities and unacceptable system losses.

HRA usually fits into a risk (or reliability) assessment, which is concerned with how a system can fail and/or lead to accidents involving loss of life or resources. Risk assessment asks several questions:

- What is the scope of the system under investigation?
- What can go wrong?
- How likely is it to go wrong?
- What will be the consequences?
- Is the risk (the product of the undesirable event's likelihood and consequences) acceptable?
- Can risk be reduced?

Figure 1 depicts the main stages of risk or safety assessment [1]. HRA's role in risk assessment is to evaluate the human contribution to risk, which can be positive or negative. HRA's functions therefore reflect those of risk analysis. In particular, HRA usually attempts to answer the following questions:

- What is the human involvement in the system?
- What can go wrong?

- How likely is it to happen?
- What can be done to reduce error impact?

The first question is usually answered by the application of *task analysis* (see **Systems Reliability**) [2], which defines the human being's tasks, procedures to be followed, equipment required, etc. For example, an air traffic controller might give the pilot of an aircraft an instruction to climb to a particular flight level, using a radiotelephone. The task would be "give instruction to pilot". The second question is served by *human error analysis* (see **Reliability Data; Probabilistic Risk Assessment**), a range of techniques that aim to identify human errors that could cause or contribute to system failure or degradation. A typical example is that the pilot mishears the flight level given by the controller, and climbs to a higher level, one which is perhaps already occupied by another aircraft. Such an error obviously can introduce a risk into the system, namely the risk of a midair collision. The third question is answered by using the techniques of *human error quantification*, which estimate how likely such an error is to happen, expressed it as a probability. The *human error probability* (HEP) (see **Behavioral Decision Studies**) is ideally a statistical entity derived from observations or tasks and errors that arise during their execution, e.g.:

$$HEP = \frac{\text{number of errors observed}}{\text{number of opportunities for error}} \quad (1)$$

In this example, on the basis of research into this type of error, it may be expected to occur approximately three times in a 1000 opportunities (instructions). This sounds reasonably low, but of course there are many thousands of flights each day. To compensate for this, there are a number of recovery and mitigation factors built into the system that the risk assessment can take account of.

First, the copilot in the aircraft may detect the error, since he or she will overhear the controller-pilot interaction. Second, the pilot should "read back" the flight level to the controller, who can therefore detect the "mishearing". Third, the controller is watching the aircraft on a radar screen, and may realize it is climbing beyond its "cleared" level. Fourth, there is an alarm to warn the controller if two aircraft get too close, and also (fifth) a completely independent alarm aboard the aircraft to alert the pilots and to tell them how to avoid the other aircraft. Sixth, the other aircraft will be listening on the same

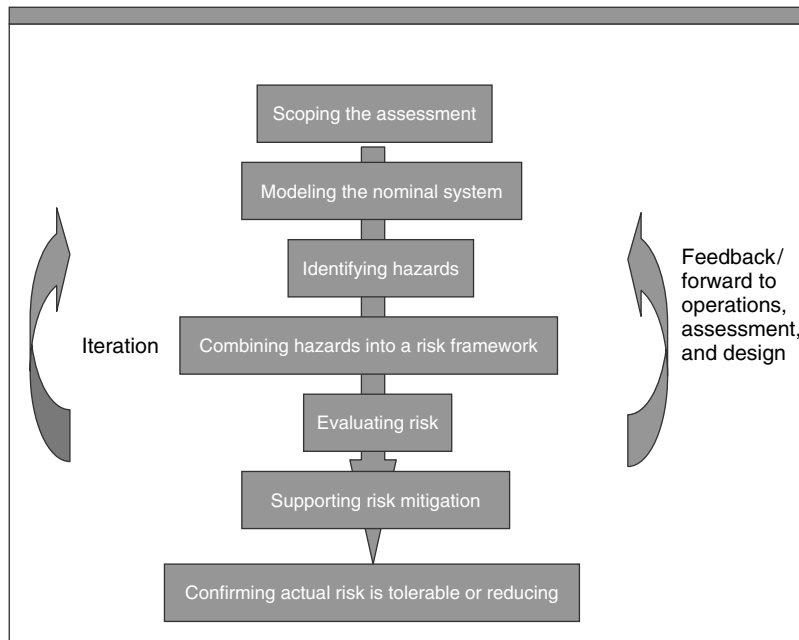


Figure 1 A generalized seven-stage safety assessment process

frequency and may hear the instruction, whereupon the pilot of that aircraft will contact the controller immediately. Last, there is sufficient space in the skies that chances of midair collision are still rare even if all these system factors fail to prevent the error propagating toward disaster. The human error rate may therefore be deemed tolerable, because its effects can be absorbed by other system defenses. Alternatives such as full automation may seem like a solution, but experience shows that automation can introduce its own problems, in particular a lack of recovery when it goes wrong, something that would be unacceptable when considering mass public transportation by air, for example.

This example at first sight suggests that human reliability is quite good (e.g., 3 errors in a 1000 opportunities is not so bad), but then when the sheer number of flights and controller–pilot interactions are considered, it appears problematic. In fact, upon detailed risk analysis, there are many recovery layers in the system, so risk is acceptable. But for the midterm future (e.g., 2020), where the number of flights is predicted to double or even triple in line with public demand for more air transportation, it may be necessary to introduce a new layer or defense. This could involve *human error reduction*, e.g., better

training to ensure the copilot or controller hears the error, or a system-based mitigation, such as sending an electronic message from the aircraft’s onboard flight management system to the ground control system, to compare what the controller input into his system with what the pilot input into his (a consistency check), and thereby alarming much earlier than current systems allow if an error occurs.

This example shows that the human reliability for a task needs to be seen in the context of the system to make meaningful decisions about risk. This tends to be true whether considering aviation, nuclear power, rail, space, or medical contexts. Whereas human factors [3] often work independently, HRA therefore mainly occurs within the framework of a risk or reliability assessment [4]. HRA analysts usually work closely with risk analysts, as well as system designers and operational personnel. The HRA analyst’s job may be described as being that of a detective investigating a negative “event”, but where the “event” has not yet occurred.

The HRA Process

A generic HRA process is shown in Table 1 [5], from scoping the problem the HRA is being asked

Table 1 The HRA process

<i>Problem scoping</i>	– Deciding what is in the scope of the assessment – e.g., the principal human interactions and functions; what nonnominal events could occur (including adverse weather and traffic patterns); what failures, whether maintenance activities are included; etc. This “context” is normally set by the safety assessor and HRA/human factors practitioner.
<i>Task analysis</i>	– The detailed analysis of the human contribution in the risk scenario – what the controllers/pilots should do, and with what resources (displays, communication media, etc.).
<i>Error identification</i>	– What can go wrong – from simple errors of omission (e.g., fail to detect wrong readback) to more complex errors of commission (e.g., right instruction given to wrong aircraft).
<i>Representation</i>	– Modeling the human errors and recoveries in the risk model, whether this uses fault and event trees, Petri nets, or some other modeling approach. This puts human errors and recoveries into the larger system context, alongside other system failures and defenses. Representation enables an accurate view of how errors propagate through the system, and whether they will indeed ultimately affect system risk.
<i>Dependence analysis</i>	– Explicit quantitative accounting for dependence between different errors and error conditions.
<i>Screening</i>	– Conservative (pessimistic) error probabilities can be used along with sensitivity analysis to determine which human errors are “risk sensitive” and which will have negligible impact on risk. The former can then be quantified in more depth, and the latter ignored.
<i>Quantification</i>	– Derivation of HEPs for errors using a formal quantification method. Quantification may also develop uncertainty bounds or confidence levels for individual HEPs.
<i>Evaluation</i>	– Determination of the impact of human error on risk, and the most important human contributions to risk, within the risk framework or representation. This usually involves comparison of the total predicted risk against a target level of safety, or else future system’s risk will be contrasted with today’s risk levels, to ensure that the future system will be at least as safe as today’s.
<i>Error reduction</i>	– On the basis of the evaluation and the prior qualitative stages (task analysis, error identification and representation), human factors and/or safety requirements may be specified to reduce the HEP and/or its impact on risk to achieve a more acceptable risk level.
<i>Documentation</i>	– Documentation of the analyses, calculations, assumptions, and recommendations to inform the designers, implementers, and operators/managers, as well as to the appropriate regulatory authority.

to address, through to documentation. Within the confines of this chapter on HRA, only 3 of these 10 steps will be covered: task analysis, error identification, and quantification (with brief treatment of dependence). Most emphasis is placed on the error identification and quantification stages in HRA, the first to try to make the concept of human error clearer and the second because it is most often seen as the quintessential element of HRA. These two steps generally dominate HRA outputs. For a fuller description of all 10 steps, see [5].

The remainder of this chapter outlines the “mechanics” of HRA – what it looks like and how it works, showing examples of actual techniques. It then considers the validity of the central approach, that of human error quantification, and finally discusses contemporary HRA issues including second generation HRA techniques, safety culture, and HRA “ethics”.

How HRA Works

This section gives examples of practical techniques and approaches of task analysis, error identification, and quantification (including dependence aspects) and error reduction. For more general background information on HRA, see [5–8].

Task Analysis

Task analysis (*see Systems Reliability*) [2] is a set of methods for capturing and describing the human operator involvements in a system. The classic form of task analysis is hierarchical task analysis (HTA), an example of which is shown in Figure 2. The figure is read top-down and starts from a goal – in this case “launch lifeboat” from an offshore oil and gas installation during an emergency. This goal is served by five subsidiary tasks which follow a plan

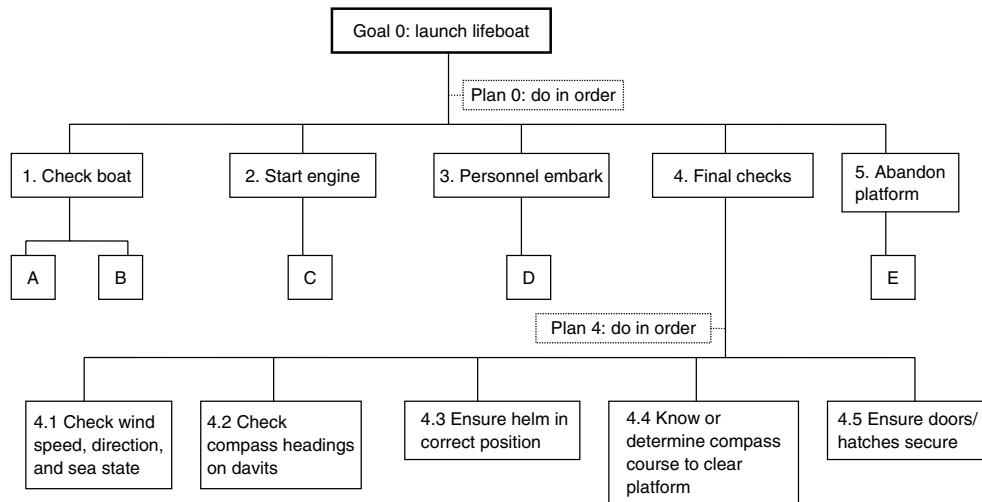


Figure 2 Example HTA for launching a lifeboat

requiring them to be carried out in order, e.g., check boat. This task requires two operations, shown in the diagram simply as “A” and “B” but in fact they are “check maintenance valve is closed” and “release retaining hook”. This simple type of hierarchical decomposition of a task is in fact a good way to define the operator’s role. Task four is decomposed in full detail to show the types of human “operations” that achieve the task “final checks”. This type of task analysis is a useful starting point for a HRA analyst and acts as a blueprint of what the human operator is supposed to do in a system.

Incidentally, this style of task analysis is also used by training departments to determine training needs, as well as by those operational groups that must write formal procedures for operators. Other formats of task analysis [2] can be used to focus on, for example, team interactions under time constraints (using vertical and horizontal timeline analysis), operator interaction with a human machine interface such as a computer screen (tabular task analysis), distributed operations in different locations (link analysis), etc. Task analysis can therefore describe the “nominal case”, i.e., what the operator(s) should do – the rest of HRA then considers what can go wrong and how to increase reliability.

Human Error Identification

Human Error Theory. Once task analysis has been completed, human error identification can then

be applied to consider what can go wrong. Human error has been defined as follows:

Any member of a set of human actions or activities that exceeds some limit of acceptability, i.e., an out-of-tolerance action where the limits of human performance are defined by the system [9].

Such a definition gives rise to the following overt error types that should be considered during any human error analysis:

- Omission error – failing to carry out a required act.
- Action error – failing to carry out a required act adequately:
 - act performed without required precision, or with too much/little force;
 - act performed at the wrong time;
 - act performed in the wrong sequence
- extraneous act – act performed which is not required (also called error of commission, EOC).

This classification tells us neatly what can happen, but gives little indication as to the causes or remedies for such errors. It also fails to capture the notion of different types of behavior and error forms, in terms of different causes and “internal” or cognitive failure mechanisms, and the important notion of intentionality of error, i.e., that some errors are intentional and others are not.

Although HRA is very applied in nature, the analyst must have an understanding of the theory of human error in order to be effective in identifying what can go wrong. Otherwise, error identification will be superficial and risk may be underestimated. To add more depth to the notion of human error, reference is usually made to four error categories [10]:

- *Slip*

Where the wrong control is activated, e.g., an error while dialing a number to make a phone call, or the wrong message is given, e.g., slips of the tongue.

- *Lapses*

Where something is forgotten, a task step missed out – e.g., when making tea, filling the kettle, but forgetting to switch it on.

- *Mistakes*

Where due to a misunderstanding or misconception, the operators do the wrong thing, e.g., in the Three Mile Island nuclear power accident in 1979, operators believed that the reactor pressure vessel had plenty of water, so switched off the emergency supply when it tried to fill the emptying vessel, thus leading to reactor core damage.

- *Violations*

Where the operators know what they are doing is not allowed, but they do it anyway, usually to make the job easier, e.g., stepping over a barrier around an automated assembly line to retrieve something dropped, without stopping the robot arm that is moving rapidly inside the barrier. Reasons for such a violation could include not wishing to lose valuable production time by stopping the process, and the risk of personal harm could be perceived as low.

Slips and lapses are the most predictable error forms, and are usually characterized as being simple errors of quality of performance or as temporary lapses in memory, often due to distraction. In the lifeboat launching example, for example, a slip might concern pressing the wrong button to start the engine (there are several buttons on the console). But this would normally be rectified quickly (since the engine would not start – nothing would happen). A lapse might involve failing to ensure that all the hatches are secure, which could have more serious consequences when the boat enters the water. General good human factors practice can eradicate most of these types of errors, and good design will tend to mean

that operators who commit them usually discover and correct them before they have adverse consequences on the system. This category, however, also includes slips and lapses in maintenance and testing activities, which may lead to *latent failures* (failures whose impact is delayed, and whose occurrence may be difficult to detect prior to an accident sequence). For example, an emergency backup system may be taken off-line for testing and then it may be tested and deemed working in good order, but the maintenance crew may omit the last critical step of placing it back on-line. This means that if it is suddenly needed in an emergency, it will be unavailable. This is why functional testing is recommended after maintenance activities have taken place. Latent failures are important because they can remove barriers to accidents, while letting the operators believe the barriers are perfectly healthy.

Mistakes refer to the case wherein a human operator or team of operators might believe (due to a misconception) that they are acting correctly, and in so doing might cause problems and prevent automatic safety systems from stopping the accident progression. In a recent air transport accident, a controller saw two aircraft on a collision course and instructed one of the aircraft to descend, unaware that the other aircraft had started descending too (in such cases, one should descend, the other should climb). The pilots onboard trusted the controller more than the onboard automatic collision avoidance system that was telling them to climb, with the result that the two aircraft did in fact collide in midair, with total loss of all passengers and crew on both aircraft. Such errors of misconception are obviously highly dangerous for any industry, since they can cause failure of protective systems.

Violations represent a phenomenon that has only recently been considered in risk assessment, most notably in the hazardous industries since the Chernobyl nuclear power plant accident in 1986. Rule violations inevitably involve some element of risk-taking or risk-recognition failure. There are two basic types [11]: the first is the *routine* or *situational* violation, where the violation is seen as being of negligible risk and therefore the violation is seen as “acceptable” or even a necessary pragmatic part of the job. The second is the *extreme* violation, where the risk is largely understood as being real, as is the fact that it is a serious violation, with punitive penalties attached if discovered. The latter types of violation are believed

to be rare in industry. An example of an extreme violation would be falsification of administrative control procedures, e.g., to say that a test had been carried out successfully, when it had not in fact occurred or had yielded a negative result. Such incidents as these are usually due to “production pressure” and may be seen as a way of keeping one’s job or ensuring the success of the company. Note that actual sabotage and terrorist acts are clearly criminal in nature, and are therefore outside the jurisdiction of HRA.

An EOC is one in which the operator does something that is incorrect and also *not required*. Examples are where a valve which should be locked open is found to be locked closed, or a computer-controlled sequence is initiated, or a safety system disabled, when none of these actions are required, and in some cases when there has not even been a task demand associated with the equipment unit. Such errors can arise due to carrying out actions on the wrong components, or can be due to a misconception, or can be due to a risk-recognition failure. An EOC can be due to a mistake or a violation, but even a simple slip could lead to an EOC, e.g., by activating a completely different (but adjacent) control rather than the one intended. EOCs are of increasing concern, for three reasons: firstly they do appear to happen, even if rarely; secondly, they can have a large impact on system risk and thirdly, they are very difficult and/or laborious to identify in the first place. This means that they may therefore be underestimated in terms of their contribution to risk, or may not even be represented in the risk assessment. Their identification and quantification is difficult, however, and is returned to under the discussion of second generation techniques in the latter part of this chapter.

Human Error Identification – Practice. For a review of human error identification techniques, see [12]. This section outlines four straightforward and accessible approaches.

Operational experience review/critical incident technique. For an existing system, the first and most obvious way to identify errors is reviewing what errors have already led to incidents or accidents in the system under investigation. Most mature industries that are safety critical also have formal safety-related event reporting, recording, and analysis systems, so that the organization can learn from incidents and

avoid their repetition or escalation toward more serious accidents. Such accidents usually involve fatalities and all the grievous losses associated with such tragic events. Additionally, large-scale accidents, particularly those which attract media attention (and most do), can also threaten the very survival of the company. However, not all industries have mature human factors causal analysis systems, instead perhaps simply recording an event as “human error” with no further analysis of why it occurred, and hence how to prevent recurrence. But if an organization has a mature “risk management culture”, then it should have formal human error classification as part of its event reporting procedures and processes.

When carrying out an HRA for a system which already exists or has an equivalent referent system, it is very beneficial to interview operators (and also trainers and supervisors, and of course incident investigators) to ask what types of events have happened. The analyst can simply ask if the interviewee (a single person or a group) knows of any human error-related events that have occurred in the existing system. While this is anecdotal, it can usually be corroborated by existing formal incident records, and such interviews often generate a richness of information, particularly about factors that can affect performance and the existing safety culture in the organization. Two important caveats are first that the analyst has to beware of people’s particular biases (we all have them) such as a potential grudge against the company or the fact that “everyone is a hero in their own story”, and second the analyst must avoid “leading the witness” or putting words into their mouths, based on the analyst’s own values and preconceptions.

At the least, reviewing incident and operational experience ensures that the analyst will be able to identify credible errors that have already happened and will gain a fuller understanding of the factors that affect human reliability in the system’s operational culture.

TRACER. In the early years of applied HRA (the early 1980s), errors were initially identified by risk assessors without any special techniques or processes. This changed by the introduction of structured approaches for error identification, an early example being the systematic human error reduction and prediction approach (SHERPA) [13], which used a series of flowcharts and questions to help the risk

or HRA analyst consider which errors could affect the system under investigation. A common approach used today is TRACER (technique for the retrospective analysis of cognitive errors) [14] which is publicly available. It is a descendant of SHERPA as well as other techniques, and uses a series of taxonomies and flowcharts to identify error possibilities for a system. It also considers error detection and recovery. TRACER is a single-analyst human error identification approach. It was developed to classify the causes of incidents, and then a further version (TRACER Lite) was developed for predictive purposes. It requires first a task analysis, and then applies a series of guide words and other error-related taxonomies to identify both errors and (qualitative) error recovery likelihoods. The framework of TRACER is shown in Figure 3 and one of its main error taxonomies is shown in Table 2.

Note that in Table 2 there is a column called *internal error mechanism*: these guide words are psychological in nature and attempt to describe likely psychological or cognitive mechanisms for the failure.

These may be helpful when later on considering error reduction. As an example, if an action is predicted to be omitted, error reduction will be different if the most likely internal error mechanism was “distraction” as opposed to “insufficient learning”. In the former case, changes to the work environment will need to occur to remove the distractions, whereas in the latter more comprehensive training is likely to be recommended.

Application of TRACER is similar to using failure mode and effects analysis (FMEA) for hardware reliability assessment. The other aspect of TRACER that is noteworthy is that TRACER also considers the likelihood of error recovery.

Human HAZOP. Another approach to identifying errors is called *human hazard and operability study* (HAZOP) [15]. This is a group-based approach which uses a set of guide words of error types applied to a task analysis, with the group comprising a chairperson or group facilitator, a note-taker, a human factors expert, a safety expert, a system designer,

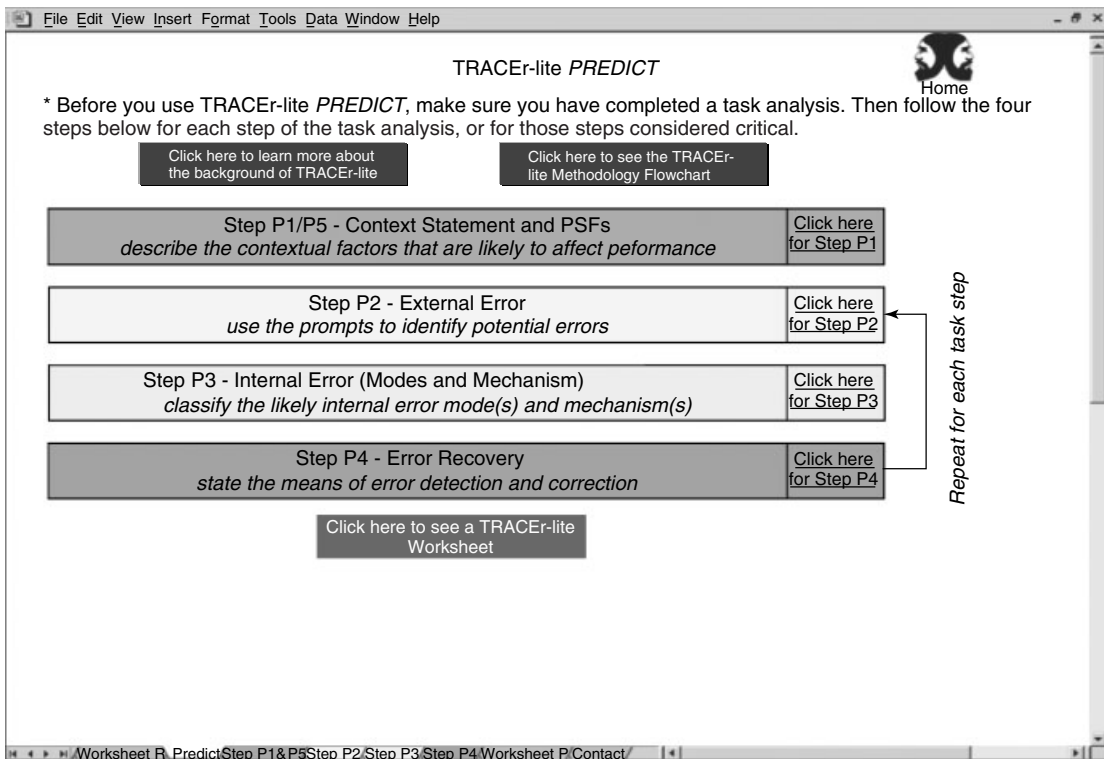


Figure 3 TRACER-Lite overview [Courtesy of Dr. S Shorrock]

Table 2 TRACER-Lite internal error taxonomy (guide words)

Internal error mode	Internal error mechanism
<i>Perception</i>	
Mishear	Expectation
Mis-see	Confusion
No detection (auditory)	Discrimination failure
No detection (visual)	Perceptual overload
	Distraction/preoccupation
<i>Memory</i>	
Forget action	Confusion
Forget information	Memory overload
Misrecall information	Insufficient learning
	Distraction/preoccupation
<i>Decision making</i>	
Misprojection	Misinterpretation
Poor decision or poor plan	Failure to consider side- or long-term effects
Late decision or late plan	Mind set/assumption
No decision or no plan	Knowledge problem
	Decision overload
<i>Action</i>	
Selection error	Variability
Unclear information	Confusion
Incorrect information	Intrusion
	Distraction/preoccupation
	Other slip

and operational experts. Table 3 shows typical guide words used in a human HAZOP. Table 4 then shows an extract of the type of discussion that can occur in a HAZOP, and Table 5 shows an example of how the results are recorded.

HAZOP is a risk and reliability approach and human HAZOP is simply an adaptation to HRA. However, in the few validations (see [16]) that have taken place of error identification techniques, human HAZOP has been shown to be very effective, particularly in identifying credible errors in the areas of violations.

The format in Table 5 can equally be used to represent the outputs from usage of TRACER or SHERPA or other human error identification techniques. It is a useful means of documenting the error's nature and impacts on the system, including barriers that should prevent the error or its effects, and a useful medium with which to identify preliminary error reduction or mitigation measures.

HAZOP has three significant advantages. Firstly, when the safety assessor is trying to identify hazards for a future system, one where there is,

Table 3 Typical human HAZOP guide words*Basic guide words*

- No action
- More action
- Less action
- Wrong action
- Part of action
- Extra action
- Other action
- More time
- Less time
- Out of sequence
- More information
- Less information
- No information
- Wrong information

Additional concepts

- Purpose/intention
- Clarity of task/scenario
- Training/procedures
- Abnormal conditions
- Maintenance
- Production pressure/risk-taking

Table 4 Excerpt from HAZOP discussion (of electronic strip system design for air traffic controllers)^(a)

HAZOP group member	Discussion
Human factors specialist	If these two objects overlap, could the controller operate on the wrong object, i.e., aiming a message for the one on top but actually communicating with the one that is now flying beneath the top one?
Designer 1	Well, we expect the controller always to move the objects so that they won't overlap, before transmitting a command.
Controller	Well, there might not always be time, but as long as you're operating on the one on top, one coming underneath won't be selected without it being clear, will it?
Designer 1	Hmm. Well, it isn't impossible actually, depending on how long you leave the cursor without entering a command. Presumably that would be no more than a couple of seconds, would it?
Controller 2	Not necessarily, I mean if I'm in the middle of something and then a higher priority call comes in, I'll leave the cursor there and then come back to it. It could be a while, up to a minute.
Designer 2	Right. Okay, we need to take another look at this, and implement some way of highlighting that the original target has been deselected and must be reacquired, otherwise the right message could be sent to the wrong aircraft.
Chairman	So, are we agreed then that we need an action on the designers here to ... ?

^(a) Reproduced from [11]. © Ashgate Publishing, 1998

as yet, no mature operational concept or procedures, this can be very difficult. There are simply too many open questions to make progress. However, in a HAZOP where experienced operators are present, these experts can usually “fill in the gaps”, interpolating between current practice and future practice of how the system should operate. This means that the preliminary hazards can be identified at a time when the design can still be changed significantly if serious hazards emerge. The second and related advantage is that the composition of the HAZOP group, comprising project design and controller reviewers, often lends itself to promptly identifying design solutions that will increase safety. Once the hazard is identified, the participants will often determine an immediate solution *via* a system change or procedural modification (as in Table 4). The third advantage is that such a process involves the project design and operational personnel in safety discussions. If they really believe a hazard is incredible they will say so, but if not, there is acceptance that such a hazard must be addressed. The approach therefore gains “buy-in” or ownership from the project team. They “own” the results as much as the safety assessors, rather than being given an “external” safety assessment that is seen as a “fait accompli”, which they have to accept only on the basis that they trust that the safety people know what they are doing.

The disadvantages of HAZOP are the resources requirements, since around six experts are needed

to run a HAZOP, and progress can sometimes seem slow. Also, experts may be biased not to identify failures (e.g., if they represent the project or development team). The ideal approach is therefore to utilize an analytic approach like TRACER as well as a human HAZOP approach – this is the most comprehensive formula for identifying all likely errors. If TRACER is carried out first, then its results can be used judiciously to steer the resources of a HAZOP team more effectively, e.g., ignoring already identified “minor” errors with little impact and focusing on the more serious, potential errors and risks of violations, etc.

JHEDI. A final checklist approach is shown in Table 6, from a nuclear power technique known as *justification of human error data information* (JHEDI) [16]. This straightforward technique can be used by a single assessor as a preliminary screening error identification tool.

Once error identification is carried out, there are two paths open to the HRA analyst or risk analyst. The first is where only a quantitative approach is being adopted, i.e., the intent is to examine the vulnerability of the system and seek ways to improve it. In such a case, it is likely that the errors identified (by whatever means) will be ranked or categorized on the order of importance and the assessors will move directly to the step of error reduction or mitigation.

The second path is taken when the HRA is part of a quantitative risk assessment approach. In such

Table 5 Extract of related ATM HAZOP output^(a)

Function	Guide word	Cause	Consequence	Indication	System defenses	Human recovery	Recommendations
Highlight object (aircraft label)	No action	Another item preventing access to target	Difficulty in hooking target aircraft	No highlighting of target	None	Drag blocking object out of way; strategic management of screen items	Design objects to “roll around” each other; use height filtering; flip system to move between object on top and the one beneath; highlight background
Other action	Clustering results in different aircraft being highlighted instead of target		Instruction may be given to wrong aircraft on the system	As above	Highlighting is color coded to indicate direction of travel; call sign is displayed on all menus	As above	As above

^(a) Reproduced from [15]. © Copernicus Publications, 2001

Table 6 JHEDI human error analysis questions

-
1. Could a misdiagnosis or misinterpretation occur, possibly due to a misperception of signals, or expectations, the complexity of the situation itself, or a failure to take note of special circumstances?
 2. Could a short-cut or even rule violation occur, due to production pressure or due to a failure to be aware of the hazard?
 3. Could a previous (latent) maintenance or calibration error lead to difficulties or errors in responding to the scenario?
 4. Could a check fail be omitted or occur on the wrong system, or simply be the wrong kind of check?
 5. Could an information–communication error occur?
 6. Could a crew-functioning error occur, such that, for example, it is not clear who is carrying out the task?
 7. Could a required act be omitted because it is at the end of the task, because of a lapse of memory or distraction, or because the operator failed to respond (e.g., due to his or her workload)?
 8. Could the task be carried out in the wrong sequence, or a subtask be repeated, or the operator lose his or her place in the procedures?
 9. Could an action or check be carried out too early or too late?
 10. Could an action be carried out with insufficient precision or force, or just inadequately?
 11. Could an action that is not required be carried out by operating on the wrong object or in the wrong location, or by carrying out an incorrect but more familiar task, or by selecting or acting on the wrong information?
 12. Could any other type of error occur?
-

cases, the human errors will be “represented” in logic trees such as fault and event trees, or in Petri nets if using dynamic risk modeling – these aspects are more properly dealt with in other chapters. But if this pathway is being pursued, the next step requires human error quantification, i.e., determining the probability of human error occurring (where possible together with its uncertainty bounds and probability distributional characteristics). This step in HRA is usually seen as the major part of HRA, and certainly most research over the past 45 years since the birth of HRA has been in this area.

Human Error Quantification

Because human error quantification has been the main part of HRA for over four decades, a brief history of the development is given, before moving on to show how typical techniques work.

Evolution of Human Reliability Assessment.

HRA began in the 1960s at the same time that risk assessment practices were also developing, in the area of missile system development, most notably in various governmental sponsored laboratories (e.g., Sandia National Laboratories) in the United States. For the first two decades, however, HRA was more of a research area than a practical discipline. It “came of age”, however, shortly after the US nuclear power plant Three Mile Island accident happened in

1979. This accident, a partial melting of the nuclear reactor core, was seen as a major demonstration of the vulnerability of systems to human error, so an approach to predict and prevent such errors was urgently required. This led to the first practical HRA technique, the technique for human error rate prediction (THERP) [6], which is still in use today.

THERP. Application of THERP entails carrying out a task analysis of the human involvements in the system, defining human event sequences that need to be carried out to assure safety, and then quantifying these sequences using a database of human error probabilities. THERP uses a database of human error probabilities as its quantification “engine”. This database contains error probabilities for such tasks as reading a display, operating a joystick, responding to an alarm, etc. Some example THERP HEPs are shown in Table 7.

Table 7 shows that human reliability is poor in the first 20 min after the onset of a system abnormality or fault condition. This unreliability is also because nuclear power plants are complex, so some basic diagnosis must take place to determine the fault according to its annunciated symptoms.

THERP recommends consideration of the effects of a large number of factors that can influence human performance (time pressure being one, in the example in Table 7), called *performance shaping factors* (PSF). These factors can range from “external”

Table 7 HEPs for failing to diagnose an annunciated (alarmed) nuclear power abnormal event^(a)

Item	<i>T</i> (minutes after onset)	Median joint HEP for diagnosis of a single, or the first, event	Error factor
1	1	1.0	—
2	10	0.5	5
3	20	0.1	10
4	30	0.01	10
5	60	0.001	10
6	1 day	0.0001	30

^(a) Reproduced from [6]. US Nuclear Regulatory Commission, 1983

factors such as lighting, work hours, and manning parameters, to more “internal” factors such as motivation, experience, and stress. The “error factor” in Table 7 associated with each nominal HEP is then used as a modification factor, e.g., if there was a significant stress on the operators, an HEP might be multiplied by a factor of 10.

THERP was also the first technique (and one of the few) to explicitly consider *dependence*. Dependence is when the success or failure of one action is related to that of a previous action or task. For example, if an operator is calibrating a set of gas detectors and makes a mistake in one because he has the wrong value in mind, then he is likely to miscalibrate all of them. Another example of dependence is when someone is checking another’s work or results – there is usually a certain degree of trust so that the check may not be as thorough as one would hope for. In such cases, adding a third or even fourth “checker” into the equation will not necessarily increase the reliability either, due to a natural dependence.

THERP models dependence in five categories: complete, high, moderate, low, and zero dependence, and modifies the second (dependent) HEP accordingly using a specific formula. The analyst judges dependence according to whether the actions under consideration are being carried out by the same or different people or teams, whether they are using the same equipment in the same location, and whether they are in the same timeframe or not. A useful guide to dependence modeling and countermeasures is given in [17].

Expert judgment methods. Although THERP was immediately successful in the early 1980s, there were

a number of criticisms of the approach, in terms of its granularity of task description and associated human error probabilities. It was seen by some as too “decompositional” in nature. Additionally, its database origins have never been published, and there were concerns from the human factors community about its “broad brush” treatment of psychological and human factors aspects, as well as its significant resource requirements when it came to analyzing a complex system such as a nuclear power plant.

By 1984, there was therefore a swing to a new range of methods based on *expert judgment* (see **Expert Judgment; Experience Feedback; Uncertainty Analysis and Dependence Modeling**). Risk assessment at the time (specifically probabilistic safety assessment or PSA), whose champion was (and still is) the nuclear power industry, was used to applying formal expert judgment techniques to deal with highly unlikely events (e.g., earthquakes and other external events considered in PSAs). These techniques were therefore adapted to HRA. In particular two techniques emerged: *absolute probability judgment* [18], in which experts directly assessed HEPs on a logarithmic scale from 1.0 to 10^{-6} ; and the *success likelihood index method* (SLIM) [19]. The main difference with this technique was that it allowed detailed consideration of PSF in the HEP calculation process. Typical PSFs used by a number of HRA techniques are the following:

- unfamiliarity of the task (to the human operator),
- time pressure or stress,
- training adequacy,
- procedures adequacy,
- adequacy of the human machine interface,
- task complexity.

Expert judgment approaches are useful for considering difficult quantification assignments, for which perhaps there are no clear data available, or where stakeholders themselves want to be sure that the numbers are judged realistic by operational experts. In this sense, expert judgment techniques have some of the advantages of HAZOPs, since they elicit ownership and involvement from the participants, as well as maximizing expertise input by having operational experts involved. Conversely, if the operational experts are not very experienced themselves (a recommended minimum operational experience level is 10 years, and to still be an active operator) or if the system being assessed is way outside their own experience range, then there is the danger of gaining results that are not accurate or are invalid. There is also the danger of biases affecting the judgment process, whether these are memory-based (e.g., giving disproportionate weight to a rare event and ignoring other more representative ones), motivational (e.g., wanting to see a new system built, so underestimating the HEPs), or cognitive (e.g., finding it hard to deal with very low probabilities such as 1 error in 100 000).

HEART, NARA, and SPAR-H. One further technique worthy of note developed in the mid-1980s was the human error assessment and reduction technique (HEART) [20]. This technique had a much smaller and more generic database of HEPs than THERP, which made it more flexible and quicker to use, and had PSF called *error producing conditions* (EPCs), each of which had a maximum effect (e.g., from a factor of 19 to a factor of 1.2) on the HEPs. HEART was based on a review of the human performance literature (field studies and experiments), so the relative strengths of the different factors that can affect human performance had some credibility with the human factors and reliability engineering domains. HEART is still in use today in a number of industries as it is very resource efficient, and has recently been redeveloped as a specific nuclear power plant HRA technique called *nuclear action reliability assessment* (NARA) [21]. As an example of how these techniques work, Tables 8 and 9 show the generic task types (GTTs) for NARA and the EPCs.

The HRA analyst assessing a nuclear power plant therefore, having identified a critical human action, e.g., responding to an alarm, must use Table 8 to pick the appropriate (best fit) GTT. Assuming it is

not a complex pattern, C1 would be selected, with a “nominal” HEP of 0.0004 or 4 errors per 10 000 opportunities (incidentally, this particular GTT is based on observed error rates with experienced UK nuclear power plant operators). The assessor must then (on the basis of prior task analysis and interviews with existing operators or prospective ones if the system is not yet built) consider if there are any human factors that could increase the likelihood of human error. As with most HRA techniques, NARA preidentifies the human factors that are most important for HRA consideration, along with “weighting factors” to determine how large the impact on the HEP will be.

Let us assume that in this case the assessor finds that the human machine interface for displaying alarms is not as good as it should be, and that the alarm condition is rare and will be relatively unfamiliar to the operator. The assessor would then select EPCs 2 (unfamiliarity) and 11 (poor or ill-matched system feedback), with maximum effects on the HEP of 20 and 4, respectively. The assessor then has to make a judgment on how unfamiliar the condition is and how poor the interface is, and this requires assessor training and experience with the technique itself (in NARA descriptive “anchor points” for each EPC also assist the assessor). Let us suppose that the assessor decides in each case to use half of the maximum effect. This is called (in HEART as well as NARA) the *assessed proportion of affect* or *APOA*. The values chosen are then substituted into the following generic NARA quantification equation:

$$HEP = GTT \times \{[(EPC_1 - 1) \times APOA_1 + 1] \times \dots \times [(EPC_n - 1) \times APOA_n + 1]\} \quad (2)$$

Inputting the chosen GTT and EPC information gives the following result:

$$\begin{aligned} HEP &= 0.0004 \times \{[(20 - 1) \times 0.5 + 1] \\ &\quad \times [(4 - 1) \times 0.5 + 1]\} \\ &= 0.0004 \times [9.5 + 1] \times [1.5 + 1] \\ &= 0.0105 \quad \text{or} \quad 0.011 \end{aligned} \quad (3)$$

Thus the final HEP is approximately 1 in 100 opportunities or system demands. This value would then be put into the appropriate fault or event tree. NARA also assesses dependence, using a version of the original THERP dependence model. The form

Table 8 NARA generic task type HEPs

NARA GTT ID	NARA GTT description	NARA GTT HEP
<i>A : Task execution</i>		
A1	Carry out simple single manual action with feedback. Skill based and therefore not necessarily with procedure.	0.005
A2	Start or reconfigure a system from the main control room following procedures, with feedback.	0.001
A3	Start or reconfigure a system from a local control panel following procedures, with feedback.	0.002
A4	Judgment needed for appropriate procedure to be followed, based on interpretation of alarms/indications and situation covered by training at appropriate intervals.	0.006
<i>B : Ensuring correct plant status and availability of plant resources</i>		
B1	Routine check of plant status.	0.02
B2	Restore a single train of a system to correct operational status after test following procedures.	0.004
B3	Set system status as part of routine operations using strict administratively controlled procedures.	0.0007
B4	Calibrate plant equipment using procedures, e.g., adjust set-point.	0.003
B5	Carry out analysis.	0.03
<i>C : Alarm/indication response</i>		
C1	Simple response to a key alarm within a range of alarms/indications providing clear indication of situation (simple diagnosis required). Response might be direct execution of simple actions or initiating other actions separately assessed.	0.0004
C2	Identification of situation requiring interpretation of complex pattern of alarms/indications. (Note that the response component should be evaluated as a separate GTT.)	0.2
<i>D : Communication</i>		
D1	Verbal communication of safety critical data.	0.006

Table 9 NARA error producing conditions

NARA EPC	NARA error producing condition (EPC) description	NARA EPC maximum effect
1	A need to unlearn a technique and apply one which requires the application of an opposing philosophy.	24
2	Unfamiliarity, i.e., a potentially important situation which only occurs infrequently or is novel.	20
3	Time pressure.	11
4	Low signal to noise ratio.	10
5	Difficulties caused by poor shift handover practices and/or team coordination problems or friction between team members.	10
6	A means of suppressing or overriding information or features which is too easily accessible.	9
7	No obvious means of reversing an unintended action.	9
8	Poor environment (lighting, heating, ventilation, physical access problems, e.g., rubble after explosion).	8
9	Operator inexperience.	8
10	Information overload, particularly one caused by simultaneous presentation of nonredundant information.	6
11	Poor, ambiguous, or ill-matched system feedback.	4
12	Shortfalls in the quality of information conveyed by procedures.	3
13	Operator underload/boredom.	3
14	A conflict between immediate and long-term objectives.	2.5
15	An incentive to use other more dangerous procedures.	2
16	No obvious way of keeping track of progress during an activity.	2
17	High emotional stress and effects of ill health.	2
18	Low workforce morale or adverse organizational environment.	2

of equation (2) prevents inadvertent reduction of the HEP by selecting a low APOA with a low maximum effect EPC. For example, if an APOA of 0.1 was selected for EPC 11, with the same GTT (C1), this would lead to a lower HEP than the original GTT:

$$HEP : 0.0004 \times 0.1 \times 2 = 0.00008 \quad (4)$$

Clearly this is not the intended outcome, so equation (2) avoids this problem. One other point to mention is that in some cases the multiplication factors push the HEP beyond the bounds of probability, e.g., producing an outcome value greater than unity (1.0), in which case the analyst must simply truncate the value at 1.0. As mentioned already, any HRA technique requires training and experience, and data sets (scenarios, factors, and actual – known – HEPs) are available with which HRA analysts can train themselves and “calibrate” their approach.

The example also highlights another useful function of this type of analysis. Of the two EPCs, which is most important if error reduction is required? In this particular example, the problem of unfamiliarity clearly dominates the human contribution to risk, so if error reduction is required, this would be the first port of call. This is a useful function, and HRA can help human factors prioritize issues – it may be far more effective and cheaper for example, to implement a refresher training regimen than provide an upgrade to the whole alarm system (HEART and SLIM also contain this function). NARA has its own error reduction guidance, and an extract for EPC 11 is given in Table 10.

A similarly flexible and quick HRA technique developed in the United States and applied on over 70 HRAs to date is the standardized plant analysis risk model human reliability assessment (SPAR-H) technique [22]. This uses an HEP calculation process

and a small number of PSF with associated quantitative multipliers to quantify HEPs for diagnosis and subsequent actions in nuclear power plants. It also contains guidance for dependence and uncertainty calculations.

Human error data. Human error probabilities would ideally be based on empirical data and stored in databases much as for reliability data in some industries. In fact, although there have been numerous human error databases over the years [23, 24], few have survived. One current human error database that is still growing is the computerized operator reliability and error database (CORE-DATA, [25]). This database has some high-quality data from nine separate industrial sectors, and has been the basis for development of NARA and a new approach for HRA in air traffic management (controller action reliability assessment, CARA [26]). Some example data from CORE-DATA is shown in Table 11.

As a general rule, for making an initial screening of HEPs, any task which is complex or time pressured may be considered as a first approximation to be in the 0.1–1.0 probability region. Routine tasks may be considered to be in the 0.01–0.1 region, and well-designed procedural tasks with good training, human interface, and checking can be considered to be in the 0.001–0.01 region. Such values should not themselves be used in an actual HRA, but are “rules of thumb” applied by assessors when first considering a range of tasks that may require quantification.

Human performance limiting values. Dependence has already been introduced *via* the THERP approach, which considers dependence between subsequent tasks. However, it is possible that a risk assessment failure path may contain several errors,

Table 10 Extract of NARA error reduction guidance for EPC 11: poor, ambiguous, or ill-matched system feedback

This EPC is clearly concerned with design aspects, so design is the optimum solution (e.g., using USNRC Doc NUREG 0700 or other ergonomic design guidance/standards). In particular, use unique coding of displays, and enhance discriminability of controls and displays *via* coding mechanisms (including color coding, grouping, etc.). For control panels, sometimes significant improvement can be made by enhanced labeling (e.g., hierarchical labeling schemes) and use of demarcation lines that highlight the functional groupings and relations between controls and control sequences. In some cases, however, the solution will entail either additional indications, including perhaps higher reliability, diversely measured, or more scenario-specific indications or alarms, or else relaying signals from external locations back to the main control room, to facilitate feedback in a centralized location. Procedures can help the operators focus on the best indicators and sets of indicators, and at least be aware of potential ambiguity problems. Training in a simulator with degraded indications can help prepare the operators for such scenarios (which are actually not uncommon in real events).

Table 11 Examples of HEPs from the CORE-DATA database^(a)

Study	Description of error	HEP	Data source
Nuclear (confidential)	During a shift the transport department brought a chemical load to the compound after permission had been arranged between two supervisors, but the correct paperwork did not arrive with the chemicals. Consequently this led to two cans of highly enriched chemical solution being processed instead of six cans of low enriched chemical.	0.0027	Real data
Manufacturing (confidential) Chemical [28]	A component has a different profile machined on each end. The operator inadvertently machines the aft end profile on the forward end. In a simulated chemical process plant, a trainee in a diagnostic task is faced with a panel of 33 instruments eight of which they were familiar with, eight they had never seen before. They had no feedback and had a limit of 5 min per problem. The error under consideration is incorrect diagnosis, including failure to diagnose.	0.00039 0.34	Hard copy data Simulator
Petrochemical [11]	Before the lifeboat is lowered to sea level, one of the internal checks that the coxswain must perform involves checking the air support system. This check was omitted. If the system is not checked and the crew find out after they have abandoned, there is the possibility of ingress of smoke. Lack of air will cause asphyxiation.	0.037	Simulator
Calculator errors [16]	An experiment is performed on a group of engineers selected from a large industry in America to investigate their performance in using electronic calculators. Each engineer is given a test and is required to calculate answers to numerical problems. The test consists of 10 problems ranging from very simple ones to more complex ones. Each problem is selected from the manual supplied with the calculator. The error of concern is that the engineer will make an error in the block of 10 problems.	0.27	Experimental data
Aviation [17]	A new system for data transfer and display for use by air traffic controllers is being evaluated in a simulation study. The simulation is based on normal flight operations at Heathrow Airport. The new system displays information to the controller on a single cathode ray tube which features a touch sensitive overlay on the display through which data is entered and controlled. The controller performs his normal duties, using the new system to transmit, obtain, and update information. There are three possible errors the controller can commit: miss touches, if the controller misses the area which should be touched; error touches, when the controller touches the wrong area; and illegal touches, when the controller touches a label out of sequence. In this scenario the air traffic controller makes a miss error.	0.064	Simulator data
Nuclear [18]	A locally operated valve has a rising stem and position indicator. An auxiliary operator, while using written procedures to check a valve line-up, fails to realize that the valve is not in the right position, after a maintenance person has performed a procedure intended to restore it to its proper position.	0.003	Expert judgment

^(a) Reproduced from [25]. © IEEE, 1997

all of which must happen, e.g., human error 1 (an initiating event error) AND human error 2 (a failure by an operator to detect the event) AND human error 3 (a failure to diagnose the event) AND human error 4 (a failure to recover the event). Even using dependence modeling as used in THERP, it is sometimes the case that such sequences (called *cutsets* in risk analysis terms) end up with a combined HEP that is unrealistically low (e.g., 0.000001). Incident and accident experience shows us in fact that although we try hard to model all failures, there are often surprises in store for us – things happen which were not predicted (e.g., an unpredicted EOC) or for reasons that were assumed to be implausible. This is also the case in hardware reliability and risk assessment, where in that domain cutsets are prevented from being optimistic by using limiting values or other methods (e.g., β factors). The same approach is necessary in HRA, so human performance limiting values (HPLVs) were developed [27]. Effectively, the general rule is not to allow the human error component of a cutset to go below 0.00001 or 10^{-5} , unless robust human error data exist to substantiate such low values or unless there are compelling qualitative arguments.

Validity of the HRA quantification approaches. Toward the end of the 1980s there were a number of techniques (around a dozen) in use in nuclear, process and chemical, and offshore oil and gas sectors of industry. In an effort to clarify the situation, a series of validations were carried out [29, 30]. In such exercises, HRA assessors were given task descriptions and asked to use their techniques to predict results for which the validators had data (and such data were not known by the assessors). Where prediction accuracy fell below a reasonable proportion (e.g., most HEPs needed to be predicted within a factor of 10, and preferably a factor of 3), the technique was invalidated. In such exercises, for example, THERP, absolute probability judgment (APJ), and HEART received positive validations, whereas SLIM's predictive accuracy depended critically on certain factors (notably "calibration" of the expert judgments) and was less assured. One approach in particular, which predicted human performance effectively solely as a function of time available to carry out the task (called *time response curve*) was invalidated [31, 32].

To show how validation works, the results of a large-scale UK validation [33] are summarized here,

which focused on the three quantification techniques THERP, HEART, and JHEDI. Totally 30 UK HRA practitioners took part in the exercise and each assessor independently used only one of the three techniques (very few assessors were practitioners in more than one technique). Each assessor quantified HEPs for 30 tasks, and for each task there were the following information:

- general description of the scenario;
- inclusion of relevant PSF information in the description;
- provision of simple linear task analysis;
- provision of diagrams where necessary and relevant;
- statement of exact human error requiring quantification.

Each assessor had 2 days to carry out the assessments, and experimental controls were exercised, so that the assessors were working effectively under invigilated examination conditions. For each of the 30 tasks the HEP was known to the experimenter, but not to any of the assessors. Tasks were chosen to be relevant to nuclear power and reprocessing industries, since all of the assessors belonged to these two areas. A large proportion of the data were from real recorded incidents, with the data range spanning five orders of magnitude (i.e., from 1.0 to $1\text{E}-5$).

The analysis of all data (i.e., all 895 estimated HEPs – there were five missing values) showed a significant correlation between estimates and their corresponding true values (Kendall's coefficient of concordance: $Z = 11.807$, $p < 0.01$). This supported the validity of the HRA quantification approach as a whole, especially as no assessors or outliers were excluded from these results. The analysis of all data for individual techniques shows a significant correlation in each case (using Kendall's coefficient of concordance): THERP, $Z = 6.86$; HEART, $Z = 6.29$; JHEDI, $Z = 8.14$; all significant at $p < 0.01$.

Individual correlations for all subjects are shown in Table 12. There are 23 significant correlations (some significant at $p < 0.01$ level) out of a possible 30 correlations. This is a very positive result, again supporting the validity of the HRA quantification approach.

Table 13 shows that there is an overall average of 72% precision (estimates within a factor of 10) for all assessors, irrespective of whether they were significantly correlated or not. This figure includes

Table 12 Correlations for each subject for the three techniques

	THERP	HEART	JHEDI
1	0.615**	0.577**	0.633**
2	0.581**	0.558**	0.551**
3	0.540**	0.473**	0.533**
4	0.521**	0.440**	0.452**
5	0.437**	0.389*	0.436**
6	0.311*	0.370*	0.423*
7	0.298	0.351*	0.418*
8	0.297	0.347*	0.401*
9	0.254	0.217	0.386*
10	0.078	0.124	0.275

* Significant at $p < 0.05$ ** Significant at $p < 0.01$

all data estimates, even the apparent outliers that have been identified in the study. This is therefore a reasonably good result, supporting HRA quantification as a whole. Furthermore, no single assessor dropped below 60% precision in the study. The precision within a factor of 3 is approximately 38% for all techniques. This is a fairly high percentage, given the required precision level of a factor of 3. The degree of optimism and pessimism is not too large, as also shown in the histogram in Figure 4, with only a small percentage of estimates at the extreme optimistic and pessimistic ends of the histogram (i.e., greater than a factor of 100 from the actual estimate). Certainly there is room for improvement, but the optimism and pessimism are not in themselves dominating the results, and estimates were more likely to be pessimistic (17.5% of the total number of estimates) than optimistic (9.7% of the total number of estimates), which means that assessors erred on the side of caution. The highest and lowest precision values for the techniques within a factor of 3 and within a factor of 10 are shown in Table 14.

The overall results were therefore positive with a significant overall correlation of all estimates with the known true values, with 23 individual (assessor)

significant correlations and a general acceptable precision range of 60–87%, the average precision being 72%. These results lend support for the HRA quantification part of the HRA process in general.

Second generation HRA. In the early 1990s a prominent HRA expert [34] suggested, on the basis of experiences of the accidents mentioned above, that most HRA approaches did not pay enough attention to context, i.e., to the detailed scenarios people found themselves in. Essentially, the argument was that considering human reliability in an abstracted format of fault and event trees was insufficient to capture the local situational factors that would actually dictate human behavior and lead to success or failure. This occurred at the same time as a growing concern in the United States in particular with EOC, wherein a human in the system does something that is erroneous and not required by procedures (e.g., shutting off emergency feed water to the reactor in Three Mile Island and disconnecting safety systems while running reactivity experiments in Chernobyl). Such errors of intention, relating to a misconception about the situation, are, as already noted earlier, severely hazardous to industries such as nuclear power and aviation. Incident experience in the United States was suggesting that it was these types of rare errors, whose very unpredictability made them difficult to defend against, which were of most real concern.

Therefore, work on a set of so-called second generation HRA techniques began in the early mid-1990s. The most notable of these were a technique for human error analysis (ATHEANA [35]), the cognitive reliability error analysis method (CREAM [36, 37]), MERMOS [38], and the connectionist approach to human reliability (CAHR) [39]. ATHEANA is notable because it has had more investment than almost any other HRA technique. The qualitative part of the method is intensive and is used for identification of safety critical human interventions in several instances, but its quantitative use has so far been marginal due to residual problems over

Table 13 Precision for the three techniques for all HRA assessors

	Factor 3 (%)	Factor 10 (%)	Optimistic (%)	Pessimistic (%)
THERP	38.33	72	13.67	13.33
HEART	32.67	70.33	11	18.33
JHEDI	43.67	75	3.33	21.33

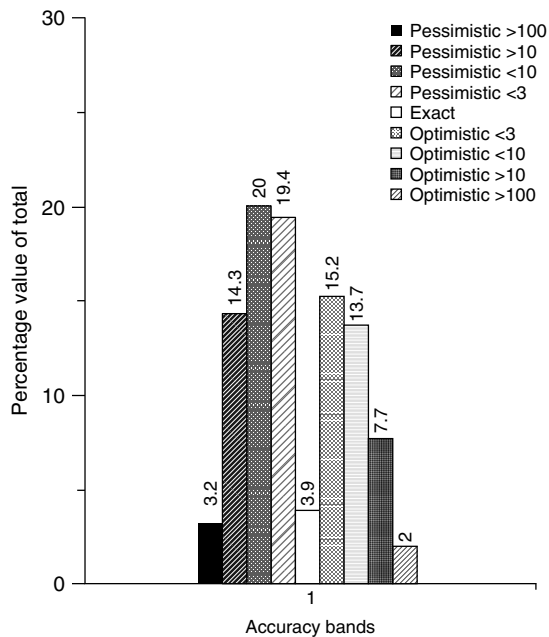


Figure 4 Overall HRA optimism and pessimism

quantification. CREAM is a more straightforward quantification technique that has had some applications in the nuclear power domain. MERMOS is used only in France and focuses on the conditions that need to coincide to give rise to errors of commission. CAHR is similar to JHEDI in that it uses real incident data calibrated to generate HEPs, and has been used in HRAs/PSAs in Germany in several industries (JHEDI is in use in only one organization in the United Kingdom). The term second generation really refers to two aspects: the first is a more profound analysis of contextual factors (including interacting PSF) to ensure the accuracy and validity of the resultant HEPs (CREAM and CAHR, and to an extent JHEDI) and the second is the ability to predict errors of commission (ATHEANA) or the conditions that lead to them (MERMOS).

A recent international HRA expert workshop (Halden Reactor Project Workshop on HRA, October 2005, Halden, Norway) provided an overview of the current level of implementation of second generation approaches. Though the methods mentioned above were all applied already in various safety assessments, there is still more work to be done regarding the evidence of empirical validation compared to “first generation” techniques (like HEART or THERP). The level of application and work to be done can also be seen from a special issue of a key reliability journal on HRA and EOCs [40].

The second generation HRA approach is clearly still evolving. Taken to one extreme, there is less focus on individual errors and more focus on determining what factors can combine to lead to an intrinsically unsafe situation, wherein errors of intention become increasingly likely. This concept has some resonance not only with accidents such as the nuclear accident at Chernobyl or the chemical accident in Bhopal, India, but also with the Überlingen midair collision in 2002 in Germany. In the latter tragic event, although there were discrete air traffic controller and pilot errors that were arguably predictable by first generation HRA methods, there was nevertheless an aggregation of an unusually large amount of factors that predisposed the whole situation to failure. This “confluence” of negative factors is now also being looked at not only by HRA, but also by proponents of the new field of “resilience engineering” (see **Managing Infrastructure Reliability, Safety, and Security**) [41].

This does not mean that second generation HRA research should stop, nor does it mean that HRA is useless without this function of predicting errors of commission – HRA can still identify and predict most types of errors and occasionally errors of commission too. What it does mean for system developers and risk managers is that other methods and measures must be utilized, adopting a “defense in depth” strategy for managing human error, including maintaining

Table 14 Precision for the three techniques

	Highest precision		Lowest precision	
	Factor of 3 (%)	Factor of 10 (%)	Factor of 3 (%)	Factor of 10 (%)
THERP	56.67	80	10	63.33
HEART	46.67	76.67	16.67	60
JHEDI	66.67	86.67	26.67	70

a good level of human factors adequacy in the operator interfaces, environment, and support systems (training, procedures, work organization, etc.), as well as a good safety culture that will avoid or reduce violations and help detect the seeds of errors of commission long before they manifest themselves in accidents.

In practical terms, first generation techniques are still useful. Rather than second generation techniques being seen as an attempt to replace first generation HRA tools, a more likely future relationship is one of complementarity. For example, SPAR-H and ATHEANA could be an effective partnership for the US nuclear industry, and the European air traffic industry is similarly considering a similar dual approach involving CARA and an ATM-version of CAHR. Such partnerships may offer a more flexible and resource-efficient approach to HRA.

Safety culture. *Safety culture* concerns the degree to which the culture of the installation's organization puts safety and correct performance as the overriding priority. It came into being as an area of study and improvement following the 1986 Chernobyl disaster [42]. If everyone in the organization is personally concerned with safety, and mindful of error, and willing to question practices if they believe them to be unsafe, then clearly this has an impact on error reduction, error recovery enhancement, violation avoidance, as well as watching out for errors of commission.

Violations, in particular, are inextricably bound together with the concept of safety culture. An installation with a good safety culture should not exhibit rule violations, whereas one with a poor safety culture will exhibit routine and even perhaps, under certain circumstances, extreme violations. The concern with rule violations is similar to that for mistakes – violations can lead to failure of multiple safety systems and barriers. As an example, a risk assessment may assume that a fully programmed machine with multiple interlocks will not be used in “maintenance mode” (where the safety interlocks are disabled) for normal operations. Such a “routine” violation would not only violate the procedures of the installation, but would also seriously violate the assumptions of the risk analysis, such that its predictions of failure for such a system could easily be incorrect by one or more orders of magnitude.

When carrying out HRA and risk assessment in general, there are usually general assumptions made about the level of training and adherence to procedures. If there is a poor safety culture, the working assumptions of the risk assessment, at a certain point, will become invalid. This is to an extent what happened with both Bhopal and Chernobyl. No risk assessor would have assumed that so many safety systems would have been switched off or disabled.

There is a link, therefore between HRA (and risk assessment) and safety culture – ideally safety culture should be measured periodically (e.g., every 2 year) by an organization to verify that the “fundamentals” of adequate behavior and safe attitudes at work, which many would consider simply professional conduct, are being maintained.

HRA, Ethics, and “Just” Culture

An important standpoint of HRA is that it is not allocating blame – almost always, accidents and incidents and human errors are the result of system design and operational constraints, complexities, and pressures. Most errors are unintentional, and in the majority of tragic accidents, for example in aviation or rail, the human operator at the “sharp end” is not able to defend him or herself after the event. HRA therefore, as with risk assessment, aims to be dispassionate – the search is always for errors and their consequences, their causes, their relative importance, and effective (and sustainable) remedies. Even when an intentional error occurs or is predicted (e.g., a violation of procedures), the assessor is looking for deeper causes in the organizational culture (called *safety culture*). Only when malice is intended, and this is exceptionally rare in work-related systems, the real connotation of blame would come into the equation, and at such a point this would be out of the hands of HRA and into criminal investigation. This standpoint of HRA practitioners is more relevant than for risk analysts, since the former will be interviewing operators and management from the system under investigation (whether an existing system or one that is on the drawing board). HRA analysts are predicting how a workforce, or a pilot, or a train driver, or a surgeon could make serious errors, possibly leading to deaths – this naturally has to be handled sensitively. The HRA analyst must convey (and believe) that (s)he is there to support these operators

by finding system vulnerabilities and helping them and the parent organization overcome them.

This point is practical as well as ethical. If people think they will be prosecuted for “honest mistakes” then they will not tend to report them. This means that the organization and the wider community will not be able to learn from such errors or even discover them until a large-scale accident occurs. This is usually far too late to learn, in terms of the losses that have already occurred (e.g., fatalities and loss of equipment etc.), as well as the financial and public confidence losses that will follow for the organization and perhaps even the industry concerned, in the wake of large-scale accidents. HRA therefore works within what is called a *just culture* framework.

As a closing point, it has to be said that most of the time people “get it right” and accidents are rare. In general, systems are very safe and productive, and this is not possible without the human operator. Human reliability, particularly in mature systems where the right balance between humans and engineered safety functions has been found, is often exceptionally good. HRA can therefore also be seen as an approach to “continuous improvement”, ensuring that the human operators (and their managers) are able to be both safe and effective – the two usually go hand in hand.

Conclusion

HRA started in the 1960s and has grown to be a useful tool in many industries where the human operators play a critical role. It is now used qualitatively and/or quantitatively in nuclear, chemical, offshore oil and gas, defense, aviation and air traffic, space, and medical sectors. This is a testament to the success of the overall approach.

HRA has been used in many assessments to decide if the design is safe or reliable enough, and to determine how to improve the human reliability in the system’s performance. Although HRA’s evolution is far from over, there is a useful methodology and toolkit available for any industry wishing to focus on supporting, protecting, and enhancing its major asset – the human in the system.

References

- [1] Kirwan, B. (2006). Safety informing design, *Safety Science* **45**, 155–197.
- [2] Kirwan, B. & Ainsworth, L.A. (eds) (1992). *A Guide to Task Analysis*, Taylor & Francis, London.
- [3] Wilson, J.R. & Corlett, N.E. (2005). *Evaluation of Human Work*, Taylor & Francis, London.
- [4] Cox, S.J. & Tait, N.R.S. (1991). *Reliability, Safety and Risk Management*, Butterworth-Heinemann, Oxford.
- [5] Kirwan, B. (1994). *A Guide to Practical Human Reliability Assessment*, Taylor & Francis, London.
- [6] Swain, A.D. & Guttman, H.E. (1983). *A Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications*, USNRC-Nureg/CR-1278, Washington, DC.
- [7] Embrey, D.E., Kontogiannis, T. & Green, M. (1994). *Preventing Human Error in Process Safety*, Centre for Chemical Process Safety (CCPS), American Institute of Chemical Engineers, New York.
- [8] US Nuclear Regulatory Commission (2005). *Good Practices for Implementing Human Reliability Analysis*, NUREG-1792, Washington, DC.
- [9] Swain, A.D. (1989). *Comparative Evaluation of Methods for Human Reliability Analysis*, GRS-81, Gesellschaft für Reaktorsicherheit (GRS)mbH, Schwertnergasse 1, 5000 Köln 1, Germany.
- [10] Reason, J.T. (1990). *Human Error*, Cambridge University Press, Cambridge.
- [11] Reason, J.T. (1998). *Managing the Risks of Organisational Accidents*, Ashgate, Aldershot.
- [12] Kirwan, B. (1998). Human error identification techniques for risk assessment of high risk systems – Part 1: review and evaluation of techniques, *Applied Ergonomics* **29**(3), 157–177.
- [13] Embrey, D.E. (1986). SHERPA – a systematic human error reduction and prediction approach, Paper presented at the *International Topical Meeting on Advances in Human Factors in Nuclear Power Systems*, Knoxville.
- [14] Shorrock, S.T. & Kirwan, B. (2002). Development and application of a human error identification tool for air traffic control, *Applied Ergonomics* **33**, 319–336.
- [15] Kirwan, B. & Kennedy, R. (2001). Assessing safety & usability of air traffic management (ATM) systems using a HAZOP approach, *Human Error and Safety System Development (HESSD 2001)*, Linköping, June 11–12.
- [16] Kirwan, B. (1997). The development of a nuclear chemical plant human reliability management approach, *Reliability Engineering and System Safety* **56**, 107–133.
- [17] Greenstreet Berman (2001). Preventing the propagation of error and misplaced reliance on faulty systems: a guide to human error dependency, UK Health & Safety Executive, Offshore Technology Report 2001/053, HMSO, London.
- [18] Seaver, D.A. & Stillwell, W.G. (1983). *Procedures for Using Expert Judgement to Estimate Human Error Probabilities in Nuclear Power Plant Operations*, NUREG/CR-2743, Washington, DC.
- [19] Embrey, D.E., Humphreys, P.C., Rosa, E.A., Kirwan, B. & Rea, K. (1984). *SLIM-MAUD: An Approach*

- to Assessing Human Error Probabilities Using Structured Expert Judgement, NUREG/CR-3518, US Nuclear Regulatory Commission, Washington, DC.
- [20] Williams, J.C. (1988). A data-based method for assessing and reducing human error to improve operational performance, *IEEE Conference on Human Factors in Power Plants*, Monterey, pp. 436–450, June 5–9.
- [21] Kirwan, B., Gibson, H., Kennedy, R., Edmunds, J., Cooksley, G. & Umbers, I. (2004). Nuclear action reliability assessment (NARA): a data-based HRA tool, in *Probabilistic Safety Assessment and Management 2004*, C. Spitzer, U. Schmocker & V.N. Dang, eds, Springer-Verlag, London, pp. 1206–1211.
- [22] Gertman, D., Blackman, H.S., Marble, J., Byers, J., Haney, L.N. & Smith, C. (2005). The SPAR-H human reliability analysis method, US Nuclear Regulatory Commission Report NUREG/CR-6883, Washington, DC.
- [23] Topmiller, D.A., Eckel, J.S. & Kozinsky, E.J. (1984). Human reliability databank for nuclear power plant operations, USNRC Report NUREG/CR-2744, Washington, DC.
- [24] Gertman, D.I. & Blackman, H. (1994). *Human Reliability and Safety Analysis Data Handbook*, John Wiley & Sons, Chichester.
- [25] Kirwan, B., Basra, G. & Taylor-Adams, S. (1997). CORE-DATA: a computerised human error database for human reliability support, *IEEE Conference on Human Factors and Power Plants*, Orlando.
- [26] Kirwan, B. & Gibson, W.H. (2007). CARA: a human reliability assessment tool for air traffic management – technical basis and preliminary architecture, in *The Safety of Systems*, F. Redmill & T. Anderson, eds, Springer-Verlag, London, pp. 163–178.
- [27] Kirwan, B., Martin, B.R., Rycraft, H. & Smith, A. (1990). Human error data collection and data generation, *International Journal of Quality and Reliability Management* 7(4), 34–66.
- [28] Kirwan, B. (1992). Human error identification in HRA. Part 2: detailed comparison of techniques, *Applied Ergonomics* 23(6), 371–381.
- [29] Kirwan, B. (1997). Validation of human reliability assessment techniques: part 1 – validation issues, *Safety Science* 27(1), 25–41.
- [30] Kirwan, B. (1997). Validation of human reliability assessment techniques: part 2 – validation results, *Safety Science* 27(1), 43–75.
- [31] Dolby, A.J. (1990). A comparison of operator response times predicted by the HCR model with those obtained from simulators, *International Journal of Quality and Reliability Management* 7(5), 19–26.
- [32] Kantowitz, B. & Fujita, Y. (1990). Cognitive theory, identifiability, and human reliability analysis, *Reliability Engineering and System Safety* 29(3), 317–328.
- [33] Kirwan, B. (1999). Some developments in human reliability assessment, in *The Occupational Ergonomics Handbook*, W. Karwowski & W.S. Marras, eds, CRC-Press, London, pp. 643–666.
- [34] Dougherty, E.M. (1990). Human reliability analysis – where shouldst thou turn?, *Reliability Engineering and System Safety* 29, 283–299.
- [35] Cooper, S.E., Ramey-Smith, A.M., Wreathall, J., Parry, G.W., Bley, D.C., Luckas, W.J., Taylor, J.H. & Barriere, M.T. (1996). *A technique for Human Error Analysis (ATHEANA) – Technical Basis and Method Description*, Nureg/CR-6350, US Nuclear Regulatory Commission, Washington, DC.
- [36] Hollnagel, E. (1993). *Human Reliability Analysis: Context and Control*, Academic Press, London.
- [37] Hollnagel, E. (1998). *CREAM: Cognitive Reliability Error Analysis Method*, Elsevier Science, Oxford.
- [38] Le Bot, P. (2003). Methodological validation of MER-MOS by 160 analyses, *Proceedings of the International Workshop Building the New HRA: Errors of Commission from Research to Application*, NEA/CSNI/R(2002)3. OECD, Issy-les-Moulineaux.
- [39] Sträter, O. (2000). *Evaluation of Human Reliability on the Basis of Operational Experience*, ISBN 3-931995-37-2, GRS-170, GRS, Köln.
- [40] Sträter, O. (ed) (2004). Human reliability analysis: data issues and errors of commission, *Special Issue of Reliability Engineering and System Safety* 83(2).
- [41] Kirwan, B., Gibson, W.H. & Hickling, B. (2008). The feasibility of human reliability assessment in air traffic management, *Reliability Engineering & System Safety* 93, 217–233.
- [42] Beswick, L. & Kettleborough, J. (2007). A proactive approach to enhancing safety culture, in *The Safety of Systems*, F. Redmill & T. Anderson, eds, Springer-Verlag, London, pp. 105–116.

BARRY KIRWAN

Imprecise Reliability

Most methods in reliability and quantitative risk assessment assume that uncertainty is quantified *via* precise probabilities, all perfectly known or determinable. For example, for system reliability, complete probabilistic information about the system structure is usually assumed, including dependence of components and subsystems. Such detailed information is often not available, owing to limited time or money for analyses or limited knowledge about a system. In recent decades, several alternative methods for uncertainty quantification have been proposed, some also for reliability. For example, fuzzy reliability theory (see **Near-Miss Management: A Participative Approach to Improving System Reliability**) [1] and possibility theory [2] provide solutions to problems that cannot be solved satisfactorily with precise probabilities. We do not discuss such methods, but restrict attention to generalized uncertainty quantification *via* upper and lower probabilities, also known as *imprecise probability* [3] (see **Ecological Risk Assessment**) or “interval probability” [4, 5]. During the last decade, upper and lower probability theory has received increasing attention, and interesting applications have been reported. See [6] for a detailed overview of imprecise reliability and many references. It is widely accepted that, by generalizing precise probability theory in a mathematically sound manner, with clear axioms and interpretations, this theory provides an attractive approach to generalized uncertainty quantification.

In classical theory, a single probability $P(A) \in [0, 1]$ is used to quantify uncertainty about event A . For statistics, probability requires an interpretation, the most common ones are in terms of “relative frequencies” or “subjective fair prices for bets”. The theory of lower and upper probabilities [3–5] generalizes probability by using lower probability $\underline{P}(A)$ and upper probability $\overline{P}(A)$ such that $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$. The classical situation, so-called precise probability, occurs if $\underline{P}(A) = \overline{P}(A)$, whereas $\underline{P}(A) = 0$ and $\overline{P}(A) = 1$ represents complete lack of knowledge about A . This generalization allows indeterminacy about A to be taken into account. Lower and upper probabilities can be interpreted in several ways. One can consider them as bounds for a precise probability, related to relative frequency of the event A , reflecting the limited information one has

about A . Alternatively, from subjective perspective, the lower (upper) probability can be interpreted as the maximum (minimum) price for which one would actually wish to buy (sell) the bet that pays 1 if A occurs and 0 if not. Generally, $\underline{P}(A)$ reflects the information and beliefs in favor of event A , while $\overline{P}(A)$ reflects such information and beliefs against A . Walley [3], from the subjective point of view, uses coherence arguments to develop theory for lower and upper probabilities and related statistical methods. His theory generalizes the Bayesian statistical theory (see **Common Cause Failure Modeling; Bayesian Statistics in Quantitative Risk Assessment; Group Decision; Subjective Probability**) for precise probabilities in a manner similar to “robust Bayesian methods” [7], where sets of prior distributions are used and data is taken into account *via* a generalized version of Bayes’ theorem. Important properties are that lower (upper) probability is superadditive (subadditive), i.e., $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B)$ ($\overline{P}(A \cup B) \leq \overline{P}(A) + \overline{P}(B)$) for disjoint A and B , and $\underline{P}(A) = 1 - \overline{P}(\bar{A})$, with $\bar{A}$ the complementary event to A [3–5]. An advantage of lower and upper probability is that one only requires limited input with regard to the uncertainties about the events of interest, and one can (in principle) always derive corresponding lower and upper probabilities for all events, *via* “natural extension” [3], if the inputs are not contradictory. This is attractive, as one normally can only meaningfully assess a limited number of characteristics of random quantities. Computation of the lower and upper probabilities according to the natural extension might be complicated, as constrained optimization problems must be solved. We discuss this topic in the section titled “Imprecise Reliability *via* Natural Extension”.

Walley [3] discussed many reasons why precise probability is too restrictive for practical uncertainty quantification. In reliability, the most important ones include limited knowledge and information about random quantities of interest, and possibly information from several sources that might appear to be conflicting if restricted to precise probabilities. Common aspects such as grouped data and censoring can be dealt with naturally *via* imprecision, and imprecise reliability offers attractive methods for inference in case data contain zero failures.

In recent years, there has been increasing research into statistical methods using lower and upper probabilities. Following Walley [3], most of this work has been on robust-Bayes-like inferences, where sets of

prior distributions are used for a parametric model. In particular, Walley's Imprecise Dirichlet Model (IDM) [8] for multinomial inferences is proving popular, though it is not undisputed (see the discussion in [8]). Some results on the application of the IDM to reliability problems are discussed in the section titled "Statistical Inference in Imprecise Reliability". An alternative statistical approach has been developed by Coolen, together with several coauthors [9–11]. It is called *nonparametric predictive inference* (NPI), and is explicitly aimed at only making limited modeling assumptions, in addition to available data. Some NPI results in reliability are also discussed in the section titled "Statistical Inference in Imprecise Reliability". In the section titled "Challenges for Research and Application", we discuss some research challenges for imprecise reliability.

Imprecise Reliability *via* Natural Extension

Most of traditional reliability theory concerns analysis of systems, with probability distributions assumed to be known precisely. An attractive generalization provided by imprecise reliability theory enables such analyses under partial knowledge of probability distributions. This involves constrained optimization problems, where the inference of interest is the function to be maximized and minimized, to provide upper and lower bounds for the inference, respectively, and where all information available is formulated *via* constraints on the set of possible probability distributions. From this perspective, many basic problems in reliability have already been studied, a detailed recent overview with many references is given in [6]. Utkin [12, 13] considered system reliability, and provided algorithms for the optimization problems. Example 1 [6] is a simple example of the possibilities of imprecise reliability in this context.

Example 1 [6]. Consider a series system consisting of two components, where only the following information about the reliability of the components is available. The probability of the first component to fail before 10 h is 0.01. The mean time to failure of the second component is between 50 and 60 h. We want to draw inferences on the probability of system failure after 100 h. The information provided does not suffice to deduce unique precise probability distributions on the times to failure of the components, so standard reliability methods which use precise

probabilities would require additional assumptions. However, this information does restrict the possible probability distributions for these components' times to failure, and imprecise reliability now enables the inference in terms of the sharpest bounds for the probability of system failure after 100 h, given that the probability distributions of the components' times to failure satisfy the constraints following from the information available. The information provided is extremely limited, and the corresponding bounds on the inferences are that the system will experience its failure only after at least 100 h with lower probability 0 and upper probability 0.99. Suppose now, however, that we add one more piece of information, namely, that the failure times of the two components are statistically independent. The upper probability of system failure after at least 100 h now becomes 0.59. The corresponding lower probability remains 0, which agrees with intuition, as the information on both components does not exclude that either one of them might fail before 100 h with probability 1.

Although very basic, such an approach is of great value. It can give answers to many questions of interest, on the basis of the information that is available. If the answer is considered to be too imprecise, then additional assumptions or information are needed. It then provides insight on the effect of the extra information on the upper and lower bounds of the inference. This approach also indicates whether information or assumptions are conflicting, as no probability distributions would exist that satisfy all constraints. In such cases, one has to reconsider the available information and assumptions. Utkin, partly in collaboration with several coauthors, has presented theory and methodology for many such imprecise reliability inferences for systems, including topics on monotone systems, multistate and continuum-state systems, repairable systems, and reliability growth for software systems [6]. Further topics in imprecise reliability, where natural extension provides exciting opportunities for inference under limited information, include the use of expert judgments (*see Default Risk; Meta-Analysis in Non-clinical Risk Assessment*) and a variety of topics in risk analysis and decision support [6]. There are still many research topics that need to be addressed before a major impact on large-scale applications is possible.

Statistical Inference in Imprecise Reliability

Walley [3] developed statistical theory based on lower and upper probabilities, where many of the models suggested are closely related to Bayesian statistical methods with sets of prior distributions instead of single priors. An increasingly popular example is Walley's IDM [8] for multinomial observations. The multinomial distribution provides a standard model for statistical inference if a random quantity can belong to any one of $k \geq 2$ different categories. In the Bayesian statistical framework [14], Dirichlet distributions are natural conjugate priors for the multinomial model, so corresponding posteriors are also Dirichlet distributions. A convenient notation is as follows. Let exchangeable random quantities X_i have a multinomial distribution with $k \geq 2$ categories $C_1, \dots, C_k$ and parameter $\theta = (\theta_1, \dots, \theta_k)$, with $\theta_j \geq 0$ and $\sum_j \theta_j = 1$, then $P(X \in C_j | \theta) = \theta_j$. A Dirichlet distribution with parameters $s > 0$ and $t = (t_1, \dots, t_k)$, where $0 < t_j < 1$ and $\sum_j t_j = 1$, for the k -dimensional parameter θ , has probability density function (pdf) $p(\theta) \propto \prod_{j=1}^k \theta_j^{s t_j - 1}$. Suppose that data are available, consisting of the categories to which the random quantities $X_1, \dots, X_n$ belong, and that n_j of these belong to category C_j , with $n_j \geq 0$ and $\sum_j n_j = n$. The likelihood function corresponding to these data is $L(\theta | n_1, \dots, n_k) \propto \prod_{j=1}^k \theta_j^{n_j}$. Bayesian updating leads to the posterior pdf

$$p(\theta | n_1, \dots, n_k) \propto \prod_{j=1}^k \theta_j^{n_j + s t_j - 1} \quad (1)$$

This is again a Dirichlet distribution, with, compared to the prior, s being replaced by $n + s$ and t_j by $\frac{n_j + s t_j}{n + s}$.

Walley [8] introduced the IDM as follows. For fixed $s > 0$, define $D(s)$ as the set of all Dirichlet (s, t) distributions, so the k -vector t varies over all values with $t_j > 0$ and $\sum_j t_j = 1$. The set of posterior distributions corresponding to $D(s)$, in case of data $(n_1, \dots, n_k)$ with $\sum_j n_j = n$, is the set of all Dirichlet distributions with pdf given by equation (1), with $t_j > 0$ and $\sum_j t_j = 1$. For any event of interest, the lower and upper probabilities according to the IDM are derived as the infimum and supremum, respectively, of the probabilities for this event corresponding to Dirichlet distributions

in this set of posteriors. For example, if one is interested in the event that the next observation, X_{n+1} , belongs to C_j , then the IDM gives lower probability $\underline{P}(X_{n+1} \in C_j | n_1, \dots, n_k) = \frac{n_j}{n+s}$ and upper probability $\overline{P}(X_{n+1} \in C_j | n_1, \dots, n_k) = \frac{n_j + s}{n+s}$. The parameter s must be chosen independently of the data, Walley [8] advocates the use of $s = 1$ or 2 , on the basis of agreement of inferences with frequentist statistical methods.

The most obvious relevance of the IDM for reliability is with $k = 2$, when it reduces to the binomial distribution with a set of beta priors [3]. Binomial data often occur in reliability when success–failure data are recorded, for example, when the number of faulty units over a given period of time is of interest. In case of system reliability, one often records data at all relevant levels (components, subsystems, system) in terms of success or failure to perform their task. Hamada *et al.* [15], showed how, in a standard Bayesian approach with precise probabilities, binomial distributions with beta priors at component level can be used, for a given system structure (e.g., expressed via a fault tree), together with failure data gathered at different levels. For example, some observations would just be “system failure”, without detailed knowledge of which component(s) caused the failure. Such multilevel data incur dependencies between the parameters of the components' failure distributions. Hamada *et al.* [15], showed how Markov Chain Monte Carlo (MCMC) (*see Nonlife Loss Reserving; Reliability Demonstration; Bayesian Statistics in Quantitative Risk Assessment*) can be used for such inferences. Researchers have recently attempted to combine this approach with the IDM. Wilson *et al.* [16], applied the IDM to the same setting with multilevel failure data, although they expanded the binary approach from [15] to data with three categories (“failure”, “degraded”, “no failure”). They solve the computational problem, which is far more complex under imprecision than when precise probabilities are used, by brute force, as they perform multiple MCMCs, by first creating a fine grid over the space of prior distributions in the IDM, then running MCMC for each to derive the corresponding posterior probabilities, and finally computing the bounds over these. This only derives an upper (lower) bound for the lower (upper) probability, but if the MCMCs have been run long enough to ensure good convergence, and very many have been run, one is confident that

such a brute method provides good approximations to the lower and upper probabilities. This indicates a major research challenge for imprecise probabilistic statistical inference, namely, fast algorithms for computation. It is likely that optimization methods can be combined with simulation-based methods like MCMC in a way that requires less computational effort than the method presented in [16].

Troffaes and Coolen [17] generalized the approach from [15] differently. First, restricting to $k = 2$, they analytically derive the IDM-based upper and lower probabilities for system and component failure for a very small system (two components) and multilevel failure data. Then, they delete the usual assumption of independence of the two components, and show how the IDM can be used for inference without any assumption on such independence, and also if there may be unknown bias in data collecting and reporting. These are situations that cannot be solved with precise probabilities without additional assumptions. Other reliability methods and applications of IDM include inference for survival data including right-censored observations [18] and reliability analysis of multistate and continuum-state systems [19].

Imprecise probability enables inferential methods based on few mathematical assumptions if data are available. Coolen, with a number of coauthors, has developed NPI, where inferences are directly on future observable random quantities, e.g., the random time to failure of the next component to be used in a system. In this approach, imprecision depends in an intuitively logical way on the available data, as it decreases if information is added, yet aspects such as censoring or grouping of data result in an increase of imprecision. Foundations of NPI, including proofs of its consistency in theory of interval probability, are presented in [9], an overview with detailed comparison to so-called objective Bayesian methods is given in [10]. A first introductory overview of NPI in reliability is presented in [11], and theory for dealing with right-censored observations in NPI in [20, 21]. This framework is also suitable for guidance on high reliability demonstration, considering how many failure-free observations are required to accept a system in a critical operation [22]. The fact that, in such situations, imprecise reliability theory allows decisions to be based on the more pessimistic one of the lower and upper probabilities, e.g., lower probability of failure-free operation (*see Availability and Maintainability*), is an intuitively attractive manner

for dealing with indeterminacy. Recently, Coolen also considered probabilistic safety assessment from similar perspective [23]. NPI can also be used to support replacement decisions for technical equipment [24], giving decision support methods that are fully adaptive to failure data, and with imprecision reflected in bounds of cost functions. NPI can provide clear insights into the influence of a variety of assumptions that are often used for more established methods, and that may frequently be rather unrealistic if considered in detail. The fact that NPI can do without most of such assumptions and still be useful under reasonable data requirement is interesting.

Recent developments of NPI in reliability include the use of a new method for multinomial data for inferences on the possible occurrence of new failure modes [25] (see Example 3), and comparison of groups of success–failure data with specific attention to groups with zero failures [26]. To illustrate the use of NPI in reliability, we include two examples with references for further details.

Example 2 [11]. Studying 400 pumps in eight pressurized water reactors in commercial operation in the United States in 1972, six pumps failed to run normally for that whole year [14]. If we assume that all pumps are replaced preventively after 1 year, for example, to avoid failure due to wear-out, and that the pumps and relevant process factors remain similar, then it may be relevant to consider NPI for the number of failing pumps out of 400 during the following year, giving the lower and upper probabilities presented in Table 1 (upper probabilities for not-reported r values are less than 0.001).

Such lower and upper probabilities provide insight into the evidence in favor and against these particular events. Lower and upper probabilities for a subset of possible values cannot be derived by summing up the corresponding values in Table 1, because of the superadditivity (sub-) of lower (upper) probabilities, see [27] for more general results. This approach has the advantage over Bayesian methods [14] that no prior distributions are required, and their properties are such that they combine attractive features from both Bayesian and frequentist statistics [10].

Example 3 [25]. Suppose that a database contains detailed information on failures experienced during warranty periods of a particular product. Currently 200 failures have been recorded, with five different failure modes specified. The producer is interested in

Table 1 Lower and upper probabilities for r out of 400 pumps failing

r	$\underline{P}(r)$	$\overline{P}(r)$
0	0.0076	0.0153
1	0.0230	0.0541
2	0.0406	0.1088
3	0.0545	0.1641
4	0.0616	0.2059
5	0.0618	0.2270
6	0.0568	0.2273
7	0.0488	0.2111
8	0.0396	0.1844
9	0.0308	0.1533
10	0.0230	0.1222
11	0.0167	0.0939
12	0.0118	0.0700
13	0.0081	0.0508
14	0.0054	0.0359
15	0.0036	0.0249
16	0.0023	0.0169
17	0.0015	0.0113

the event that the next failure of such a product during its warranty period is caused by a failure mode other than the five already recorded. Let us first assume that there is no clear assumption or knowledge about the number of possible failure modes. Suppose that interest is in the event that the next reported failure is caused by an as yet unseen failure mode, then the NPI upper probability for this event is equal to $5/200$ [25]. If, however, the producer has actually specified two further possible failure modes, which have not yet been recorded so far, and interest is in the event that the next failure mode will be one of these two, then the NPI method gives upper probability $2/200$. This holds for any pair of such possible new failure modes that one wishes to specify, without restrictions on their number: this is possible because of the subadditivity of upper probability, which is an advantage over the use of precise probabilities. The NPI lower probabilities for both these events are zero, reflecting no strong evidence in the data that new failure modes can actually occur.

Now suppose that these 200 failures were instead caused by 25 different failure modes. Then the upper probability for the next failure mode to be any as yet unseen failure mode changes to $25/200$, but the upper probability for it to be either of any two described new failure modes remains $2/200$. It is in

line with intuition that the changed data affect this first upper probability, because the fact that more failure modes have been recorded suggests that there may be more failure modes. For the second event considered, the reasoning is somewhat different, as effectively interest is in two specific, as yet unseen, failure modes, and there is no actual difference in the data available that is naturally suggesting that either of these two failure modes has become more likely. If the producer has specific knowledge on the number of possible failure modes, that can also affect these upper probabilities [25].

Challenges for Research and Application

Imprecise probabilistic methods in reliability and risk assessment have clear practical advantages over more established and more restricted theory with precise probabilities. Owing to the possible interpretations of lower and upper probabilities, they are often convenient for practical risk assessment, as they reflect both “pessimistic” and “optimistic” inference. It is often clear on which of these to base decisions, for example, one would focus on the upper probability of an accident to occur.

Imprecise reliability is still at an early stage of development [6, 28], as are general theory and applications of lower and upper probabilities. There are many research challenges, ranging from foundational aspects to development of implementation tools. Upscaling from “textbook style” examples to larger scale problems provides many challenges, not only for computational methods but also e.g., for effective elicitation of expert judgment. At the foundational level, further study of the relation between imprecision and (amount of) information is required [29].

There is huge scope for new models and methods in imprecise reliability, and further development of some methods is required. For example, Utkin and Gurov [30] presented attractive classes of lifetime distributions, with $H(r, s)$, for $0 \leq r \leq s$, the class of all lifetime distributions with cumulative hazard function $\Lambda(t)$ such that $\Lambda(t)/t^r$ increases and $\Lambda(t)/t^s$ decreases in t . For example, $H(1, \infty)$ are all distributions with increasing hazard rates. Some results for related inference have been presented [30], but interesting research problems are still open, including how to fit such classes to

available data. The IDM and NPI have shown promising opportunities for applications in reliability, but statistical theory must be developed further to enhance applicability, for example, on the use of covariates and generalization to multivariate data.

References

- [1] Cai, K.Y. (1996). *Introduction to Fuzzy Reliability*, Kluwer Academic Publishers, Boston.
- [2] Dubois, D. & Prade, H. (1988). *Possibility Theory: An Approach to Computerized Processing of Uncertainty*, Plenum Press, New York.
- [3] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, London.
- [4] Weichselberger, K. (2000). The theory of interval-probability as a unifying concept for uncertainty, *International Journal of Approximate Reasoning* **24**, 149–170.
- [5] Weichselberger, K. (2001). *Elementare Grundbegriffe einer Allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als Umfassendes Konzept*, Physica, Heidelberg.
- [6] Utkin, L.V. & Coolen, F.P.A. (2007). Imprecise reliability: an introductory overview, in *Computational Intelligence in Reliability Engineering, Volume 2: New Metaheuristics, Neural and Fuzzy Techniques in Reliability*, G. Levitin, ed, Springer, Chapter 10, pp. 261–306.
- [7] Berger, J.O. (1990). Robust Bayesian analysis: sensitivity to the prior, *Journal of Statistical Planning and Inference* **25**, 303–328.
- [8] Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles (with discussion), *Journal of the Royal Statistical Society B* **58**, 3–57.
- [9] Augustin, T. & Coolen, F.P.A. (2004). Nonparametric predictive inference and interval probability, *Journal of Statistical Planning and Inference* **124**, 251–272.
- [10] Coolen, F.P.A. (2006). On nonparametric predictive inference and objective Bayesianism, *Journal of Logic, Language and Information* **15**, 21–47.
- [11] Coolen, F.P.A., Coolen-Schrijner, P. & Yan, K.J. (2002). Nonparametric predictive inference in reliability, *Reliability Engineering and System Safety* **78**, 185–193.
- [12] Utkin, L.V. (2004). Interval reliability of typical systems with partially known probabilities, *European Journal of Operational Research* **153**, 790–802.
- [13] Utkin, L.V. (2004). A new efficient algorithm for computing the imprecise reliability of monotone systems, *Reliability Engineering and System Safety* **86**, 179–190.
- [14] Martz, H.F. & Waller, R.A. (1982). *Bayesian Reliability Analysis*, John Wiley & Sons, New York.
- [15] Hamada, M.S., Martz, H., Reese, C.S., Graves, T., Johnson, V. & Wilson, A.G. (2004). A fully Bayesian approach for combining multilevel failure information in fault tree quantification and optimal follow-on resource allocation, *Reliability Engineering and System Safety* **86**, 297–305.
- [16] Wilson, A.G., Huzurbazar, A.V. & Sentz, K. (2007). The Imprecise Dirichlet Model for Multilevel System Reliability, *In submission*.
- [17] Troffaes, M. & Coolen, F.P.A. (2008). Applying the Imprecise Dirichlet Model in Cases with Partial Observations and Dependencies in Failure Data, *International Journal of Approximate Reasoning*, to appear.
- [18] Coolen, F.P.A. (1997). An imprecise Dirichlet model for Bayesian analysis of failure data including right-censored observations, *Reliability Engineering and System Safety* **56**, 61–68.
- [19] Utkin, L.V. (2006). Cautious reliability analysis of multi-state and continuum-state systems based on the imprecise Dirichlet model, *International Journal of Reliability, Quality and Safety Engineering* **13**, 433–453.
- [20] Coolen, F.P.A. & Yan, K.J. (2004). Nonparametric predictive inference with right-censored data, *Journal of Statistical Planning and Inference* **126**, 25–54.
- [21] Coolen, F.P.A. & Yan, K.J. (2003). Nonparametric predictive inference for grouped lifetime data, *Reliability Engineering and System Safety* **80**, 243–252.
- [22] Coolen, F.P.A. & Coolen-Schrijner, P. (2005). Nonparametric predictive reliability demonstration for failure-free periods, *IMA Journal of Management Mathematics* **16**, 1–11.
- [23] Coolen, F.P.A. (2006). On probabilistic safety assessment in case of zero failures, *Journal of Risk and Reliability* **220**, 105–114.
- [24] Coolen-Schrijner, P. & Coolen, F.P.A. (2004). Adaptive age replacement based on nonparametric predictive inference, *Journal of the Operational Research Society* **55**, 1281–1297.
- [25] Coolen, F.P.A. (2007). Nonparametric prediction of unobserved failure modes, *Journal of Risk and Reliability* **221**, 207–216.
- [26] Coolen-Schrijner, P. & Coolen, F.P.A. (2007). Nonparametric predictive comparison of success-failure data in reliability, *Journal of Risk and Reliability* **221**, 319–328.
- [27] Coolen, F.P.A. (1998). Low structure imprecise predictive inference for Bayes' problem, *Statistics and Probability Letters* **36**, 349–357.
- [28] Coolen, F.P.A. (2004). On the use of imprecise probabilities in reliability, *Quality and Reliability Engineering International* **20**, 193–202.
- [29] Coolen, F.P.A. & Newby, M.J. (1994). Bayesian reliability analysis with imprecise prior probabilities, *Reliability Engineering and System Safety* **43**, 75–85.
- [30] Utkin, L.V. & Gurov, S.V. (2002). Imprecise reliability for the new lifetime distribution classes, *Journal of Statistical Planning and Inference* **105**, 215–232.

FRANK P.A. COOLEN AND LEV V. UTKIN

Individual Risk Models

Insurance is protection against uncertainty. A contract (*policy*) between the insurance company (*insurer*) and a policyholder (*insured*) states that while the insurer pays a specified amount of money (also called *premium*) to the insurance company, in return, the company will compensate any loss suffered by the insured as a result of a specified random event (*claim*). A policy may include specific provisions that reduce the insurer's obligation (e.g., *deductibles*). Similar policies are grouped in an insurance portfolio.

If at a certain point of time the amount of surplus of the insurer becomes negative, then it is ruined. To avoid ruin, an insurance company sets apart some financial reserves and also buys risk protection from a reinsurance company (*see Credit Risk Models; Reinsurance; Securitization/Life*). There are various forms of reinsurance covers for the insurance portfolios.

To track the surplus, the claim payments must be modeled. This is why, for an insurance company, studying the aggregate claims distribution (*see Bonus–Malus Systems; Comonotonicity*), that is, the distribution of the aggregate amount of insurance claims occurring in an insurance portfolio within a given period, is of crucial importance (e.g., for premiums calculation, for evaluating the ruin probability and the needed reserves, for studying the effect of deductibles and reinsurance covers).

There are two main ways to model aggregate claims distributions; individual and the collective models. In the collective risk model, the aggregate claims of the portfolio is the sum of the individual claims (considered as independent and identically distributed), their number being a random variable (i.e., the number of claims arising in the considered period) independent of the payment amounts (*severities*). This decomposition allows consideration of claims number and claims amounts separately.

In an individual risk model, the aggregate claims of the portfolio is the sum of the aggregate claims of each of the policies. Thus, while in the collective model the claims are not related to the policies and the group of policies is taken as a whole (hence the name *collective*), in the individual model, the claims are related to the individual policies and

then aggregated to the portfolio (therefore the name *individual*).

In the following, we restrict ourselves to the individual risk model (*see Inequalities in Risk Theory*). Assuming that the policies (also called *risks*) in the portfolio are independent, it is straightforward to calculate the distribution of the sum of their aggregate claims by using convolutions. But this technique can be quite laborious, hence, alternative methods, such as recursive methods, have been investigated. Most of the recursions that have been derived for the general individual risk model can be considered as generalizations of [1].

A survey of recursions for aggregate claims distributions can be found in [2], while an overview of these recursive methods, in the framework of the individual risk model, is given in [3]. For a more detailed study of recursions grouping in the existing literature, we refer to [4].

The structure of this paper is as follows: we start by describing the individual risk model, then we present two exact recursions for it, followed by the three main approximative recursions. We end with other alternative methods and a numerical example.

In the following, we shall refer to probability functions as distributions, the sum over an empty set is assumed to be equal to zero, and we let $[x]$ denote the largest integer less than or equal to x .

Model Description

In the general individual risk model, the portfolio consists of n independent policies of different types. For each type, the aggregate claims distribution and the number of policies of that type is known; the distribution of the aggregate claims S of the entire portfolio is to be evaluated. A particular case of this model, well studied in the literature, assumes that the portfolio is divided into classes by clustering all policies with the same probability of a strictly positive claim amount and with the same severity distribution, as displayed in Table 1. The severity distribution (*see Dependent Insurance Risks; From Basel II to Solvency II – Risk Management in the Insurance Sector*) of a policy corresponds to the conditional distribution of the claim amount, given that a claim has occurred.

For the resulting two-way classification model, the class (i, j) , $i = 1, \dots, a$, $j = 1, \dots, b$, contains

Table 1 Portfolio structure for the two-way classification model

Claim amount distribution		Claim probability				
		q_1	q_2	$\cdots$	q_j	$\cdots$
	g_1					
	g_2				$\cdots$	
	$\cdots$					
	g_i			$\cdots$	n_{ij}	
	$\cdots$					
	g_a					

all policies with severity distribution g_i and claim probability q_j . The number of policies in class (i, j) is denoted by n_{ij} . It is assumed that for each j , $0 < q_j < 1$, and that the claim amounts of the individual policies are positive integers representing multiples of some convenient monetary unit, so that for each i , $g_i(x)$ is defined for $x = 1, 2, \dots$.

The probability that the aggregate claims $S = s$ is denoted by $f(s)$. We also use the notation $p_j = 1 - q_j$ and $n_{+j} = \sum_{i=1}^a n_{ij}$.

The case where the individual claim amounts can take only one strictly positive value, i.e., for each i , g_i is concentrated in one value, has been called the *individual life model* because of its straightforward application in life insurance: the portfolio consists of life insurance policies, and the corresponding strictly positive value is the benefit payable if death occurs within a certain exposure period.

Back to the general model, to see the applicability of the distribution of the aggregate claims S , we consider the following example in which the surplus U of a portfolio after a certain time period (e.g., 1 year), is given by

$$U = u + \pi - S \quad (1)$$

where π and S are respectively the portfolio's premium income and aggregate claims at the end of the period, while u is the initial reserve. Then the ruin probability (see **Longevity Risk and Life Annuities; Large Insurance Losses Distributions; Risk Measures and Economic Capital for (Re)insurers; Solvency; Ruin Probabilities: Computational Aspects**) at the end of the considered time period is

$$\psi(u) = \Pr(U < 0) = \Pr(S > u + \pi) \quad (2)$$

which can be evaluated knowing the distribution of S . From formula (2), one can also easily determine

the minimum reserve u needed to keep the ruin probability ψ under a maximum threshold imposed by the insurer, and hence decide whether reinsurance is needed. Similar formulae can be used to study how different types of reinsurance could affect the ruin probability of this particular portfolio.

In the following, we will present exact and approximative recursions for the two-way classification model, as developed initially. These recursions can be extended to the general individual risk model (see [5]).

Exact Recursions

We start by presenting recursions for the exact computation of the aggregate claims probabilities f .

De Pril's Exact Recursion

Using probability generating functions, De Pril [6] obtained the following exact recursion,

$$f(0) = \prod_{j=1}^b p_j^{n_{+j}} \quad (3)$$

$$f(s) = \frac{1}{s} \sum_{i=1}^a \sum_{k=1}^s A(i, k) \times \sum_{x=k}^s x g_i^{*k}(x) f(s-x), \quad s = 1, 2, \dots \quad (4)$$

with

$$A(i, k) = \frac{(-1)^{k+1}}{k} \sum_{j=1}^b n_{ij} \left(\frac{q_j}{p_j} \right)^k \quad (5)$$

and g_i^{*k} denoting the k -fold convolution of g_i . In our case,

$$g_i^{*k}(s) = \sum_{x=1}^{s-k+1} g_i^{*(k-1)}(s-x) g_i(x), \quad k = 2, \dots, s \quad (6)$$

since $g_i(x)$ is defined for $x \geq 1$, we must have $s-x \geq k-1$.

Dhaene–Vandebroek's Recursion

Dhaene and Vandebroek [7] proposed the following exact recursive formula

$$f(s) = \frac{1}{s} \sum_{i=1}^a \sum_{j=1}^b n_{ij} v_{ij}(s), \quad s = 1, 2, \dots \quad (7)$$

with the initial value given as before by formula (3), and where the $v_{ij}s$ are determined by

$$v_{ij}(s) = \frac{q_j}{p_j} \sum_{x=1}^s g_i(x) (x f(s-x) - v_{ij}(s-x)), \quad s = 1, 2, \dots \quad (8)$$

and $v_{ij}(s) = 0$ otherwise.

When $a = b = 1$, Dhaene–Vandebroek's recursion can be reduced to Panjer's recursion [8] with binomial counting distribution.

Approximations

To reduce the computing time, approximating f by a function $f^{(r)}$ was suggested. The three most well-known classes of such approximations are De Pril, Kornya, and Hipp.

Although f is a distribution, unfortunately, this is not always the case with $f^{(r)}$ (e.g., De Pril's and Hipp's approximations are not proper distributions). Another possible lack of an approximation is that its moments don't usually match the corresponding moments of the exact distribution, even for lower orders (this is the case with De Pril's and Kornya's approximations). Therefore, to measure the quality of different approximations, error measures were introduced and studied in the literature.

De Pril's Approximation

The De Pril approximation was introduced by De Pril [9] for the individual life model and De Pril [6] for the general model. This r th order approximation is obtained by simply summing up to $k = r$ in formula (4), that is

$$f^{(r)}(s) = \frac{1}{s} \sum_{i=1}^a \sum_{k=1}^{\min\{r,s\}} A(i, k) \sum_{x=k}^s x g_i^{*k}(x) f^{(r)}(s-x),$$

$$s = 1, 2, \dots \quad (9)$$

with the same initial value of formula (3), and the coefficients A given by formula (5).

Kornya's Approximation

Introduced by Kornya [10] for the individual life model and by Hipp [11] for the general model, Kornya's approximation is equal to De Pril's approximation up to a multiplicative constant chosen so that $f^{(r)}$ sums to one like a probability distribution. Thus, formula (9) is the same, but now we have the initial value

$$f^{K(r)}(0) = \exp \left(\sum_{j=1}^b n_{+j} \sum_{k=1}^r \frac{(-1)^k}{k} \left(\frac{q_j}{p_j} \right)^k \right) \quad (10)$$

Hipp's Approximation

Introduced by Hipp in [11], this approximation has the property that its moments match the moments of the exact distribution up to the order of the approximation, as shown in [12]. It follows that

$$f^{H(r)}(0) = \exp \left(- \sum_{j=1}^b n_{+j} \sum_{k=1}^r \frac{q_j^k}{k} \right) \quad (11)$$

$$f^{H(r)}(s) = \frac{1}{s} \sum_{x=1}^s x h^{(r)}(x) f^{H(r)}(s-x), \quad s = 1, 2, \dots \quad (12)$$

with

$$h^{(r)}(x) = \sum_{i=1}^a \sum_{j=1}^b n_{ij} \sum_{k=1}^{\min\{r,x\}} \sum_{l=1}^k \frac{(-1)^{l+1}}{k} \binom{k}{l} \times q_j^k g_i^{*l}(x), \quad x = 1, 2, \dots \quad (13)$$

Compared with the other approximations, Hipp's approximation requires more computation time.

Error Bounds

De Pril [6] compared these three approximations and deduced error bounds based on the error measure $\varepsilon(f, f^{(r)}) = \sum_{s=0}^{\infty} |f(s) - f^{(r)}(s)|$, separately

for each of the approximations. The deduction of the error bounds was unified in [13]. Dhaene and Sundt [14] deduced more inequalities for the error measure ε .

Computational Aspects

The efficiency of the different exact and approximate recursions, in terms of the number of computations required (especially multiplications and divisions), and the size of the array of the data to be kept, has been investigated partly in [7] and complemented in [3]. These authors concluded that if an exact recursive method is required, Dhaene–Vandebroek’s recursion will be the optimal choice, if the portfolio is not too heterogeneous. A similar comparison with two alternative binomial methods was conducted in [15], who showed that in some situations, a binomial or a combined method can perform better. In most practical cases, however, depending on the level of accuracy required, an “ r th order approximation” can be the best choice. From numerical examples (see [16]) and also from the error bounds for the approximations, it follows that in practical situations, with values of the claim probabilities q_j small enough (e.g., smaller than 1/2), the approximations for the probabilities converge very quickly to their respective exact values, so that a choice of r equal to 3, 4, or 5 will give almost exact results.

Also, a practical comparison of the compound Poisson approximation with De Pril’s, Kornya’s, and Hipp’s methods is given in [17].

Alternative Methods

Transforming the Individual Model

In the literature, there are other methods leading to recursive formulae for the individual model. For example, the individual model could be replaced by a collective one, and then the recursion of Panjer [8] to evaluate the associated compound distribution could be used. Such an approach leading to a compound Poisson model (see **Volatility Smile**) is described e.g., in [18] and [19]; the compound binomial model is presented in [20]; and generalizations can be found in [21]. (See also [22–27].) An overview and extensions of such results are presented in [28].

Sundt [29] proposed a “natural” approximation in which the original portfolio is replaced by a

homogeneous one. Here the distribution of S is approximated on the basis of convolutions calculated using the recursive formula described in [30].

The De Pril Transform

Inspired by De Pril [6], Sundt [31] introduced the De Pril transform of a distribution defined on the nonnegative integers with positive probability at zero. He noticed that for any such distribution f , there exists a unique function φ_f such that

$$f(s) = \frac{1}{s} \sum_{x=1}^s \varphi_f(x) f(s-x), \quad s = 1, 2, \dots \quad (14)$$

He called φ_f the De Pril transform of f . Solving equation (14) for $\varphi_f(s)$ gives

$$\varphi_f(s) = \frac{1}{f(0)} \left(s f(s) - \sum_{x=1}^{s-1} \varphi_f(x) f(s-x) \right) \\ s = 1, 2, \dots \quad (15)$$

As f should sum to one, its De Pril transform is unique. Several results concerning De Pril’s transform are presented in [31–33].

The De Pril transform can be used to express the exact and approximative recursions described in the sections titled “Exact Recursions” and “Approximations” in a unified form, based on formula (14). For example, De Pril’s exact recursion (formula 4) can be rewritten as formula (14) with

$$\varphi_f(x) = x \sum_{k=1}^x \frac{(-1)^{k+1}}{k} \sum_{j=1}^b \left(\frac{q_j}{p_j} \right)^k \sum_{i=1}^a n_{ij} g_i^{*k}(x), \\ x = 1, 2, \dots \quad (16)$$

Similarly, for De Pril’s approximative recursion (formula 9), one can use formula (14) with f replaced by $f^{(r)}$, and the sum in formula (16) up to $k = \min\{r, x\}$. The initial value of formula (3) is the same.

We notice that Hipp’s approximation is already in the form (14) with $\varphi_{f^{H(r)}}(x) = x h^{(r)}(x)$.

Dhaene and Sundt [32] also extended the definition of the De Pril transform to the class of all functions on the nonnegative integers with positive mass at zero.

Special Case: Individual Life Model

$$s = 1, 2, \dots \quad (18)$$

Recursions for this particular case have been derived in [34, 35] with

As mentioned before, the individual life model is obtained from the two-way classification individual risk model considered above, by choosing

$$g_i(x) = \begin{cases} 1, & x = c_i \\ 0, & \text{otherwise} \end{cases}, \quad i = 1, 2, \dots, a \quad (17)$$

with c_i the amount at risk (e.g., for a life insurance policy, the benefit payable if death occurs) of the policies in row i . In this case, the recursion in formula (4) reduces to

$$f(s) = \frac{1}{s} \sum_{i=1}^a \sum_{k=1}^{[s/c_i]} a(i, k) f(s - kc_i),$$

$$a(i, k) = (-1)^{k+1} c_i \sum_{j=1}^b n_{ij} \left(\frac{q_j}{p_j} \right)^k \quad (19)$$

This recursion was first derived in [34].

Table 2 Gerber's portfolio

	q_j			
	0.03	0.04	0.05	0.06
1	2	—	—	—
2	3	1	2	2
3	1	2	4	2
4	2	2	2	2
5	—	1	2	1

Number of policies: 31

Table 3 Probability function (pf) and cumulative distribution function (cdf): exact and De Pril's approximation

s	Exact, $D(4)$, $K(4)$, $H(4)$		$D(1)$		$D(2)$	
	pf	cdf	pf	cdf	pf	cdf
0	0.23819	0.23819	0.23819	0.23819	0.23819	0.23819
1	0.01473	0.25292	0.01473	0.25292	0.01473	0.25292
2	0.08773	0.34066	0.08796	0.34089	0.08773	0.34066
3	0.11318	0.45384	0.11319	0.45408	0.11317	0.45384
4	0.11070	0.56455	0.11297	0.56705	0.11070	0.56455
5	0.09632	0.66088	0.09656	0.66362	0.09632	0.66087
6	0.06154	0.72243	0.06519	0.72882	0.06146	0.72234
7	0.06902	0.79145	0.07031	0.79913	0.06901	0.79136
8	0.05481	0.84627	0.05918	0.85832	0.05479	0.84615
9	0.04314	0.88941	0.04543	0.90375	0.04300	0.88915
10	0.03010	0.91952	0.03421	0.93796	0.03006	0.91922
11	0.02352	0.94305	0.02644	0.96441	0.02346	0.94268
12	0.01828	0.96133	0.02100	0.98541	0.01813	0.96082
13	0.01250	0.97384	0.01525	1.0006	0.01243	0.97325
14	0.00871	0.98255	0.01084	1.0115	0.00862	0.98188
15	0.00591	0.98846	0.00776	1.0192	0.00578	0.98766
16	0.00415	0.99261	0.00554	1.0248	0.00407	0.99174
17	0.00271	0.99533	0.00388	1.0287	0.00263	0.99438
18	0.00174	0.99707	0.00262	1.0313	0.00167	0.99605
19	0.00111	0.99819	0.00177	1.0331	0.00105	0.99711
20	0.00071	0.99890	0.00119	1.0343	0.00066	0.99777
21	0.00044	0.99934	0.00079	1.0351	0.00040	0.99818
22	0.00026	0.99961	0.00051	1.0356	0.00023	0.99842
23	0.00016	0.99977	0.00033	1.0359	0.00013	0.99856
24	0.00009	0.99987	0.00021	1.0361	0.00007	0.99864
25	0.00005	0.99992	0.00013	1.0363	0.00004	0.99868
ε				0.03653		0.00126

Table 4 Probability function and cumulative distribution function: Kornya's and Hipp's approximation

s	$K(1)$		$K(2)$		$H(1)$		$H(2)$	
	pf	cdf	pf	cdf	pf	cdf	pf	cdf
0	0.22979	0.22979	0.23849	0.23849	0.24659	0.24659	0.23847	0.23847
1	0.01421	0.24401	0.01475	0.25324	0.01479	0.26139	0.01473	0.25321
2	0.08486	0.32887	0.08784	0.34109	0.08675	0.34814	0.08764	0.34085
3	0.10920	0.43807	0.11332	0.45441	0.11122	0.45936	0.11302	0.45387
4	0.10899	0.54706	0.11084	0.56526	0.11039	0.56976	0.11073	0.56461
5	0.09316	0.64023	0.09644	0.66171	0.09285	0.66262	0.09610	0.66071
6	0.06289	0.70313	0.06154	0.72325	0.06100	0.72363	0.06158	0.72230
7	0.06783	0.77097	0.06910	0.79236	0.06542	0.78906	0.06885	0.79115
8	0.05709	0.82807	0.05485	0.84722	0.05457	0.84363	0.05495	0.84610
9	0.04383	0.87190	0.04306	0.89028	0.04132	0.88495	0.04301	0.88912
10	0.03300	0.90491	0.03010	0.92038	0.03057	0.91553	0.03026	0.91938
11	0.02551	0.93042	0.02349	0.94387	0.02330	0.93884	0.02358	0.94297
12	0.02026	0.95068	0.01816	0.96203	0.01834	0.95718	0.01827	0.96124
13	0.01471	0.96540	0.01245	0.97448	0.01314	0.97033	0.01260	0.97384
14	0.01046	0.97586	0.00863	0.98312	0.00921	0.97955	0.00875	0.98259
15	0.00748	0.98335	0.00579	0.98891	0.00650	0.98606	0.00591	0.98851
16	0.00535	0.98871	0.00408	0.99300	0.00459	0.99065	0.00416	0.99268
17	0.00374	0.99245	0.00263	0.99563	0.00317	0.99383	0.00272	0.99540
18	0.00253	0.99499	0.00167	0.99731	0.00212	0.99595	0.00174	0.99714
19	0.00171	0.99670	0.00105	0.99837	0.00141	0.99736	0.00110	0.99825
20	0.00115	0.99785	0.00066	0.99904	0.00093	0.99830	0.00070	0.99895
21	0.00076	0.99862	0.00040	0.99944	0.00061	0.99892	0.00043	0.99938
22	0.00050	0.99912	0.00023	0.99968	0.00039	0.99932	0.00025	0.99964
23	0.00032	0.99944	0.00013	0.99982	0.00025	0.99957	0.00015	0.99979
24	0.00020	0.99965	0.00007	0.99990	0.00016	0.99973	0.00008	0.99988
25	0.00013	0.99978	0.00004	0.99995	0.00010	0.99983	0.00005	0.99993
ε		0.04366		0.00190		0.02629		0.00171

In the individual life model, the $v_{ij}s$ in formula (8) reduce to

$$v_{ij}(s) = \frac{q_j}{p_j} (c_i f(s - c_i) - v_{ij}(s - c_i)),$$

$$s = 1, 2, \dots \quad (20)$$

and the recursion in formula (7) with formula (20) is due to Waldmann [35]. Waldmann obtained his results by a rearrangement of the recursive scheme of De Pril [34] and showed that his exact recursive scheme is more efficient than the original algorithm of De Pril.

Recursions for the number of policies that lead to a strictly positive claim amount are obtained from formulas (18) and (20) by setting all c_i equal to 1. In this case, the recursion of De Pril [34] given in formula (18) reduces to the recursion of White and Greville [1].

Numerical Example

To illustrate the accuracy of the three approximations presented in the section titled “Approximations”, we consider the portfolio of Gerber [18], also used by several authors (see [11, 16, 20, 21, 29]). This portfolio corresponds to the individual life model with $a = 5$ and $c_i = i, i = 1, \dots, 5$ (see Table 2).

In Tables 3 and 4 we give the probability function and cumulative distribution $\Pr(S \leq s)$ of the portfolio calculated according to various algorithms, where *exact* stands for exact values, while $D(r)$, $K(r)$, and $H(r)$ respectively stand for De Pril's, Kornya's, and Hipp's approximations of order r . In the last line, we also show the value of the ε error measure defined in the section titled “Error Bounds”.

All the numerical values, ε excepted, are truncated after five decimals. This explains the fact that for $r = 4$ the values of the three approximations equal the exact values for $s = 0, \dots, 25$. Also, because of

a close similarity with the exact values, we did not show the approximated values for $r = 3$. Yet, we have the following error values:

$$\begin{aligned}\varepsilon(\text{Exact}, D(3)) &= 5.2212 \times 10^{-5}; \\ \varepsilon(\text{Exact}, D(4)) &= 2.3722 \times 10^{-6} \\ \varepsilon(\text{Exact}, K(3)) &= 8.7261 \times 10^{-5}; \\ \varepsilon(\text{Exact}, K(4)) &= 4.2989 \times 10^{-6} \\ \varepsilon(\text{Exact}, H(3)) &= 1.3529 \times 10^{-4}; \\ \varepsilon(\text{Exact}, H(4)) &= 1.0774 \times 10^{-5}.\end{aligned}$$

In conclusion, this numerical example supports the fact that a choice of $r \geq 4$ shall give almost exact results, and from the error values ε , it seems that De Pril's approximation converges faster.

References

- [1] White, R.P. & Greville, T.N.E. (1959). On computing the probability that exactly k out of n independent events will occur, *Transactions of the Society of Actuaries* **11**, 88–99. Discussion 96–99.
- [2] Sundt, B. (2002). Recursive evaluation of aggregate claims distributions, *Insurance: Mathematics and Economics* **30**, 297–322.
- [3] Dhaene, J., Ribas, C. & Vernic, R. (2006). Recursions for the individual risk model, *Acta Mathematicae Applicatae Sinicae – English series* **14**, 632–652.
- [4] Sundt, B. & Vernic, R.. *Recursions for Convolutions and Compound Distributions With Insurance Applications*. Book manuscript in progress.
- [5] Sundt, B. (2004). De Pril Recursions and Approximations, in *Encyclopedia of Actuarial Science*, John Wiley & Sons, Vol. 1, pp. 425–430.
- [6] De Pril, N. (1989). The aggregate claims distribution in the individual model with arbitrary positive claims, *ASTIN Bulletin* **19**, 9–24.
- [7] Dhaene, J. & Vandebroek, M. (1995). Recursions for the individual model, *Insurance: Mathematics and Economics* **16**, 31–38.
- [8] Panjer, H.H. (1981). Recursive evaluation of a family of compound distributions, *ASTIN Bulletin* **12**, 22–26.
- [9] De Pril, N. (1988). Improved approximations for the aggregate claims distribution of a life insurance portfolio, *Scandinavian Actuarial Journal* 61–68.
- [10] Kornya, P.S. (1983). Distribution of aggregate claims in the individual risk model, *Transactions of the Society of Actuaries* **35**, 823–836. Discussion: 837–858.
- [11] Hipp, C. (1986). Improved approximations for the aggregate claims distribution in the individual model, *ASTIN Bulletin* **16**(2), 89–100.
- [12] Dhaene, J., Sundt, B. & De Pril, N. (1996). Some moment relations for the Hipp approximation, *ASTIN Bulletin* **26**, 117–121.
- [13] Dhaene, J. & De Pril, N. (1994). On a class of approximative computation methods in the individual risk model, *Insurance: Mathematics and Economics* **14**, 181–196.
- [14] Dhaene, J. & Sundt, B. (1997). On error bounds for approximations to aggregate claims distributions, *ASTIN Bulletin* **27**, 243–262.
- [15] Sundt, B. & Vernic, R. (2006). Two binomial methods for evaluating the aggregate claims distribution in De Pril's individual risk model, *Belgian Actuarial Bulletin* **6**(1), 5–13.
- [16] Vandebroek, M. & De Pril, N. (1988). Recursions for the distribution of a life portfolio: a numerical comparison, *Bulletin of the Royal Association of Belgian Actuaries* **82**, 21–35.
- [17] Kuon, S., Reich, A. & Reimers, L. (1987). Panjer vs. Kornya vs. De Pril: a comparison from a practical point of view, *ASTIN Bulletin* **17**(2), 183–191.
- [18] Gerber, H.U. (1979). *An Introduction to Mathematical Risk Theory; Huebner Foundation Monographs* 8, Richard D. Irwin, Homewood.
- [19] Gerber, H.U. (1997). *Life Insurance Mathematics*, Springer-Verlag, Berlin.
- [20] Jewell, W.S. & Sundt, B. (1981). Improved approximations for the distribution of a heterogeneous risk portfolio, *Bulletin of the Association of Swiss Actuaries* **2**, 221–241.
- [21] Chan, F.Y. (1984). On a family of aggregate claims distributions, *Insurance: Mathematics and Economics* **3**, 151–155.
- [22] Bühlmann, H., Gagliardi, B., Gerber, H.U. & Straub, E. (1977). Some inequalities for stop-loss premiums, *ASTIN Bulletin* **9**, 75–83.
- [23] Gerber, H.U. (1984). Error bounds for the compound Poisson approximation, *Insurance: Mathematics and Economics* **3**, 191–194.
- [24] Hipp, C. (1985). Approximation of aggregate claims distributions by compound Poisson distributions, *Insurance: Mathematics and Economics* **4**, 227–232. Correction note: *Insurance: Mathematics and Economics* 1985, 6, 165.
- [25] Kuon, S., Radtke, M. & Reich, A. (1993). An appropriate way to switch from the individual model to the collective one, *ASTIN Bulletin* **23**, 23–54.
- [26] Michel, R. (1987). An improved error bound for the compound Poisson approximation, *ASTIN Bulletin* **17**, 165–169.
- [27] Panjer, H.H. & Willmot, G.E. (1992). *Insurance Risk Models*, Society of Actuaries, Schaumburg.
- [28] De Pril, N. & Dhaene, J. (1992). Error bounds for compound Poisson approximations of the individual risk model, *ASTIN Bulletin* **19**(1), 135–148.
- [29] Sundt, B. (1985). On approximations for the distribution of a heterogeneous risk portfolio, *Bulletin of the Association of Swiss Actuaries* 189–203.

- [30] De Pril, N. (1985). Recursions for convolutions of arithmetic distributions, *ASTIN Bulletin* **15**, 135–139.
- [31] Sundt, B. (1995). On some properties of De Pril transforms of counting distributions, *ASTIN Bulletin* **25**, 19–31.
- [32] Dhaene, J. & Sundt, B. (1998). On approximating distributions by approximating their De Pril transforms, *Scandinavian Actuarial Journal* 1–23.
- [33] Sundt, B. (2000). The multivariate De Pril transform, *Insurance: Mathematics and Economics* **27**, 123–136.
- [34] De Pril, N. (1986). On the exact computation of the aggregate claims distribution in the individual life model, *ASTIN Bulletin* **16**, 109–112.
- [35] Waldmann, K.-H. (1994). On the exact calculation of the aggregate claims distribution in the individual life model, *ASTIN Bulletin* **24**, 89–96.

RALUCA VERNIC

Inequalities in Risk Theory

Consider an insurance portfolio containing n risks. The risks may have different distribution functions belonging to a known parametric family of distributions. For example, consider a life insurance portfolio with hidden heterogeneity, due to different backgrounds of the insureds, that is unknown to the insurer. We would like to compare the premia between two portfolios with different amounts of heterogeneity, when all the insureds pay the same premium. In this paper we examine the impact of heterogeneity on the portfolio's risk measure. A risk measure function ρ assigns to each nonnegative random variable a nonnegative number. For example, a premium can be viewed as a risk measure. The larger the value of $\rho(X)$, the more dangerous the risk X , and thus more capital is required to hedge this risk. Specifically, we expect that $\rho(X) \geq \rho(\mathbb{E}(X))$, where $\mathbb{E}(X)$ is the mean of the random variable X . Let g be an increasing convex function. The risk measure defined by $\rho(X) = \mathbb{E}g(X)$ has many desirable properties; especially note that according to Jensen's inequality $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$. For a comprehensive overview on risk measures (see **Risk Classification/Life**), see Denuit and Frostig [1].

Consider two portfolios: portfolio I contains the risks $(X_1, \dots, X_n)$ and portfolio II contains the risks $(Y_1, \dots, Y_n)$. Assume that the distribution functions of X_i and Y_i , $i = 1, \dots, n$ belong to the same parametric family of distributions. Let $\theta = (\theta_1, \dots, \theta_n)$ and $\eta = (\eta_1, \dots, \eta_n)$, be the vectors of parameters of the distribution functions of the risks in portfolio I and II, respectively. Let η be more heterogeneous or, more diverse, in a sense that will be defined later, than θ . The goal of this paper is to study the impact of the increase in the heterogeneity on the portfolio's risk measure.

There is no definite answer to the question whether heterogeneity reduces or increases risk, and the answer depends on the type of heterogeneity. An example where heterogeneity increases risk was studied by Denuit and Vermandele [2]. They considered a portfolio containing n independent identically distributed (i.i.d.) risks $X_1, \dots, X_n$, and functions $\psi_1, \dots, \psi_n$ with $\bar{\psi}(x) = \frac{1}{n} \sum_{j=1}^n \psi_j(x)$. They proved that the loss $\sum_{j=1}^n \bar{\psi}(X_j)$ is preferred over

$\sum_{j=1}^n \psi_j(X_j)$ by all decision makers, with nondecreasing and concave utility functions.

On the other hand, the following example demonstrates that heterogeneity reduces risk. Consider two portfolios each containing two risks. Let the first one be (X_1, X_2) and the second one be (Y_1, Y_2) . Let $X_i = J_{p_i} C_i$ and $Y_i = J_{q_i} C_i$, $i = 1, 2$, where C_1, C_2 denote the claim amounts and are identically distributed with mean μ and variance σ^2 . J_{p_i} equals 1 with probability p_i , and 0 with probability $1 - p_i$. It indicates whether a claim arose for policy i . All the random variables involved are independent. Assume further that $p_1 + p_2 = q_1 + q_2$ and that $p_2 - p_1 < q_2 - q_1$. Thus, the vector $\mathbf{q} = (q_1, q_2)$ is more heterogeneous than $\mathbf{p} = (p_1, p_2)$. Note that the mean of total payout is the same for the two portfolios. Let us measure the risk by the variance of the total payout.

$$\begin{aligned} \text{Var}[X_1 + X_2] &= (\sigma^2 + \mu^2)(p_1 + p_2) - \mu^2(p_1^2 + p_2^2) \\ &= (\sigma^2 + \mu^2)(p_1 + p_2) - \frac{\mu^2}{2} \\ &\quad \times \left((p_1 + p_2)^2 + (p_2 - p_1)^2 \right) \quad (1) \end{aligned}$$

Similarly

$$\begin{aligned} \text{Var}[Y_1 + Y_2] &= (\sigma^2 + \mu^2)(p_1 + p_2) - \frac{\mu^2}{2} \\ &\quad \times \left((p_1 + p_2)^2 + (q_2 - q_1)^2 \right) \quad (2) \end{aligned}$$

The variance of the more heterogeneous one is less than the variance of the more homogeneous one. In this example heterogeneity is preferable.

To illustrate the last example numerically, assume that $p_1 = p_2 = 1/3$ and $q_1 = 1/6$, $q_2 = 1/2$. Then,

$$\text{Var}[X_1 + X_2] = \frac{2}{3}(\sigma^2 + \mu^2) - \frac{\mu^2}{2} \frac{4}{9} \quad (3)$$

$$\text{Var}[Y_1 + Y_2] = \frac{2}{3}(\sigma^2 + \mu^2) - \frac{\mu^2}{2} \frac{5}{9} \quad (4)$$

In the sequel we will generalize this example to portfolios containing n risks. Throughout the paper bold capital letters denote n -dimensional random vectors, while bold small letters denote n -dimensional vectors with real components. We apply the concepts of majorization and Schur functions to compare the heterogeneity in two portfolios. The comparison is

made for different examples. We assume that the random variables involved in the various examples are statistically independent.

Majorization and Schur Functions

Majorization is a tool to measure the diversity of the components of an n -dimensional vector. Comprehensive references for majorization order and its applications in probability and statistics appear in Marshall and Olkin [3] and Arnold [4].

Conceptually, the vector $\mathbf{y} = (y_1, \dots, y_n)$ majorizes the vector $\mathbf{x} = (x_1, \dots, x_n)$ if the components of $\mathbf{y}$ are more diverse than the components of $\mathbf{x}$. Formally, let $x_{(1,n)} \geq x_{(2,n)} \geq \dots \geq x_{(n,n)}$ and $y_{(1,n)} \geq y_{(2,n)} \geq \dots \geq y_{(n,n)}$ denote the decreasing rearrangement of $\mathbf{x}$ and $\mathbf{y}$; then $\mathbf{y}$ is said to majorize $\mathbf{x}$, denoted as $\mathbf{x} \leq_{\text{maj}} \mathbf{y}$ if

$$\sum_{i=1}^k y_{(i,n)} \geq \sum_{i=1}^k x_{(i,n)}, \quad k = 1, \dots, n-1 \quad (5)$$

and

$$\sum_{i=1}^n y_{(i,n)} = \sum_{i=1}^n x_{(i,n)} \quad (6)$$

Let $1/2 \leq \alpha \leq 1$. Define the $\mathbf{T}$ transform as follows:

$$\begin{aligned} \mathbf{T}\mathbf{x} = \mathbf{T}(x_1, \dots, x_n) &= (x_1, \dots, \alpha x_i \\ &+ (1 - \alpha)x_j, \dots, x_{j-1}, \alpha x_j \\ &+ (1 - \alpha)x_i, \dots, x_n) \end{aligned} \quad (7)$$

Under the $\mathbf{T}$ transform, two components of the vector $\mathbf{x}$ are replaced by their weighted averages such that their sum is not changed. If $\mathbf{x} \leq_{\text{maj}} \mathbf{y}$, then $\mathbf{x}$ can be derived from $\mathbf{y}$ by a finite number of $\mathbf{T}$ transforms, (cf. Marshall and Olkin [3], Arnold [4], and Ma [5]).

Schur increasing functions are functions that increase in the majorization order. This can be formally stated as follows.

Definition 1 Let $A \subset \mathbb{R}^n$. A function $\phi : A \rightarrow \mathbb{R}$ is said to be Schur increasing (decreasing) if $\phi(\mathbf{x}) \leq (\geq) \phi(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in A$, where $\mathbf{x} \leq_{\text{maj}} \mathbf{y}$.

Note that Schur functions are symmetric, that is, if π is a permutation of the numbers $1, \dots, n$,

then $g(x_1, \dots, x_n) = g(x_{\pi(1)}, \dots, x_{\pi(n)})$. Consider the following examples of Schur increasing and Schur decreasing functions:

1. $g : \mathbb{R}_+^n \rightarrow \mathbb{R}_+$, $g(\mathbf{x}) = \prod_{i=1}^n x_i$ is Schur decreasing.
2. A function $\phi : \mathbb{R}_+^n \rightarrow \mathbb{R}_+$, is convex (concave) if for any $\mathbf{y}, \mathbf{z} \in \mathbb{R}_+^n$ and for $0 \leq \alpha \leq 1$

$$\phi(\alpha \mathbf{y} + (1 - \alpha)\mathbf{z}) \leq (\geq) \alpha \phi(\mathbf{y}) + (1 - \alpha)\phi(\mathbf{z}) \quad (8)$$

Marshall and Olkin [3] showed that if ϕ is symmetric and convex then it is Schur increasing.

3. If g is a univariate convex (concave) function, then $\phi(\mathbf{x}) = \sum_{j=1}^n g(x_j)$ is Schur increasing (decreasing).

In the sequel we require a multivariate extension of the majorization ordering. We say that the vectors $(\bar{\mathbf{w}}_1, \dots, \bar{\mathbf{w}}_r)$ are majorized in the Robin Hood (RH) sense by $(\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_r)$, written as $(\bar{\mathbf{w}}_1, \dots, \bar{\mathbf{w}}_r) \leq_{\text{RH}} (\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_r)$, if there exists a finite set of $\mathbf{T}_1, \dots, \mathbf{T}_K$ of $\mathbf{T}$ transforms, such that $\bar{\mathbf{w}}_i = \mathbf{T}_1 \dots \mathbf{T}_K \bar{\mathbf{v}}_i$, for $i = 1, \dots, r$. If the components of $\bar{\mathbf{w}}_i$ and $\bar{\mathbf{v}}_i$ are arranged in increasing order and the set of the $\mathbf{T}$ transforms preserve that ordering, we say that $(\bar{\mathbf{w}}_1, \dots, \bar{\mathbf{w}}_r)$ are majorized in the ordered Robin Hood (ORH) sense by $(\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_r)$, written as $(\bar{\mathbf{w}}_1, \dots, \bar{\mathbf{w}}_r) \leq_{\text{ORH}} (\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_r)$. In both definitions $\bar{\mathbf{w}}_i$ is obtained from $\bar{\mathbf{v}}_i$ by the same averaging scheme, for all $i = 1, \dots, r$.

For example, the vectors $((1, 3, 11), (1, 2, 5))$ are bigger than $((6, 3, 6), (3, 2, 3))$ in the sense of RH, but not in the sense of ORH. Clearly ORH ordering implies RH ordering.

Heterogeneity in General Insurance Portfolios

We compare two portfolios: portfolio I containing n risks $(X_1, \dots, X_n)$ and portfolio II containing n risks $(Y_1, \dots, Y_n)$. Assume that $X_i = \sum_{j=1}^{N_i} C_{i,j}$ and $Y_i = \sum_{j=1}^{M_i} D_{i,j}$. Thus, N_i or M_i is the number of claims in a certain period for insured i , or business i , and $C_{i,j}$ or $D_{i,j}$ is the j th claim amount of the i th insured, or the i th business line.

In the sequel we will assume that the claim amount distributions of $C_{i,j}$ and $D_{i,j}$ belong to one of the following parametric family of distributions:

C1: Location-scale family

Let $Z_{i,j}$, $i = 1, \dots, n$, $j \geq 1$ be i.i.d. random variables. Let the claim amount in portfolio I be $C_{i,j}$, where $C_{i,j} = a_i Z_{i,j} + s_i$, and let the claim amount in portfolio II be $D_{i,j} = b_i Z_{i,j} + t_i$. We compare the two sets of vectors $(\mathbf{a}, \mathbf{s})$ and $(\mathbf{b}, \mathbf{t})$ by the RH or by the ORH ordering.

Example 1 Let $Z_{i,j}$ be standard normal random variables, and let $\mathbf{a} = (1.5, 1.5, 1.5)$, $\mathbf{s} = (30, 30, 30)$, $\mathbf{b} = ((0.5, 1, 3)$, and $\mathbf{t} = (10, 20, 60))$. Note that

$$\begin{aligned} ((0.5, 1, 3), (10, 20, 60)) &\succeq_{\text{orh}} ((0.5, 2, 2), (10, 40, 40)) \\ &\succeq_{\text{orh}} ((1, 1.5, 2), (20, 30, 40)) \\ &\succeq_{\text{orh}} ((1.5, 1.5, 1.5), \\ &\quad (30, 30, 30)) \end{aligned} \quad (9)$$

Thus, $(\mathbf{a}, \mathbf{s})$ is majorized by $(\mathbf{b}, \mathbf{t})$ in ORH ordering, and hence in RH ordering.

C2: Semigroup property with respect to convolution

$C_{i,j}$ and $D_{i,j}$, $j \geq 1$ are distributed as Z_{θ_i} and Z_{η_i} with distribution functions F_{θ_i} and F_{η_i} , respectively, belonging to the same parametric family of distributions $\mathcal{F} = \{F_\alpha, \alpha \in \Theta\}$. This family has the following property: if Z_α and Z_β are independent with distribution functions F_α and F_β respectively, then $Z_\alpha + Z_\beta$ is distributed as $Z_{\alpha+\beta}$. This property is called the *semigroup property* with respect to convolution. Assume that the vector of parameters for portfolio II is more heterogeneous than that of portfolio I, i.e., $\boldsymbol{\theta} \preceq_{\text{maj}} \boldsymbol{\eta}$.

In Example 2 we present a family of distributions for which assumption C2 holds.

Example 2 Consider the family of gamma distributions with an identical scale parameter, say λ , and shape parameter θ_i or η_i . Thus, the density function of C_i is $f_{C_i}(x) = \frac{1}{\Gamma(\theta_i)} e^{-\lambda x} \lambda^{\theta_i} x^{\theta_i-1}$, and the density function of D_i is the same with the parameter η_i replacing θ_i . Consider two portfolios containing three risks. Let $\lambda = 0.1$, $\boldsymbol{\eta} = (1, 2, 6)$, and $\boldsymbol{\theta} = (3, 3, 3)$. Then

$$(1, 2, 6) \succeq_{\text{maj}} (1, 4, 4) \succeq_{\text{maj}} (2, 3, 4) \succeq_{\text{maj}} (3, 3, 3) \quad (10)$$

When comparing portfolios I and II we consider three types of heterogeneity: (a) heterogeneity in the claims arrival process, (b) heterogeneity in the claim

amounts, and (c) heterogeneity in both the arrival process and the claim amounts.

We consider the individual risk model first and then the collective risk model.

Heterogeneity in the Individual Risk Model

In this case N_i is 1 with probability p_i and 0 otherwise; similarly, M_i is 1 with probability q_i and 0 otherwise. Thus, $X_i = J_{p_i} C_i$, where J_{p_i} equals 1 if individual i has a claim during a given period and 0 otherwise; similarly, $Y_i = K_{q_i} D_i$, with K_{q_i} equals 1 with probability q_i and 0 otherwise. Given that a claim has occurred, its amount is a positive random variable C_i or D_i for the i th insured in portfolio I or II, respectively.

Case a: Heterogeneity in claim occurrence probabilities

We generalize the example given in the introduction to a portfolio containing n risks. Assume that (a) C_i and the D_i , $i = 1, \dots, n$, are i.i.d. random variables. Thus, all the claim amounts are statistically identical. (b) The claim occurrence probabilities $\mathbf{p} = (p_1, \dots, p_n)$ of portfolio I vector is less heterogeneous than $\mathbf{q} = (q_1, \dots, q_n)$ – the claim occurrence probabilities vector of portfolio II in the sense of majorization ordering, i.e., $\mathbf{p} \preceq_{\text{maj}} \mathbf{q}$.

Ma [5] proved that, for any convex function g , $\mathbb{E}[g(\sum_{i=1}^n X_i)] \leq \mathbb{E}[g(\sum_{i=1}^n Y_i)]$. Thus, the more heterogeneous the vector of the claim occurrence probabilities is, the less risky is the portfolio with respect to a convex risk measure. Note that $\mathbb{E}[\sum_{i=1}^n X_i] = \mathbb{E}[\sum_{i=1}^n Y_i]$. Thus $\text{Var}[\sum_{i=1}^n X_i] \leq \text{Var}[\sum_{i=1}^n Y_i]$.

An extension of this result to the case where the claim amounts are not identically distributed can be found in Denuit and Frostig [1].

Case b: Heterogeneity in the Claim Amounts

Assume that portfolios I and II have the same occurrence probabilities. Thus, $X_i = J_{p_i} C_i$ and $Y_i = J_{p_i} D_i$. Further $p_1 \leq p_2 \leq \dots \leq p_n$. Then if

1. assumption C1 holds, i.e., C_i and D_i belong to the same location-scale family of distributions with $(\mathbf{a}, \mathbf{s}) \preceq_{\text{ORH}} (\mathbf{b}, \mathbf{t})$, or,
2. C_i and D_i belong to the same parametric family of distributions with the semigroup property with

respect to convolution, as described in C2, and $\theta \preceq_{\text{ORH}} \eta$,

then portfolio II is riskier than portfolio I, in the sense that for every increasing convex function g the following holds:

$$\mathbb{E} \left[g \left(\sum_{i=1}^n X_i \right) \right] \leq \mathbb{E} \left[g \left(\sum_{i=1}^n Y_i \right) \right] \quad (11)$$

Example 3 Consider two portfolios I and II with the same claim occurrence probabilities $\mathbf{p} = (0.1, 0.2, 0.3)$ and the claim size distributions as described in Example 1 or Example 2. Then portfolio II is riskier than portfolio I in the sense that inequality (11) holds for any increasing convex function.

Case c: Heterogeneity in both the claim occurrence probabilities and the claim amounts

Let the distribution functions of C_i and D_i , $i = 1, \dots, n$ belong to the same location-scale family, i.e., C1 holds, and,

$$(\mathbf{a}, \mathbf{s}, \mathbf{q}) \preceq_{\text{RH}} (\mathbf{b}, \mathbf{t}, \mathbf{p}) \quad (12)$$

Then equation (11) holds for any convex function g . Thus, the risk increases as the heterogeneity of the occurrence probability decreases, and the heterogeneity in the claim amounts distributions increases. In this case, $\mathbb{E} [\sum_{i=1}^n X_i] = \mathbb{E} [\sum_{i=1}^n Y_i]$, while $\text{Var} [\sum_{i=1}^n X_i] \leq \text{Var} [\sum_{i=1}^n Y_i]$. For illustration consider the following example:

Example 4 Consider claim size distributions as per Example 1, and let the claim occurrence probabilities for portfolio I be $\mathbf{p} = (0.1, 0.1, 0.7)$ and for portfolio II be $\mathbf{q} = (0.3, 0.3, 0.3)$. Note that $(\mathbf{a}, \mathbf{s}, \mathbf{q}) \preceq_{\text{RH}} (\mathbf{b}, \mathbf{t}, \mathbf{p})$, since

$$\begin{aligned} & ((0.5, 1, 3), (10, 20, 60), (0.1, 0.1, 0.7)) \\ & \succeq_{\text{RH}} ((0.5, 2, 2), (10, 40, 40), (0.1, 0.4, 0.4)) \\ & \succeq_{\text{RH}} ((1, 1.5, 2), (20, 30, 40), (0.2, 0.3, 0.4)) \\ & \succeq_{\text{RH}} ((1.5, 1.5, 1.5), (30, 30, 30), (0.3, 0.3, 0.3)) \end{aligned} \quad (13)$$

In this example, $\mathbb{E}(X_1 + X_2 + X_3) = \mathbb{E}(Y_1 + Y_2 + Y_3) = 27$, while $\text{Var}(X_1 + X_2 + X_3) = 83.025 < \text{Var}(Y_1 + Y_2 + Y_3) = 504.075$.

Thus, portfolio II is riskier than portfolio I. Contrary to the results that show that the risk decreases

as the heterogeneity in the occurrence probabilities increases, increasing the heterogeneity of some functions of the occurrence probabilities might increase the risk of the portfolio.

Let $\mathbf{h}(\mathbf{p}) = (h(p_1), \dots, h(p_n))$, and $\mathbf{h}(\mathbf{q}) = (h(q_1), \dots, h(q_n))$. Assume that the claim amount distributions belong to a parametric family of distributions with the semigroup property with respect to convolution, as described in assumption C2. Let $h(x) = \ln(x)$, or for $h(x) = \frac{1-x}{x}$, Denuit and Frostig [1] showed that

$$(\mathbf{h}(\mathbf{p}), \theta) \preceq_{\text{ORH}} (\mathbf{h}(\mathbf{q}), \eta) \text{ implies that } \mathbb{E} \left[g \left(\sum_{i=1}^n X_i \right) \right] \leq \mathbb{E} \left[g \left(\sum_{i=1}^n Y_i \right) \right] \quad (14)$$

for any nondecreasing convex function g . Thus, the more heterogeneous portfolio is the riskier one with respect to increasing convex risk measure. Consider the following example for the case where $h(x) = \ln(x)$.

Example 5 Consider portfolio I with claim occurrence probabilities with $\ln(\mathbf{p}) = (-1, -1, -1)$, (corresponding to the vector of occurrence probabilities $\mathbf{p} = (0.3679, 0.3679, 0.3679)$), and portfolio II containing three risks with claim occurrence probabilities with $\ln(\mathbf{q}) = (-2, -0.9, -0.1)$, (corresponding to the vector of occurrence probabilities $\mathbf{p} = (0.1353, 0.4066, 0.9048)$). The claim amount distributions corresponding to portfolios I and II are as per Example 2, i.e., all the claim amounts are gamma distributed with scale parameter 0.1, and shape parameter $\theta = (3, 3, 3)$ for portfolio I, and $\eta = (1, 2, 6)$ for portfolio II. Note that

$$\begin{aligned} & ((-2, -0.9, -0.1), (1, 2, 6)) \\ & \succeq_{\text{ORH}} ((-2, -0.5, -0.5), (1, 4, 4)) \\ & \succeq_{\text{ORH}} ((-1.5, -1, -0.5), (2, 3, 4)) \\ & \succeq_{\text{ORH}} ((-1, -1, -1), (3, 3, 3)) \end{aligned} \quad (15)$$

Thus, $(\ln(\mathbf{p}), \theta) \preceq_{\text{ORH}} (\ln(\mathbf{q}), \eta)$, and portfolio II is riskier than portfolio I.

Heterogeneity in the Collective Risk Model

In this case, N_i and M_i are independent nonnegative integer-valued random variables. N_i or M_i are the number of claim arrivals of business line i during a

certain period for portfolios I and II, respectively. Consider three cases: 1. heterogeneity only in the arrival process, 2. heterogeneity only in the distribution of the claim amounts, and 3. heterogeneity both in the arrival process and claim amounts.

Case 1: Heterogeneity only in the arrival process

Assume the following:

1. The number of type i claim arrivals to portfolios I and II have distribution functions G_{λ_i} and G_{v_i} , respectively, belonging to the same family of parametric distributions $\mathcal{G} = \{G_\lambda, \lambda \in \Lambda\}$, with the semigroup property with respect to convolution. Let $\lambda = (\lambda_1, \dots, \lambda_n)$ and $v = (v_1, \dots, v_n)$.
2. $\lambda \preceq_{\text{maj}} v$ (16)

Then for any symmetric convex function g ,

$$\mathbb{E}[g(\mathbf{X})] \leq \mathbb{E}[g(\mathbf{Y})] \quad (17)$$

In particular, for a univariate convex function g , equation (11) holds.

Thus, the portfolio with the more heterogeneous arrival process is the riskier one.

Example 6 Assume that the number of claim arrivals of business i is a Poisson random variable with parameter λ_i for portfolio I, and parameter v_i for portfolio II. The family of Poisson distributions has the semigroup with respect to convolution. Let $\lambda = (300, 300, 300)$ and $v = (100, 100, 700)$. Then $(100, 100, 700) \preceq_{\text{maj}} (100, 400, 400) \preceq_{\text{maj}} (200, 300, 400) \preceq_{\text{maj}} (300, 300, 300)$.

Case 2: Heterogeneity only in the distribution of the claim amounts

Let $X_i = \sum_{j=1}^{N_i} C_{i,j}$ and $Y_i = \sum_{j=1}^{N_i} D_{i,j}$, where N_i are i.i.d. If $C_{i,j}$ and $D_{i,j}$ belong to the same location-scale parameter, as described in assumption C1 with $(\mathbf{a}, \mathbf{s}) \preceq_{\text{RH}} (\mathbf{b}, \mathbf{t})$, then,

$$\mathbb{E} \left[\sum_{i=1}^n g(X_i) \right] \leq \mathbb{E} \left[\sum_{i=1}^n g(Y_i) \right] \quad (18)$$

for any convex function g .

When $C_{i,j}$ and $D_{i,j}$ belong to the same parametric family with the semigroup property with respect to convolution, as described in assumption C2, then equation (18) holds for any convex function g . In addition equation (11) holds for any convex function g .

Case 3: Heterogeneity both in the arrival process and claim amounts

When both the distributions of the number of claim arrivals and the claim amounts have the semigroup property as described in case 1 and assumption C2 respectively, then increasing heterogeneity both in the claim arrival parameters and in the claim amounts parameters increases the sum of the risks.

$(\lambda, \gamma) \preceq_{\text{RH}} (v, \eta)$ implies that for any convex increasing function g ,

$$\mathbb{E} \left[\sum_{i=1}^n g(X_i) \right] \leq \mathbb{E} \left[\sum_{i=1}^n g(Y_i) \right] \quad (19)$$

The results for the collective risk model yield that heterogeneity either in the claim amounts or in the arrival processes increases risk. For a proof of the results, see Frostig and Denuit [6].

Let $\text{VaR}[X; \alpha] = \inf\{x \in \mathbb{R} | \Pr[X \leq x] \geq \alpha\}$. $\text{VaR}(X)$ is a popular risk measure. Another risk measure that has more desirable properties than the VaR is TVaR , where $\text{TVaR}[X; \alpha] = \frac{1}{1-\alpha} \int_{\alpha}^{\infty} \text{VaR}[X; s] ds$. It can be shown (see Denuit *et al.* [7]) that $\text{TVaR}[X; \alpha] \leq \text{TVaR}[Y; \alpha]$ if and only if $\mathbb{E}[g(X)] \leq \mathbb{E}[g(Y)]$ for any nondecreasing convex function g .

Thus, as the vector of the parameters of the claim amount distributions is more heterogeneous, the portfolio becomes riskier in the individual model and in the collective risk model. Increasing the heterogeneity in the arrival process parameters increases the risk in the collective risk model; however, it decreases the risk in the individual risk model.

Heterogeneity in Life Insurance Portfolios

Unobserved heterogeneity in life insurance portfolios holds for a group of insured individuals with different backgrounds. Thus, different subgroups might have different survival life distributions. In this section we study the impact of this heterogeneity on the premium. Spreeuw [8] considered a portfolio with n contracts, $n \geq 2$ for one period, say a year. Let p_r be the probability that insured r survives after 1 year. Assume that all insureds pay the same premium $\Pi = \Pi(\mathbf{p})$ at the beginning of the period. At the end of the period a proportion δ of the mortality profit or loss is allocated to those who survived the period. Let D be the amount paid at the end of the period by

the insurer to the insured upon death. Assumption of zero expected loss implies that the premium $\Pi(\mathbf{p})$ is

$$\Pi(\mathbf{p}) = D \left(1 - \frac{(1 - \delta) \sum_{m=1}^n p_m}{n \left[1 - \delta \left[1 - \prod_{m=1}^n (1 - p_m) \right] \right]} \right) \quad (20)$$

Since $\prod_{j=1}^n (1 - p_j)$ is Schur decreasing, $\Pi(\mathbf{p})$ is also Schur decreasing. Then consider another portfolio with survival vector $\mathbf{q}$, where $\mathbf{q} \leq_{\text{maj}} \mathbf{p}$ implies that $\Pi(\mathbf{q}) \geq \Pi(\mathbf{p})$. Especially if $\bar{p} = \frac{1}{n} \sum_{i=1}^n p_i$ and $\bar{\mathbf{p}} = (\bar{p}, \dots, \bar{p})$, then $\Pi(\bar{\mathbf{p}}) \geq \Pi(\mathbf{p})$.

Thus, in the Spreuw model the premium decreases as the portfolio's heterogeneity increases. Hence, the hidden heterogeneity causes the insurer to overestimate the premium.

Similar results were obtained by Dahan *et al.* [9, 10]. They considered a heterogeneous portfolio with n contracts, $n \geq 2$, with survival probabilities as in the Spreuw model. A common death benefit D is paid at time $(k+1)/l$ if the insured dies in the interval $(k/l, (k+1)/l)$, $k = 0, \dots, l-1$, and survival benefit S is paid at the end of the year. The life pays a premium $\pi(\mathbf{p})$ at k/l as long as he is alive. Let ${}_t p_r$ be the probability that the r th insured survives up to age $x+t$. Let $f(\mathbf{p}) = \sum_{r=1}^n p_r$, and $g(\mathbf{p}) = \sum_{k=0}^{l-1} \sum_{r=1}^n (k/l) p_r \tilde{v}^k$, where $\tilde{v} = (1+i)^{-1/l}$, and i is the annual interest rate. The equation of value yields

$$\pi(\mathbf{p}) = -(1 - \tilde{v})Dg(\mathbf{p}) + \frac{m \cdot D + \tilde{v}^l \cdot (S - D) \cdot f(\mathbf{p})}{g(\mathbf{p})} \quad (21)$$

In order to calculate the premium one needs to estimate ${}_t p_r$, for $0 \leq t \leq 1$. The most known approximations in the actuarial literature are

1. Exponential approximation: ${}_t p_r = (p_r)^t$
2. Balducci approximation: ${}_t p_r = \frac{p_r}{(1-t)p_r + t}$
3. linear approximation: $1 - {}_t p_r = t(1 - p_r)$.

Under the exponential or Balducci approximations, ${}_t p_r$ is a concave function of p_r , while under the linear approximation, ${}_t p_r$ is linear in p_r and therefore both convex and concave. Since the sum of Schur decreasing functions is Schur decreasing, the function $g(\mathbf{p})$ is Schur decreasing. Therefore, assuming

the exponential or the Balducci approximations, equation (21) yields that the premium is Schur increasing in $\mathbf{p}$. Thus, the more heterogeneous the population (in the sense of majorization ordering), the higher the premium. When applying the linear approximation, the premium is unaffected by the heterogeneity (in the majorization ordering sense). Under different types of approximations presented by Jones and Mereu [11], and Dahan *et al.* [10] arrived at the opposite conclusion.

Dahan *et al.* [12] considered a portfolio containing n endowment insurance contracts issued for M years. They assumed that all insureds are at the same age at the policy issue time and that their survival probabilities belong to a parametric family of survival functions. Under some conditions on the force of mortality, they showed that as heterogeneity increases the premium increases. Thus, assuming homogeneous portfolio the insurer underestimates the premium. Dahan *et al.* [13] also studied the case of n endowment insurance contracts issued for M years, where the insureds are subject to the same mortality table but are at different ages at the policy issue date. In this case also, the premium increases as the heterogeneity of the insureds' ages increases.

Conclusion

We have demonstrated that heterogeneity in the distribution functions of the risks influence the risk measure of the corresponding portfolios. However, to see whether heterogeneity increases or decreases risk, one has to check the type and the source of heterogeneity. In the individual and collective risk models, we saw that heterogeneity in the claim amounts increases any nondecreasing convex risk measure. In the collective risk model, heterogeneity in the claim arrival process also increases risk. The contrary happens when increasing heterogeneity in the claim occurrence probabilities in the individual model.

As for life insurance portfolios, we saw that ignoring heterogeneity might cause the insurer to overestimate or underestimate the premium depending on the type and the source of increases in heterogeneity.

References

- [1] Denuit, M. & Frostig, E. (2006). Heterogeneity and the need for capital in the individual model, *Scandinavian Actuarial Journal* **2006**(1), 42–66.

- [2] Denuit, M. & Vermandele, C. (1998). Optimal reinsurance and stop-loss order, *Insurance: Mathematics and Economics* **22**, 229–233.
- [3] Marshall, A.W. & Olkin, I. (1979). *Inequalities : Theory of Majorization and its Applications*, Academic Press, New York.
- [4] Arnold, C. (1987). Majorization and the Lorentz order: a brief introduction, in *Lecture Notes in Statistics*, Springer-Verlag, Berlin.
- [5] Ma, C. (2000). Convex orders for linear combinations of random variables, *Journal of Statistical Planning and Inference* **84**, 11–25.
- [6] Frostig, E. & Denuit, M. (2006). Monotonicity results for portfolios with heterogeneous claims arrival processes, *Insurance: Mathematics and Economics* **38**, 484–494.
- [7] Denuit, M., Dhaene, J., Goovaerts, M. & Kaas, R. (2005). *Actuarial Theory for Dependent Risks*, John Wiley & Sons.
- [8] Spreeuw, J. (1998). Majorization order applied to a system of mortality profit distribution. Paper presented at the *Second International Congress on Insurance: Mathematics and Economics*, Lousannne.
- [9] Dahan, M., Frostig, E. & Langberg, N.A. (2002). Life insurance policies with statistical heterogeneous population, *Scandinavian Actuarial Journal* **3**, 212–222.
- [10] Dahan, M., Frostig, E. & Langberg, N.A. (2003). Analysis of heterogeneous endowment policies under fractional approximations, *Insurance: Mathematics and Economics* **33**, 567–584.
- [11] Jones, B.L. & Mereu, J.A. (2002). A family of fractional age assumptions, *Insurance: Mathematics and Economics* **27**, 261–276.
- [12] Dahan, M., Frostig, E. & Langberg, N.A. (2004). Insurance contracts portfolios with heterogeneous parametric life distribution, *Scandinavian Actuarial Journal* **6**, 431–447.
- [13] Dahan, M., Frostig, E. & Langberg, N.A. (2004). Insurance contracts portfolios with insured ages, *Insurance: Mathematics and Economics* **35**, 137–153.

Further Reading

- Hu, T. & Ruan, L. (2004). A note on multivariate comparisons of Bernoulli random variables, *Statistics and Probability Letters* **126**, 281–288.

Related Articles

Risk Classification in Nonlife Insurance

Risk Measures and Economic Capital for (Re)-insurers

Risk-Neutral Pricing: Importance and Relevance

ESTHER FROSTIG

Inferiority and Superiority Trials

A superiority clinical trial (*see Randomized Controlled Trials*) is often conducted to establish that an experimental treatment (e.g., a therapy, preventive agent, or medical device) is superior to a control treatment (e.g., an accepted “standard” treatment, or a placebo), by some measure of treatment effect or *efficacy*. A noninferiority clinical trial is typically conducted to establish that the treatment effect of an experimental treatment is not much worse than that of an accepted “standard” treatment or an active-control treatment [1]. When an experimental treatment is anticipated to be nearly as effective as an active-control treatment, it might be worth developing if it has other advantages such as reduced *toxicity* and/or cost. An equivalence trial (a two-way noninferiority trial) can be used to establish that an experimental treatment is neither much better nor much worse than an active-control treatment [1].

Superiority and noninferiority clinical trials are often conducted as *Phase III comparative clinical trials*. In these trials, subjects that meet the study criteria are recruited and often assigned to either the experimental treatment arm or the control treatment arm using *randomization* procedures. Sometimes, *blinding* or *masking* is implemented to ensure that the participating patients or both the participating patients and their physicians are unaware of the *treatment allocation* as it can have a strong effect on the observed treatment outcome. It is important to select appropriate randomization procedures and/or blinding methods in order to minimize the risk of leading to some kind of *bias in clinical research*.

Superiority Trials

A prominent example of a superiority trial is the Salk vaccine trial of 1954. This trial was conducted to assess the effectiveness of the Salk vaccine as a protection against paralysis or death from poliomyelitis as described in more detail by Meier [2]. In one part of this study, 401 973 children were randomly assigned to either a treatment arm injected with Salk vaccine or a placebo arm injected with a salt solution. The trial was double blind, that is, neither the

children nor the diagnosing physicians were aware of who had been given the Salk vaccine or the salt solution.

When designing such superiority clinical trials, one must guard against the risk of committing any of the two common types of errors in typical hypothesis-testing-based trial designs. A type I error is made if the study leads to a false conclusion that the experimental treatment is superior to the control and a type II error is made if the study fails to establish superiority when the experimental treatment is indeed superior in efficacy. The probabilities of type I and type II errors are often denoted as α and β , respectively. When a trial is designed with the proportion of events (e.g., the proportion of deaths in Salk vaccine trial) as the primary efficacy *endpoint*, the investigators often have some idea of the order of magnitude of the proportions they are studying. A difference $\Delta = p_E - p_C$, between the projected proportion of events p_E for the experiment treatment and p_C for the control treatment, that is worth detecting is often specified [3]. The required sample size (i.e., number of subjects needed) for the study to ensure desired power $(1 - \beta)$ for detecting a specified difference Δ among the two treatment arms, at a prespecified significance level α (or type I error), using a one-sided z -test, can be obtained as

$$n_C = \frac{\left\{ z_\alpha \sqrt{\bar{p}\bar{q}} \left(1 + \frac{1}{k} \right) + z_\beta \sqrt{p_C q_C + \frac{p_E q_E}{k}} \right\}^2}{(p_E - p_C)^2},$$

$$n_E = k n_C \quad (1)$$

where the sample size (n_E) for the experimental treatment arm is k times as large as the sample size (n_C) for the control arm, $q_C = 1 - p_C$, $q_E = 1 - p_E$, $\bar{p} = (p_C + k p_E)/(1 + k)$, $\bar{q} = 1 - \bar{p}$, and z_α and z_β are the upper percentiles of the standard normal distribution corresponding to α and β [4]. More details on superiority trial design and sample-size formulas with various efficacy endpoints can be found in Piantadosi [5] and Chow and Liu [6].

Frequently, a formal hypothesis test is conducted to assess if superiority is met on the basis of data from the superiority trial. In practice, a two-sided test is often recommended even though the investigators are only interested in a one-sided alternative hypothesis [3]. The true value of the efficacy parameter can be

estimated and the corresponding confidence interval obtained for each treatment arm. The magnitude of the efficacy differences among treatment groups can be evaluated using appropriate measures such as, difference in response rates, means, hazard rates, median survival times, etc., along with corresponding confidence intervals [3–6].

Recent advances in medical science have lead to effective treatments for many diseases. Thus, superiority trials are often required to compare experimental treatment to an existing effective treatment instead of a placebo. Much effort is also being devoted to conducting noninferiority trials for developing experimental treatment which, without being necessarily more effective than the standard treatment, can have other important advantages such as in reducing *toxicity* (see **Toxicity (Adverse Events)**) and/or cost, which is discussed next.

Noninferiority Trials

In a noninferiority clinical trial, it is common to specify a threshold for the margin of indifference when comparing the treatment effects between the experimental treatment and the control treatment [1]. For example, when comparing the proportion of success in the experimental treatment (P_E) with that of the control treatment (P_C), the noninferiority margin (or the margin of indifference), δ , is often defined as the maximum difference amount by which the experimental treatment is not considered clinically much worse than the control treatment if $P_E > P_C - \delta$. The noninferiority margin δ , should be prespecified and determined on the basis of clinical judgment [1, 7, 8].

One must guard against the risk of committing any of the two types of errors in typical hypothesis-testing-based noninferiority trial designs [8]. A type I error is made if the experimental treatment is falsely concluded as noninferior to the control. A type II error is made if one fails to conclude noninferiority when the experimental treatment is indeed noninferior compared to the control treatment. For a prespecified noninferiority margin δ , one can determine the sample size (the number of subjects) needed to ensure desired type I and II errors as discussed in Farrington and Manning [9], and Chow and Liu [6], among others.

Frequently, a hypothesis test is performed to assess if noninferiority is met. For example, when efficacy is

in terms of proportion of successes, one may perform a z -test with the following test statistic,

$$T_n = \frac{p_E - p_C + \delta}{\sqrt{\frac{\hat{P}_E(1 - \hat{P}_E)}{n} + \frac{\hat{P}_C(1 - \hat{P}_C)}{n}}} \quad (2)$$

where δ is the noninferiority margin, p_C and p_E are the observed efficacy proportions and $\hat{P}_C$ and $\hat{P}_E$ are the maximum-likelihood estimates of the efficacy proportion in the control and experimental treatment arms, respectively [6, 9]. Noninferiority can be claimed by comparing the observed value of T_n to the percentiles of the standard normal distribution. The magnitude of the efficacy difference between the treatment arms can be evaluated using appropriate measures such as, difference in response rates, means, hazard rates, median survival time, along with the corresponding confidence intervals [6, 9].

Noninferiority trials are frequently conducted to develop less toxic and/or less costly new treatments compared to an accepted active-control treatment. Thus, the active control should truly be effective. Otherwise, a treatment that is worse than placebo may be approved on the basis of the results of a noninferiority test. Furthermore, noninferiority trials often need a large sample size, thus are costly and may require many years to be completed. Therefore, it is important that the trial not only studies an experimental treatment, which offers meaningful advantages over the standard control, but also is feasible and can be completed in a timely fashion. Finally, proper *monitoring* of such trials is important to minimize the risk in denying patients an effective standard treatment.

References

- [1] International Conference on Harmonization (1998). E9: guidance on statistical principles for clinical trials, *Federal Register* **63**(179), 49583–49598.
- [2] Meier, P. (1989). The biggest public health experiment ever: the 1954 field trial of the Salk poliomyelitis vaccine, in *Statistics: A Guide to the Unknown*, 3rd Edition, J.M. Tanur, F. Mosteller, W.H. Kruskal, E.L. Lehmann, R.F. Link, R.S. Pieters & G.R. Rising, eds, Duxbury, California.
- [3] Fleiss, J.L., Levin, B. & Paik, M.C. (2003). *Statistical Methods for Rates and Proportions*, 3rd Edition, John Wiley & Sons.
- [4] Rosner, B. (2006). *Fundamentals of Biostatistics*, 6th Edition, Duxbury.

- [5] Piantadosi, S. (2005). *Clinical Trials: A Methodologic Perspective*, 2nd Edition, John Wiley & Sons.
- [6] Chow, S.-C. & Liu, J.-P. (2003). *Design and Analysis of Clinical Trials: Concepts and Methodologies*, 2nd Edition, John Wiley & Sons.
- [7] European Agency for the Evaluation of Medicinal Products, Committee for Proprietary Medicinal Products (2005). *Guideline on the Choice of the Non-inferiority Margin*, European Medicines Agency (EMA) at <http://www.emea.eu.int/pdfs/human/ewp/215899en.pdf>.
- [8] Blackwelder, W.C. (1982). "Proving the null hypothesis" in clinical trials, *Controlled Clinical Trials* **3**, 345–353.
- [9] Farrington, C.P. & Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with

null hypothesis of non-zero risk difference or non-unity relative risk, *Statistics in Medicine* **9**, 1447–1454.

Related Articles

Compliance with Treatment Allocation

Meta-Analysis in Clinical Risk Assessment

VANDANA MUKHI AND YONGZHAO SHAO

Influence Diagrams

Influence diagrams (IDs) are a very useful tool for representing decision problems, especially when the problems exhibit considerable symmetry. Unlike the decision tree (*see* **Decision Trees**), whose topology represents relationships between events and particular decisions taken, the ID represents relationships between measurement and state *random variables* and *decision variables*. Consequently, IDs exhibit the following characteristics:

- They typically have a topology that is much more compact than, for example, the decision tree – there usually being far fewer variables in a problem than events and decisions.
- They represent qualitative information that can be elicited from assertions in common language and do not need early specifications from the client of numerical values.
- Unlike the decision tree, they can express independencies between random variables in the description of a problem;
- Like the decision tree, they can be used as a framework for neatly encoding quantitative information about the decision process so that Bayes decision rules (*see* **Decision Modeling**) can be calculated.
- They can represent continuous, discrete, or mixtures of continuous and discrete problems equally well.

The nodes of an ID represent decision variables, measurement and state random variables, and utility functions, while the existence of a directed edge or arrow from one node N to another M suggests that the value of N might influence the value of M . A simple example of an ID is given in Figure 1. An experiment is performed, giving a realization $X = x$ before an action $a \in A$ need be taken. The decision maker (DM) has a utility function that is a function of a consequence C and a only. The prior distribution of C and the conditional distribution of $X|C$ have both been elicited from the DM. The value of C is not determined before the action a is taken.

Conventionally, actions are associated with a decision node $\square$, random variables by a node $\circ$, and utility functions by a node $\diamond$. The dependence of X on C is depicted by the arrow from C to X . Because

a can only be chosen as a function of X and not C , there is an arrow from X to a but not one from C to a . Finally, because U is a function of a and C but not X , there is an arrow from a to U and from C to U , but not from X to U . A formal definition of this construction is given below.

Note that unlike the decision tree, the ID is very compact. This is because, unlike the decision tree, it does not represent the state space of variables associated with the nodes in the graph. It does represent *dependencies* or influences between the nodes. These dependence relationships can be elicited directly and are valuable for interrogating model plausibility (*see* below) and for determining simple subclasses of optimal decision rules (*see* [1, 2]). They provide a useful framework around which efficiently optimal decision rules can be calculated.

The genesis of IDs was a rather complicated one. Bayes nets (BNs), which represent various dependence relationships between random variables (but not actions or utility functions), by directed acyclic graphs (DAG) were being studied with some vigor, first by mathematical biologists and then by statisticians, by the 1980s [3, 4]. Meanwhile, several researchers in Artificial Intelligence, including Pearl, had developed similar graphical tools for representing dependence structures (*see* [5] for references). Parallel to this, decision analysts developed IDs to represent decision problems [6]. Pioneering work by Shachter [7] noted the connection between the two fields – where the BN could be seen as a special case of an ID having no decision or utility nodes – since when the field has become very cross disciplinary. Increasingly fast computational algorithms have been developed for BNs over the last 20 years. Many such algorithms have been implemented in code and these are now adapted to decision problems as well as inference problems (*see*, for example [8, 9]).

In this article, rather than describing the technical features of fast algorithms for easy calculation of Bayes decision rules, which is probably not central to the interests of this audience and is well documented elsewhere [1, 8–10], the focus is on the use of IDs for *problem representation*. Of course, once a problem has been faithfully represented as an ID, the algorithms mentioned above can be implemented and calculations made.

Partly because of the cross disciplinary nature of their development, the semantics of IDs is still not fully agreed upon. Here, we most closely follow the

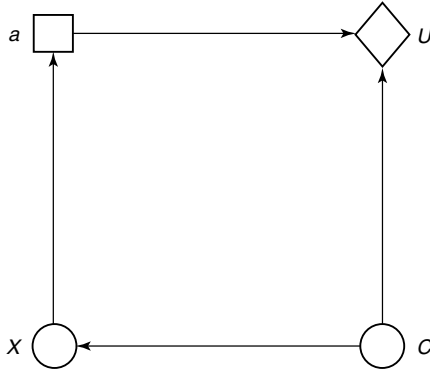


Figure 1 A simple influence diagram

notation of Shachter. It is simplest to first consider the BN (sometimes also called the *Bayesian network*, the *chance influence diagram*, or *probability directed acyclic graph* (DAG)) and only after this, the ID itself.

Irrelevance and Conditional Independence

One key element of any qualitative description of a problem is whether two uncertain measurements X and Y are related. Only if they are related to one another is it necessary to quantify *how* they are related. If learning the value of X is irrelevant to predictions of Y , then this can be expressed probabilistically by stating that the probability density or mass function $p_{X,Y}(x, y) = p_X(x)p_Y(y)$, where $p_X(x)$ is the marginal mass function of X and $p_Y(y)$ is the marginal density/mass function of Y . Note that from this observation we can make the deduction that if X is irrelevant to predictions of Y then, by the symmetry of the relation above, Y must be irrelevant to predictions about X – a deduction not immediately apparent from the statements in common language.

Independence is not quite a rich enough concept to be of much practical use. However *conditional* independence is. Furthermore, like independence, it can also be expressed in common language. Thus, suppose the DM asserts that once she knows the value of Z , whatever that value of Z is, the value of X is then irrelevant to any predictions she might make concerning Y . Then, if she is a Bayesian and therefore thinking probabilistically, we can translate this statement as asserting that Y is independent of X given Z . This, in turn, is equivalent to her asserting that her subjective conditional density or mass

function $p_{Y|Z,X}(y|x, z)$ is not an explicit function of x . We write this conditional independence assertion as $Y \perp\!\!\!\perp X|Z$.

It is easily shown that if $Y \perp\!\!\!\perp X|Z$ and if $p_{X,Y,Z}(x, y, z)$ is the joint density/mass function of (X, Y, Z) , then

$$p_{X,Y,Z}(x, y, z) = p_{Y|Z}(y|z)p_{X|Z}(x|z)p_Z(z) \quad (1)$$

where $p_{X|Z}(x|z)$, $p_{Y|Z}(y|z)$ are respectively the densities of X given Z and Y given Z , and $p_Z(z)$ is the marginal density/mass function of Z .

Conditional independence has two properties that translate directly into statements about conditions a Bayesian should obey while considering irrelevance. The first is *symmetry*, which is a generalization of the symmetry we observed for independence. Thus, for any three vectors of measurements $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ we must have that the statements $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}$ and $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$ are equivalent. Thus, if the Bayesian DM states that given the value of $\mathbf{Z}$, $\mathbf{X}$ provides no further useful information about $\mathbf{Y}$, then she also should believe that given the value of $\mathbf{Z}$, $\mathbf{Y}$ provides no further useful information about $\mathbf{X}$.

A second useful *decomposition* property is the following: if $(\mathbf{W}, \mathbf{X}, \mathbf{Y}, \mathbf{Z})$ are random vectors, then the statement $\mathbf{Z} \perp\!\!\!\perp (\mathbf{X}, \mathbf{Y})|\mathbf{W}$, is logically equivalent to the two statements (asserted together) that $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y}|\mathbf{W}$ and $\mathbf{Z} \perp\!\!\!\perp \mathbf{X} |(\mathbf{W}, \mathbf{Y})$. Pearl illustrates why this equivalence should be expected to hold not only for probabilistic systems but also for any reasonable semantics of irrelevance using a simple example. Suppose your current relevant knowledge about $\mathbf{Z}$ is encapsulated in the random vector $\mathbf{W}$. You now read a book in two volumes, which might be informative about $\mathbf{Z}$. Let the information in the first volume be $\mathbf{Y}$ and the information in the second volume be $\mathbf{X}$. The assertion $\mathbf{Z} \perp\!\!\!\perp (\mathbf{X}, \mathbf{Y})|\mathbf{W}$ says that, given what you already know $\mathbf{W}$, the two volumes give no additional information relevant to any predictions about $\mathbf{Z}$. This surely implies that $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y}|\mathbf{W}$ – given what you already know, $\mathbf{W}$, the first volume gives no additional information relevant to any predictions about $\mathbf{Z}$; and also $\mathbf{Z} \perp\!\!\!\perp \mathbf{X} |(\mathbf{W}, \mathbf{Y})$ – given what you already know, $\mathbf{W}$, and the information in the first volume, the second volume gives no additional information relevant to any predictions about $\mathbf{Z}$. Conversely, if either one of $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y}|\mathbf{W}$ and $\mathbf{Z} \perp\!\!\!\perp \mathbf{X} |(\mathbf{W}, \mathbf{Y})$ is not true, then surely the DM must believe that the pair of volumes, taken together, have some relevant information about $\mathbf{Z}$, additional to $\mathbf{W}$.

The point is, then, that qualitative dependence relations made by a Bayesian DM must obey certain rules (the ones described above are called the *semigraphoid axioms* [5, 11, 12]). If elicited statements do not obey these rules, then they are self contradictory. The analyst then needs to help the DM to explore the reasons for these apparent contradictions and help her develop her model so that these contradictions disappear. One big practical problem is how the analyst can converse with the DM in such a way that she can fully engage in the development of the model. Only then can she take ownership of the resulting model. For example, she may well be scared off by the sort of notation used above. There is a way of doing this when the model describes the outworkings of a process – the BN. On the one hand, this graph depicts the process in an entirely natural way; however on the other, it is an entirely formal construction and so supports logical interrogation.

Representing Dependencies Using a Bayes Net

For simplicity, assume that our problem has a discrete formulation and let the n variables of interest be denoted by $\mathbf{X} = (X_1, X_2, \dots, X_n)$, these having a joint probability mass function $p(\mathbf{x})$. Then the usual rules of probability tell us that we can calculate

$$p(\mathbf{x}) = p_1(x_1)p_2(x_2|x_1)p_3(x_3|x_1, x_2) \dots \times p_n(x_n|x_1, \dots, x_{n-1}) \quad (2)$$

Suppose that the ordering of these variables follows the natural development of the process in the sense that the events measured by X_j do not happen after those measured by X_i if $j < i$. Quite often, the DM may be initially hesitant about specifying a functional form for $p_i(x_i|x_1, \dots, x_{i-1})$ in this formula. However, she is often quite happy that, for each X_i , $2 \leq i \leq n$, given the values of the subset $Q_i(\mathbf{X})$ – called the *parents* of X_i – of the variables $\{X_1, \dots, X_{i-1}\}$ listed before X_i , the values of the remaining variables $R_i(\mathbf{X})$ listed before X_i are irrelevant for forecasting X_i . In the notation of the last section, this simply states that

$$X_i \perp\!\!\!\perp R_i(\mathbf{X}) \mid Q_i(\mathbf{X}) \quad (3)$$

where

$$Q_i(\mathbf{X}) \subseteq \{X_1, \dots, X_{i-1}\} \quad (4)$$

$$R_i(\mathbf{X}) = \{X_1, \dots, X_{i-1}\} \setminus Q_i(\mathbf{X}) \quad (5)$$

In terms of the probability breakdown, when $X_i \perp\!\!\!\perp R_i(\mathbf{X}) \mid Q_i(\mathbf{X})$, this tells us that $p_i(x_i|x_1, \dots, x_{i-1})$ can be written as an explicit function of $(x_i, Q_i(\mathbf{x}))$.

This set of statements about the relationships between $\{X_1, X_2, \dots, X_n\}$ can now be written as a DAG, G . The vertices or nodes of this graph are labeled by the set $\{X_1, \dots, X_n\}$ of variables describing the problem. We connect a directed edge from $X_j \rightarrow X_i$ if and only if X_j is a *parent* of X_i , that is $X_j \in Q_i(\mathbf{X})$. Notice the simplicity of this representation: a variable remains unattached to X_i only if it does not provide its own unique useful information, which adds to that given by the parents of X_i .

1. A BN is said to be *valid* if all the irrelevance statements (equation 3) that can be read in its DAG are believed by the DM.
2. A DAG of a BN is always acyclic because, by definition, parents are listed before children. Note that there are (albeit less powerful) analogous potentially cyclic graphs (e.g. [13]) that relate to simultaneous equation models.
3. It is important to realize that it is the *missing* edges in a DAG G of a BN that are significant. In particular, the DAG where all nodes are connected to each other is completely uninformative about the irrelevance relationships between its variables.
4. The *d*-separation theorem – first proved by Verma and Pearl [5, 9, 11, 14] – allows for *all* valid deductions to be read directly off the DAG. It relies only on the graphoid properties discussed above. This result enables the DAG of a BN to be used to query whether a given irrelevance statement of the form $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$, where $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are subsets of the variables of the problem – as represented as vertices of G – can be logically deduced from the DM's original irrelevance statements. This is very useful because the analyst can take the statements given by the DM and confront her with statements that can be logically deduced from what she has said. Only if these deductions look plausible to her can

the model be requisite [15]. If the DM does not believe them, then she should be able to explain why. On the basis of this explanation, the analyst can then support the DM in modifying her BN – see the example below. The modified DAG of the new BN contains either more relationships (edges) or more variables (vertices) than the original. Briefly, to check the validity of $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{Z}$, where the subsets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are all disjoint from one another the analyst acts as follows:

- Takes the subgraph G' of G , whose vertices include only those variables $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$ in the query, together with those other variables – represented as vertices in G – that appear *before* a variable in $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$ in G . Let there be a directed edge in G' between two of its vertices if and only if that directed edge appears between the corresponding edges in G .
 - Joins each pair of unconnected parents of vertices in G' by an undirected edge to form a new (mixed) graph G'' .
 - Substitutes each directed edge in G'' by an undirected edge to form an undirected graph G''' .
 - If every path in G''' between any vertex X_i in $\mathbf{X}$ and any vertex X_j in $\mathbf{Y}$ passes through some vertex X_k in $\mathbf{Z}$, then it can be deduced that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{Z}$, i.e., that the DM implicitly believes that, once the values of variables in $\mathbf{Z}$ are known, the values of variables in $\mathbf{X}$ give no further information relevant to prediction about $\mathbf{Y}$. On the other hand, if there is a path in G''' between an X_i in $\mathbf{X}$ and an X_j in $\mathbf{Y}$ that does not pass through any of the variables in $\mathbf{Z}$, then there is at least one probability distribution consistent with the elicited irrelevance statements such that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{Z}$ is not true – i.e., where $\mathbf{X}$ provides further information to $\mathbf{Z}$ about $\mathbf{Y}$. More details of this result, including a proof, can be found in [14], with simpler proofs being available in [11] and [9].
5. The conditional independence structure encoded in G can be used as an efficient framework for storing elicited conditional probabilities, and when there are significant numbers of elicited irrelevances, this framework is usually far more efficient than, for example, the tree. Furthermore, the framework can also often be used to quickly

ensure the fast calculation of margins of interest even in very high dimensional problems, provided the number of parents of all its nodes are small. Details of such algorithms can be found in [9] and [8].

Consider the following grossly simplified but, nevertheless, illuminating example of how eliciting a BN can help the DM express important features of an underlying process that might influence its development.

Example 1 A utility company needs to determine an optimal portfolio of its programs of work. For any promising scheme, the protocol the company produces provides a rough estimate R of the cost of the work. On the basis of this, it may proceed to obtain a more detailed estimate D from its civil engineers. The program is then put out to tender, and the firm placing the lowest bid T receives the work. The work is then completed and the final outturn cost O – which may include any unforeseen costs – is charged. The company needs to monitor its current evaluation of the outturn cost O so that it can appropriately balance financial risks between this and other projects.

The company currently uses the detailed estimate to arrive at the probability distribution of the tender price, and the outturn price is estimated using the winning tender bid T . It follows that they are implicitly assuming that the BN with DAG given in Figure 2 is valid.

To check this model, the analyst can ask two questions related to $T \perp\!\!\!\perp R | D$ and $O \perp\!\!\!\perp (R, D) | T$, respectively, which can be stated in words as follows:

- “Knowing the detailed estimate, can you think of any scenario where also knowing the rough estimate would assist in refining an estimate for the tender price?” and
- “Are there any scenarios in which the detailed or rough estimate might provide additional useful information about the outturn price over and above that provided by the winning tender price?”

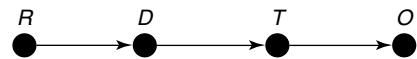


Figure 2 An initial DAG for Example 1

When confronted with the second query, it is noted that for projects where the tender price is much less than the detailed estimate, the outturn price is often much higher than would normally be estimated from the tender price. The company can also provide an explanation for why this is so. In periods of economic recession, work is hard for the contractors to find. In such circumstances, it is not unusual for a firm to place a low tender to secure the work and then endeavor to find many “add-ons”, so that the outturn price they finally charge is economic to them.

Note from this example that presenting the company with the BN has provided a qualitative framework from which it has modified its description of the problem: incorporating a new variable E , representing the economic environment into the description. The new BN is now given in Figure 3.

If relevant ways of measuring E can be found, then more reliable estimates of the outturn price based on T and E should be forthcoming. Notice that, using the semigraphoid properties directly, other queries derivable from the d -separation theorem could be fed back and checked for their plausibility. So, for example, the analyst could ask “If you had to resurrect the rough estimate of a project because this had been lost, and you had both your detailed estimate and the tender bid available, then would the detailed estimate alone be sufficient to resurrect the rough estimate as accurately as possible?” This should be so, because the symmetry property demands that – as the DM has asserted above – $T \perp\!\!\!\perp R|D$, then it can be deduced that $R \perp\!\!\!\perp T|D$, i.e., that the answer to the above question is “yes”. However, note that this question is not obviously equivalent to the original assertion. So, other forgotten dependencies might be teased out of the DM just by asking this question. Similarly, to check whether our model now implies $O \perp\!\!\!\perp D|T$, use the transform G to G''' as described above (Figure 4).

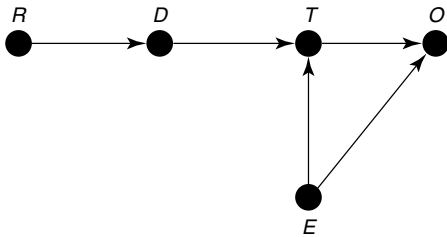


Figure 3 A modified BN

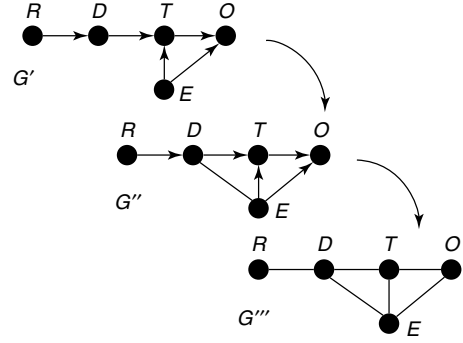


Figure 4 Checking independence using a moralized BN

Note that the path (D, E, O) in G''' between D and O is not blocked by T and so we can conclude that, if E is not known for this project, then the detailed estimate D may well be informative about the outturn price O . However, if we have recorded both T and the measure E of the economic environment, then note that (T, E) block all paths between D and O , so under the adapted graph, D is assumed unnecessary to record, if both (T, E) are recorded.

The Semantics of Influence Diagrams

An ID is a generalization of a BN, extended to include both decision variables and utilities. These IDs are useful for coding a particular but common type of decision problem. Such “uniform” problems have two characteristics. First, they can be formulated so that each possible decision rule $d(\mathbf{X}) \in A$ – the set of possible decision rules – can be expressed as a totally ordered sequence of subdecisions $(d_1(\mathbf{X}), d_2(\mathbf{X}), \dots, d_k(\mathbf{X}))$, where $d_j(\mathbf{X}) \in A_j : 1 \leq j \leq k$, and where the DM knows and remembers the decisions $\{d_1, d_2, \dots, d_{j-1}\}$ she has acted out or committed to when she commits to $d_j(\mathbf{X})$, $2 \leq j \leq k$. We assume here that the commitment to any acts $\{d_1, d_2, \dots, d_{j-1}\}$ does not constrain the possible choice $d_j(\mathbf{X}) \in A_j : 2 \leq j \leq k$. Technically, this can be written as the condition that $A = A_1 \times A_2 \times \dots \times A_k$, where the *decisions variables* $d_1(\mathbf{X}), d_2(\mathbf{X}), \dots, d_k(\mathbf{X})$ are allowed to take any value in their range.

The next issue is to code what information is available when one of these decision variables has to be enacted. Order the components of $\mathbf{X}$ so that they are consistent with the order in which they

happen under a particular reference decision rule $d(\mathbf{X}) \in A$. Let $T_1(\mathbf{X}|d(\mathbf{X}))$ denote the set of random variables that is the union of the set $Q'_1(\mathbf{X}|d_1(\mathbf{X}))$ of all the random variables whose values are known before $d_1(\mathbf{X})$ is enacted together with all variables that measure events that have happened before variables in $Q'_1(\mathbf{X}|d_1(\mathbf{X}))$, but are still unknown to the DM when she commits to $d_1(\mathbf{X})$. Similarly for $i = 1, 2, \dots, k$, let $T_i(\mathbf{X}|d(\mathbf{X}))$ denote the union of the set of variables that have happened which contains variables that happen between when $d_j(\mathbf{X})$ is taken and $d_{j-1}(\mathbf{X})$ is taken and $Q'_j(\mathbf{X}|d_j(\mathbf{X}))$, $j = 1, 2, \dots, k$. Finally, let $T_{k+1}(\mathbf{X}|d(\mathbf{X}))$ denote the remainder of the random variables that have not yet appeared as components of $\{T_1(\mathbf{X}|d(\mathbf{X})), \dots, T_k(\mathbf{X}|d(\mathbf{X}))\}$. Thus, $T_{k+1}(\mathbf{X}|d(\mathbf{X}))$ contains a subset of the variables that remain unknown for any decision rule $d(\mathbf{X}) \in A$ and happen after the DM has committed herself to all aspects of the decision process. Our second condition can now be stated. The partition of variables $\{T_i(\mathbf{X}) : 1 \leq i \leq k+1\}$ defined above must not depend on the DM's choice of $d(\mathbf{X}) \in A$ and so satisfies the condition above for *all* possible $d(\mathbf{X}) \in A$ (not just the reference decision). This second condition demands that whatever value a subdecision is set to, it does not disrupt the process in a way that influences the *order* in which the random variables describing the process happen. This order is the same regardless of the decisions taken – although setting subdecision variables to different values may well influence the *value* a subsequently occurring variable takes. Here, we call a decision problem *uniform* when it satisfies the two conditions stated above.

Example 2 A company plans to launch a cosmetic product, but are concerned that one of its constituents might provoke an attributable allergic reaction in some users if its concentration is higher than a currently unknown threshold B . If the company launches the product when it is possibly allergenic, then they will have to withdraw the product with a loss of profit and reputation. However, the higher the concentration of this constituent, the more effective the product and so the higher their predicted sales and also their reputation. They, therefore, plan to perform tests of the constituent. Their first subdecision is the duration l of the experiment. The longer the experiment, the better the quality of information X about the potential to cause an allergic reaction as a function of concentration, but the lesser the profit

from the product, since competing products marketed by others can be expected to appear in the medium term. The second subdecision is the concentration c to use in the final marketed product.

It is easy to check that this is a uniform decision problem in which the number of subdecisions $k = 2$. The danger threshold of the concentration waits to be discovered, so B is our first variable, whilst the second variable X is known before the second subdecision c . So in the notation above, $T_1 = X_1 = B$, $d_1 = l$, $T_2 = X_2 = X$, $d_2 = c$, and T_3 is empty.

Note that, for a uniform decision problem, any decision rule $d(\mathbf{X}) \in A$ can be specified in terms of an ordered sequence of subdecisions $(d_1(\mathbf{X}), d_2(\mathbf{X}), \dots, d_k(\mathbf{X}))$. Furthermore, when the DM commits to an act $d_j(\mathbf{X})$, this will be a function of a subset $Q'_j(\mathbf{X}|d_j(\mathbf{X}))$ of the variables in $X_j \in T^{(j)}(\mathbf{X}) = \bigcup_{i=1}^j T_i(\mathbf{X})$ – the variables known by the time $d_j(\mathbf{X})$ is enacted – together with her acts $d^{(j-1)} = \{d_1, d_2, \dots, d_{j-1}\}$ so far committed to and remembered, $2 \leq j \leq k$. Let

$$Q_j = \bigcup \{Q'_j(\mathbf{X}|d_j(\mathbf{X})) : d(\mathbf{X}) \in A\} \cup \{d_1(d(\mathbf{X})), d_2(d(\mathbf{X})), \dots, d_{j-1}(d(\mathbf{X}))\} \quad (6)$$

and call this set Q_j the *parents* of d_j . The parents of d_j are then simply the set of variables that have been seen and decisions that have been enacted so far, by the time $d_j(\mathbf{X})$ needs to be chosen.

Fortunately, a decision problem is often uniform. When this is so, it is possible to represent many of its qualitative features by drawing a very evocative and useful graph, similar to the BN. To see this, first index the variables defining the problem so that

$$T_1(\mathbf{X}) = \{X_1, X_2, \dots, X_{i(1)}\} \quad (7)$$

$$\vdots$$

$$T_2(\mathbf{X}) = \{X_{i(1)+1}, X_{i(1)+2}, \dots, X_{i(2)}\} \quad (8)$$

$$T_j(\mathbf{X}) = \{X_{i(j-1)+1}, X_{i(j-1)+2}, \dots, X_{i(j)}\} \quad (9)$$

$$\vdots$$

$$T_{k+1}(\mathbf{X}) = \{X_{i(k)+1}, X_{i(k)+2}, \dots, X_n\} \quad (10)$$

Notice that by construction, the random variables and the decision variables are now listed in an order consistent with how the DM believes the historical

ordering in which events happen and decisions are taken, i.e.,

$$\begin{aligned} &X_1, X_2, \dots, X_{i(1)}, d_1, X_{i(1)+1}, X_{i(1)+1}, \dots, \\ &X_{i(2)}, d_2, \dots, X_{i(k-1)+1}, X_{i(k-1)+2}, \dots, X_{i(k)}, d_k, \\ &X_{i(k)+1}, X_{i(k)+2}, \dots, X_n \end{aligned} \quad (11)$$

For each $d(\mathbf{X}) \in A$, it is now possible to write down a factorization of densities, valid for all $d(\mathbf{X}) \in A$ and consistent with the ordering above that in turn can form the basis of a graphical representation of a decision problem. Thus, let $p(\mathbf{x}, u|d(\mathbf{X}))$ denote the joint mass function or density over the variables $\mathbf{X} = (X_1, X_2, \dots, X_n)$ and u the utility function, given that the DM takes the decision rule $d(\mathbf{X})$. By the familiar conditioning rules,

$$p(\mathbf{x}, u|d(\mathbf{X})) = p(u|\mathbf{x}, d(\mathbf{X})) \cdot p(\mathbf{x}|d(\mathbf{X})) \quad (12)$$

where

$$\begin{aligned} p(\mathbf{x}|d(\mathbf{X})) &= p_1(T_1(\mathbf{x}))p_2(T_2(\mathbf{x})|T_1(\mathbf{x}), d_1(d(\mathbf{X}))) \dots \\ &\quad p_j(T_j(\mathbf{x})|T^{(j-1)}(\mathbf{x}), d^{(j-1)}(d(\mathbf{X}))) \dots \\ &\quad p_k(T_k(\mathbf{x})|T^{(k-1)}(\mathbf{x}), d^{(k-1)}(d(\mathbf{X}))) \\ &\quad p_{k+1}(T_{k+1}(\mathbf{x})|T^{(k)}(\mathbf{x}), d(\mathbf{X})) \end{aligned} \quad (13)$$

Here, for $j = 1, 2, \dots, k+1$ the term $p_j(T_j(\mathbf{x})|T^{(j-1)}(\mathbf{x}), d^{(j-1)}(d(\mathbf{X})))$ is simply the joint density or mass function of the subset of variables $T_j(\mathbf{x})$ conditional on events that have already happened and all decisions $d^{(j-1)}(\mathbf{x})$ committed to under $d(\mathbf{X})$ before these variables. These component densities can be factorized further. Thus

$$\begin{aligned} &p_j(T_j(\mathbf{x})|T^{(j-1)}(\mathbf{x}), d^{(j-1)}(d(\mathbf{X}))) \\ &= \prod_{s=i(j-1)+1}^{i(j)} p_{j,s}(x_s|T^{(j-1)}(\mathbf{x}), x_{i(j-1)+1}, \dots, x_{s-1}, \\ &\quad d^{(j-1)}(d(\mathbf{X}))) \end{aligned} \quad (14)$$

Let the *parents* Q_s of X_s be defined as the set of arguments $\{T^{(j-1)}(\mathbf{x}), x_{i(j-1)+1}, \dots, x_{s-1}, d^{(j-1)}(d(\mathbf{X}))\}$ that appear *explicitly* in the conditional mass function/density $p_{j,s}(x_s|T^{(j-1)}(\mathbf{x}), x_{i(j-1)+1}, \dots, x_{s-1}, d^{(j-1)}(d(\mathbf{X})))$, for at least one $d(\mathbf{X}) \in A$. The parents of X_s are then simply those variables and

subdecisions that, at least for some possible decision rules happening earlier in the listing, might affect the distribution of X_s . Note this set of parents can be discovered directly from the DM, without reference to anything quantitative.

Finally, let the *parents* of the DM's utility function u be the subset of those components of $\mathbf{X} \cup \{d_1(\mathbf{X}), d_2(\mathbf{X}), \dots, d_k(\mathbf{X})\}$ that appear *explicitly* as arguments of u . Again note that the parents of u can be directly elicited as those features of the problem that might influence how well the DM assesses she has performed after everything has happened.

Note that for complicated problems in which the order decisions need to be made is clear, to simplify the ID, it is suggested in [1] that we omit the edges from one decision node to another – this is sometimes called the *no forgetting* construction.

Once the analyst has elicited the parents of decision variables, random variables, and utilities in the DM's formulation of her problem, a graph can be drawn that depicts the interrelationships between all the variables in the problem and provides a framework within which the DM's Bayes decision rule can be calculated.

Drawing and Analyzing an Historic Influence Diagram

An ID has three sorts of nodes as its vertices: chance nodes (represented by $\circ$), decision nodes (represented by $\square$), and utility nodes (represented by $\diamond$). In the notation above, the chance nodes are the ordered components $X_i, 1 \leq i \leq n$, the decision nodes are the subdecisions $(d_1, d_2, \dots, d_k)$, and the utility node is U . To draw the ID I – here called the *historic* ID – simply draw a connected edge from each parent of a vertex to that vertex, where a node's parents are defined above. By doing this, we simply connect each vertex, whether a chance, decision, or utility node to other factors that might have “influenced” it. Thus, if the vertex is a random variable, it is attached by an edge from random variables and subdecisions that might influence its distribution. If the vertex is a decision node, then a directed edge attaches to it from all variables and decisions that are known to the DM – and therefore, possibly taken account of – at the time she has to commit to her subdecision. Finally, all the variables and decisions influencing the evaluation of the efficacy of a chosen decision rule connect into the utility node.

The example in the section titled “The Semantics of Influence Diagrams” has a historic ID I shown in Figure 5.

Being first, the concentration B has no parents and the decision I also has no parent because B is unknown when the size of the experiment is decided upon. The test results X depend both upon the length of the experiment and the toxicity of the constituent, and so has parents (I, B) . The decision c about the concentration to use in the product can be made having observed X and taking into account the resolution I of the experiment, so it has parents (I, X) . Finally, the utility has parents I because the longer the experimental period the greater the loss, and also c and B , since both profits and reputation depend upon marketing a product with as high as possible safe concentration.

There are various observations we can make from this simple graph:

1. Note that this expresses pictorially and very directly the explanation given above and the semantics are easy to explain to the DM. Notice also that unlike the decision tree, there is no problem that the variables in the problem may be continuous.
2. Like the BN, this graph provides an elegant framework around which to store elicited probabilistic information. Thus, the two probability tables here are the margin of B : the DM’s prior probability of the highest safe concentration and the probability distribution of the test results X , given the value of B and the length of the experiments (i.e., given all the different configurations of the pairs of value of its parents).

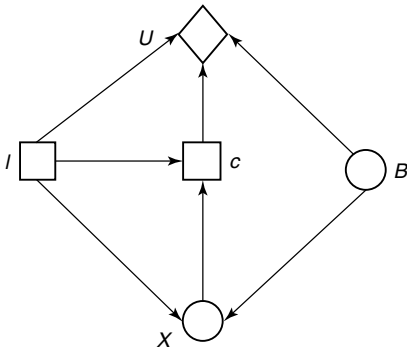


Figure 5 A historic influence diagram for Example 2

3. IDs can often be drawn to different levels of resolution. For example, in the illustration above, we could include explicitly as two chance nodes the profit and the reputation of the company or more information could be represented about both the underlying toxicity, and the test result X . In practice, both of these would often be vectors whose components could be vertices of I and whose lack of mutual dependence structure could be coded through missing edges in the graph. Typically, as for the ID described above, such encoded dependence structures can then be used to interrogate the model, elegantly store further elicited information, and also be used in fast computational algorithms for identifying Bayes decision rules. However, there is obviously a trade-off here between the detail presented and the transparency of the graph. Care also needs to be taken that the elaborated decision problem remains uniform.
4. The description of a decision problem can sometimes be dramatically simplified once the dependence structure becomes clear. The easiest such simplification to perform is when, on drawing an ID, you notice that a decision or chance node – called a *barren node* in the diagram – is the parent of no other node. It can be then proved (see [1]) that, for purposes of identifying a Bayes decision rule, we can equally work with the ID that omits this barren node and its connected edges.
5. As more vertices and edges are added, the diagram starts to become opaque. Various authors suggest different methods for simplifying the ID. For example, Shachter [1] suggests omitting edges between decision nodes (the no *forgetting convention*), while Dawid [16] foregoes the inclusion of any parents of decision nodes. However, these modifications make the graph less expressive. The ID can also be made more expressive. Thus analogs of the d -separation theorem for BNs can be used to determine whether a subdecision associated with a Bayes decision rule *needs* to depend on a particular variable (decision or chance) whose values have been observed or whether it can be ignored. In the latter case, edges can be dropped from the irrelevant variable into that decision (see e.g. [2, 17]); then new barren nodes can appear that can be deleted from the ID.

Extensive Form Influence Diagrams and Calculating a Bayes Rule

An ID can also be used as a framework for rollback and thus for discovering a Bayes decision rule. However, the historic influence diagram is not appropriate for this calculation, in general, because it is constructed consistently with the order variables are believed to happen rather than the order in which they are observed. So – like the causal decision tree – the historic ID is a great tool for problem representation and for storing elicited prior information (see **Decision Trees**). However, to be useful as a framework for backward induction, it needs to be transformed so that it is more like the dynamic programming tree (see above).

To draw an *extensive form ID* from an historic ID, we need to reorder the components of $\mathbf{X}$ so that they are consistent with the order in which they *become known* to the DM. In particular, this reordering must ensure that

1. as with historic IDs, any variable X_i known when d_i is taken appears before d_i in the analogous conditioning order (equation 11), $1 \leq i \leq k$;
2. any component variable X_s whose value is *unknown* when d_i is taken, appears *after* d_i in the analogous conditioning order (equation 11) (even if it has actually happened before this decision).

Typically, to transform an historic ID to an extensive form ID, we simply need to permute the order of appearance of some of the component chance variables, using the law of total probability and Bayes rule. Note that reversing conditioning sometimes induces further dependencies between variables in the problem.

By doing this, we obtain a new graphical representation with the same vertices, some edges sometimes reversed, and with some additional edges into chance nodes, sometimes added. The extensive form ID of the running example is given in Figure 6.

Notice that, in this case, the extensive form ID is identical to the historic ID except the directed edge from B to X has been replaced by one from X to B , i.e., this arc has been reversed. The distribution of X given each decision I and the distribution of B given X , both need to be calculated using the law of total probability and Bayes rule, making use of the densities encoded in the historic ID.

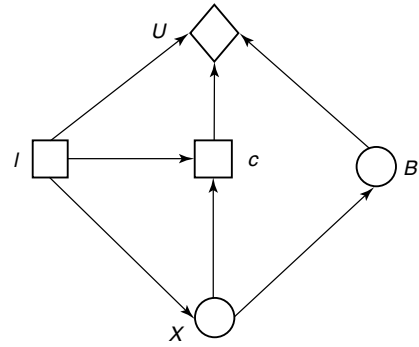


Figure 6 An extensive form influence diagram

A Bayes decision rule can now be calculated using an analogous algorithm to the rollback algorithm described above for decision trees (see **Decision Trees**). This algorithm again uses the existence of a Bayes decision rule that assumes that from all situations arrived at in a decision sequence, future decisions can be supposed to be expected utility maximizing. We note the following:

- Being rather technical, a precise description of the rollback algorithm for extensive form IDs is beyond the scope of this article, but is described in [1]. Chapter 8 in [9] contains a good review of these methods. It has subsequently been discovered that sometimes it is possible to evaluate a Bayes decision rule more efficiently using methods not based on rollback [8]. However, for the practitioner it suffices to know that software algorithms are available to automatically calculate a Bayes decision rule for any given historic ID.
- It is actually possible to generalize the construction above so that it applies to a somewhat larger class of problems than those discussed above (see [2, 9]), although we would then need to be more careful how the objects in the graph are defined to keep the construction formal. It is also often possible to construct extra variables artificially and so construct a uniform decision problem from one that is not – see below. However, this latter type of constructive approach can introduce new technical difficulties and loses some of the interpretable appeal of the graph.
- There are other methods for depicting decision problems in an equally formal way (see [9]).

Causal Bayes Nets

There has recently been much interest in the study of models that are called *causal*. Several authors subsequently realized [16, 18] that such models could be seen as a particular example of the historic ID described above. Consider the following hypothetical scenario.

Example 3 Males diagnosed with low sperm count can take an available fertility treatment F . The treatment – a one-off course of pills – is known to sometimes increase the sperm count S of some men. However, taking the treatment has a risk: in a small number of individuals, taking F increases the probability of high blood pressure B in later life, leading to a higher risk of a heart attack H . Expert judgment is that across the population of males being administered the drug, the DAG of the BN below (Figure 7) is valid, with an associated factorization formula of the joint mass function $p(f, s, b, h)$ of these variables and obvious notation:

$$p(f, s, b, h) = p_F(f)p_S(s|f)p_B(b|f)p_H(h|b) \quad (15)$$

The BN G_1 contains two statements: once given his blood pressure, risk of future heart attack is not affected by whether the man took the treatment; and given he takes the treatment, his sperm count is uninformative about his future blood pressure and risk of heart attack.

Note, it can be checked, using the semigraphoid axioms, that G_1 contains exactly the same irrelevance statements as, for example, G_2 (see Figure 8).

Now assume that the expert judgment encoded in G_1 is correct. Then it could be argued that by writing G_1 , the expert might be prepared to not only assert the conditional independence in the observed (or *idle*) system as above *but also* the probability distribution of the process, if it is manipulated – i.e., about the causal connections between the variables above. Unlike the BN, such a graph would distinguish between G_1 and G_2 , assigning a meaning to the directionality of edges in the graph. How various

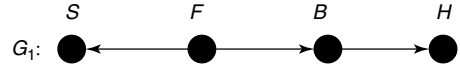


Figure 7 A DAG for Example 3

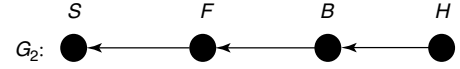


Figure 8 An equivalent DAG

Consider the situation described in G_1 , but modified to the prediction of what would happen if the fertility treatment was imposed $\{f = 1\}$ on all such men. Call this manipulation $\{do F = 1\}$. Then the effects of this policy might be that in this manipulated system, $\{f = 1\}$ would happen with probability 1 and the density $p(f, s, b, h|do f = 1)$ would be given by

$$p(s, b, h|do f = 1) = p_S(s|f = 1)p_B(b|f = 1)p_H(h|b) \quad (16)$$

While, if we banned the sales of the treatment, $\{f = 0\}$ would happen with probability 1 and

$$p(s, b, h|do f = 0) = p_S(s|f = 0)p_B(b|f = 0)p_H(h|b) \quad (17)$$

Note that in this case, $p(s, b, h|do f = 1) = p(s, b, h|f = 1)$ and $p(s, b, h|do f = 0) = p(s, b, h|f = 0)$. However, if the man receives another treatment in the future, which forces his blood pressure to be high $\{do b = 1\}$, then this cannot affect whether or not he previously took the fertility drug or whether he previously had an increased sperm count. Furthermore, we might expect his probability of a heart attack to be as high as in the idle system. Thus, given G_1 accurately represents the causal chains in the system, we could expect to infer that

$$p(s, f, h|do b = 1) = p_F(f)p_S(s|f)p_H(h|b = 1) \quad (18)$$

Note that this time this is not the same formula as for $p(s, f, h|b = 1)$, which is given by

$$p(s, f, h|b = 1) = \frac{p_F(f)p_B(b = 1|f) \times p_S(s|f)p_H(h|b = 1)}{p_F(f = 0)p_B(b = 1|f = 0) + p_F(f = 1)p_B(b = 1|f = 1)} \quad (19)$$

authors [18–20] suggest this might be done is given below.

This difference should be expected. If we subsequently learn that a man has high blood pressure, we

would infer that it is more likely he took the treatment that tends to cause this effect. On the other hand, if the man's blood pressure was made high by a subsequent treatment this tells us nothing about what it might have been before and so nothing about whether he had taken the original treatment.

In general, [19] and [18] suggest that we can legitimately call a BN G "causal". Suppose that for *each* variable X in G and each of its possible values x the manipulation $\{do\ X = x\}$ has the effect described above. Thus, after all such manipulations, the DM believes that (a) the new mass function conditional on the manipulation $\{do\ X = x\}$ of all variables Y occurring after X in G , will be equal to the corresponding mass function conditional on *observing* $\{X = x\}$ (in the factorization of the unmanipulated system associated with G), (b) the new mass function, conditional on the manipulation of all other variables Y (except X), will be as it was in the unmanipulated system, had $\{X = x\}$ **not** been observed, (c) the new mass function conditional on the manipulation sets X to x with probability 1. If this truly expresses what the DM believes about the system, then G is called a *causal Bayes Net*. Note that causal BNs assert much more than the BN alone. In particular, unlike the Bayes Net, if the DAGs of two causal BNs are different, then they assert different things. For example, the DAGs G_1 and G_2 above are equivalent representations of the same BN, but as causal BNs they assert very different things about the expert's predicted effects of controlling the system.

We can, in fact, represent a causal BN as an ID. Using the no forgetting convention [1], this can be drawn as in Figure 9.

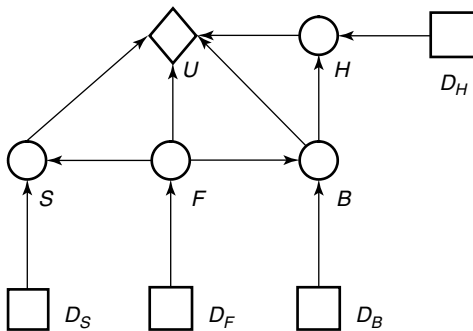


Figure 9 Causal BN as influence diagram

Here, for example, the decision variable D_B chooses between three options: to leave blood pressure to develop naturally, to medicate to lower the pressure, or to force it to be high. The probability tables associated with these different decisions simply follow the convention outlined above. Causal BNs give a far from general framework for expressing *all* types of causal structures. Their semantics are severely restricted and in particular, imply that the class of possible causal manipulations can only convert the BN into a uniform decision problem. Furthermore, in any given context, it may not be at all clear *which* parameterization of variables to use and exactly *what* doing a variable would actually mean. Thus, in the example above, what we mean by forcing blood pressure to be high is ambiguous, unless further elaborated (see [21, 22] for further discussion of this point).

However, in some problems, especially those associated with outputs of networks of simulators where variables are functionally defined and interventions unambiguous, they have provided a very useful and evocative tool to unpick and explain complicated control problems.

Links between Influence Diagrams and Decision Trees

Even when decision trees are used to represent a problem, it is now common to represent the dependence structure between some of its variables using a BN. Thus, for example, the historic BN of the decision problem described in the article on *decision trees* is simply represented as in Figure 10.

However, what is not sometimes realized, is that many problems expressible as decision trees are not uniform and so cannot be directly expressed as IDs. The decision problem used as a running example for the decision tree section above is such a problem. It is sometimes possible to construct dummy variables, so that within this new framework, the decision problem *is* uniform. However, such constructions are often contrived, involve the addition of many extra

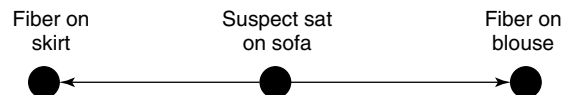


Figure 10 Another historic BN

edges, and more variables as functions of others in the ID, and so are less than transparent and can lose the conditional independence structure that was there originally. This might not matter so much for computational algorithms, but for model description and exploration, it does.

Various hybrid forms of trees, BNs, and IDs now exist (for example, see [23–25]), as well as other competing graphical representations of problems (see e.g. [26–28]). These methods are typically very powerful, but are either semantically more complicated or suited to the representation of more specific kinds of decision problems than the ID and decision tree.

References

- [1] Shachter, R.D. (1986). Evaluating influence diagrams, *Operations Research* **34**, 871–882.
- [2] Smith, J.Q. (1989). Influence diagrams for Bayesian decision analysis, *European Journal of Operational Research* **40**, 363–376.
- [3] Canning, C., Thompson, E.A. & Skolnick, M.H. (1978). Probability functions on complex pedigrees, *Advances in Applied Probability* **10**, 26–61.
- [4] Lauritzen, S.L. & Spiegelhalter, D.J. (1988). Local computation with probabilities on graphical structures and their application to expert systems (with discussion), *Journal Royal Statistical Society B* **50**, 157–224.
- [5] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo.
- [6] Howard, R.A. & Matheson, J.E. (eds) (1984). *Readings on the Principles and Applications of Decision Analysis*, Strategic Decision Group, Menlo Park, pp. 719–762, Vol. 2.
- [7] Shachter, R.D. (1988). Probabilistic inference and influence diagrams, *Operations Research* **36**(4), 589–604.
- [8] Jensen, F.V. & Nielsen, T.D. (2007). *Bayesian Networks and Decision Graphs*, 2nd Edition, Springer-Verlag, New York.
- [9] Cowell, R.G., Dawid, A.P., Lauritzen, S.L. & Spiegelhalter, D.J. (1999). *Probabilistic Networks and Expert Systems*, Springer-Verlag.
- [10] Marshall, K.T. & Oliver, R.M. (1995). *Decision Making and Forecasting*, McGraw-Hill.
- [11] Lauritzen, S.L. (1996). *Graphical Models*, 1st Edition, Oxford Science Press, Oxford.
- [12] Studeny, M. (2005). *Probabilistic Conditional Independence Structures*, Springer-Verlag.
- [13] Koster, J.T.A. (1996). Markov properties of nonrecursive causal models, *Annals of Statistics* **24**, 2148–2177.
- [14] Verma, T. & Pearl, J. (1988). Causal networks: semantic and expressiveness, *Proceedings of 4th Uncertainty in Artificial Intelligence 4*, Minneapolis, pp. 352–359.
- [15] Phillips, L.D. (1984). A theory of requisite decision models, *Acta Psychologica* **56**, 29–48.
- [16] Dawid, A.P. (2002). Influence diagrams for causal modelling and inference, *International Statistical Reviews* **70**, 161–189.
- [17] Vomelelova, M. & Jensen, F.V. (2002). An extension of lazy evaluation for influence diagrams avoiding redundant variables in the potentials, in *Proceedings of the First European Workshop on Probabilistic Graphical Models*, J.A. Gamez & A. Salmeron, eds, Cuenca.
- [18] Pearl, J. (2000). *Causality. Models, Reasoning and Inference*, Cambridge University Press, Cambridge.
- [19] Spirtes, P., Glymour, C. & Scheines, R. (1993). *Causation, Prediction, and Search*, Springer-Verlag, New York.
- [20] Pearl, J. (2003). Statistics and causal inference: a review (with discussion) *Sociedad de Estadística e Investigación, Operativa Test* **12**(2), 281–345.
- [21] Shafer, G.R. & Pearl, J. (eds) (1990). *Readings in Uncertainty Reasoning*, Morgan Kaufmann.
- [22] Smith, J.Q. & Anderson, P.E. (2008). Conditional independence and chain event graphs, *Artificial Intelligence* **172**, 42–68.
- [23] Salmaron, A., Cano, A. & Moral, S. (2000). Importance sampling in Bayesian Networks using probability trees, *Computational Statistics and Data Analysis* **24**, 387–413.
- [24] McAllister, D., Collins, M. & Periera, F. (2004). Case factor diagrams for structured probability modelling, *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence (UAI-04)*, Banff, pp. 382–391.
- [25] Poole, D. & Zhang, N.L. (2003). Exploiting contextual independence in probabilistic inference, *Journal of Artificial Intelligence Research* **18**, 263–313.
- [26] Covaliu, Z. & Oliver, R.M. (1995). Representation and solution of decision problems using sequential decision diagrams, *Management Science* **41**, 1860–1881.
- [27] Jensen, F.V., Nielsen, T.D. & Shenoy, P.P. (2004). Sequential Influence Diagrams: a unified asymmetry framework, in *Proceedings of the second European Workshop on Probabilistic Graphical Models*, P. Lucas, ed, Lorentz Center, Leiden, pp. 121–128.
- [28] Smith, J.Q. & Figueroa-Quiroz, L.J. (2007). A causal algebra for dynamic flow networks, *Advances in Probabilistic Graphical Models*, P. Lucas, J.A. Gamez & A. Salmeron, eds, Springer, pp. 39–54.

Further Reading

- De Finetti, B. (1975). *Theory of Probability*, John Wiley & Sons, Vol. 1.
- Geiger, D., Verma, T.S. & Pearl, J. (1990). Identifying independence in Bayesian networks, *Networks* **20**, 507–534.
- Jaegar, M. (2004). Probability decision graphs – combining verification and AI techniques for probabilistic inference, *International Journal of Uncertainty Fuzziness and Knowledge Based Systems* **12**, 19–42.
- Jensen, F., Jensen, F.V. & Dittmar, S.L. (1994). From influence diagrams to junction trees, *Proceedings of the 10th*

- Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco pp. 367–373.
- Oliver, R.M. & Smith, J.Q. (1990). *Influence Diagrams, Belief Nets, and Decision Analysis*, John Wiley & Sons.
- Puch, R.O. & Smith, J.Q. (2004). FINDS: a training package to assess forensic fibre evidence, in *Advances in Artificial Intelligence*, Springer-Verlag, 420–429.
- Savage, L.J. (1972). *The Foundations of Statistics*, 2nd Edition, Dover.
- Smith, J.Q. (1988). *Decision Analysis: A Bayesian Approach*, Chapman & Hall.
- Smith, J.Q. (1995). Handling multiple sources of information using influence diagrams, *European Journal of Operational Research* **86**, 189–200.
- Thwaites, P.A. & Smith, J.Q. (2006). Evaluating Causal Effects using Chain Event Graphs, *Proceedings of the*

Third Workshop on Probabilistic Graphical Models, Prague, 291–300.

Related Articles

Bayes' Theorem and Updating of Belief

Bayesian Statistics in Quantitative Risk Assessment

Utility Function

JIM Q. SMITH AND PETER THWAITES

Informational Value of Corporate Issuer Credit Ratings

The quality of corporate issuer credit ratings^a provided by Standard & Poor's (S&P) (*see Credit Scoring via Altman Z-Score; Enterprise Risk Management (ERM)*) and Moody's (*see Risk in Credit Granting and Lending Decisions: Credit Scoring; Default Risk*) has come under closer scrutiny. Whether justified or not, the most talked about Enron and WorldCom cases intensified the debate on rating quality. The majority of investors believe that ratings are accurate measures of the issuer's creditworthiness. However, investors are less satisfied about their timeliness. Several surveys, conducted in the United States, reveal that most investors believe that rating agencies are too slow in adjusting their ratings to changes in the creditworthiness.^b At the same time, investors prefer some degree of rating stability to avoid frequent rebalancing of their portfolios, even when the underlying credit risk fundamentals justify rating migrations. Apparently, investors want both stable and timely ratings, which are two conflicting objectives. It is a dilemma for investors. Moody's tries to find a compromise: *"Moody's analysts attempt to balance the market's need for timely updates on issuer risk profiles, with its conflicting expectation for stable ratings"* [1].

Some critics argue that the unsatisfactory timeliness in ratings is mainly the result of a lack of competition in the National Recognized Statistical Rating Organization (NRSRO) rating industry. Because of legal and contractual restrictions, most institutional investors are only allowed to invest in companies rated by an agency with an NRSRO status. In order to have access to the capital market, companies are forced to have ratings from NRSRO agencies. By the end of 2006 only five agencies, namely, S&P, Moody's Investor Services, Fitch, A.M. Best, and the Canada's Dominion, had obtained this status from the SEC. Even though the legal basis of the NRSRO status is only found in the United States, an NRSRO rating is a necessary condition for large international companies to have sufficient access to the international capital market. In practice the two largest agencies S&P and Moody's are close to a duopoly,

since large companies prefer to have at least two ratings and market participants expect large companies to have more than one rating opinion. In practice this works fine as the ratings assignments of major rating agencies largely agree with each other.

To allow more competition in the rating industry, the US Congress approved the Credit Rating Agency Reform Act in September 2006. This act introduces a new system of voluntary registration, so any rating agency that can show a good track record – subject to objective criteria – can obtain an NRSRO designation. To what extent this act will change the rating industry depends on the use of ratings. Do ratings provide credit risk information in the first place or do ratings serve the certification need – to check whether portfolio's meet the portfolio eligibility standards set by regulators, fund trustees, or boards of directors? The first objective requires ratings to be superior by being timely and accurate while the second objective demands for some level of rating stability. A benchmark study with credit scores which proxies for the actual point-in-time credit quality shows that ratings exhibit a high rating stability: the number of rating migrations is lowered by a factor 2.5. However, the stability objective – achieved by the agencies' through-the-cycle methodology is at a high cost: ratings are delayed by 8–9 months and the default prediction performance of ratings is seriously affected: for a 1-year prediction horizon the accuracy ratios drop by 9.9% relative to a long-term point-in-time perspective. A shift to a more balanced equilibrium between rating stability and rating timeliness might result when the rating industry becomes more competitive – as intended by the new act.

The sections titled "Motivations to Increase Rating Stability" and "Rating Stability Enhancement by a Through-the-Cycle Methodology" describe the motivations to increase rating stability and how agencies achieve rating stability by a through-the-cycle methodology. The section titled "Costs of Rating Stability in Terms of Rating Timeliness and Default Prediction Performance" reports the costs of rating stability in terms of rating timeliness and default prediction performance. The section titled "The added value of rating outlooks and rating reviews" completes this quantitative picture by examining the added value of rating outlooks and rating reviews. The section titled "A New Balance Between Rating Timeliness and Stability due to More Competition?" concludes with an outlook for the rating industry.

Motivations to Increase Rating Stability

A number of arguments can be put forward to strive for some level of rating stability.

1. Investors do not want ratings to be changed to reflect small changes in the companies' financial condition, which are likely to be reversed in the near future. This will force investors to rebalance their portfolio's unnecessarily to meet the portfolio eligibility standards. By following this line of reasoning, S&P [2] is convinced that stable ratings are of most value to investors. Companies – paying for the agency rating services – also prefer rating stability, which prevents continuous renegotiation with investors and banks.
2. From a supervisory perspective, rating stability is desirable to prevent procyclicality effects. A prompt and full response to changes in actual creditworthiness could deepen a financial crisis. Linkages of portfolio strategies and portfolio mandates with ratings and, in future, linkages of bank capital requirements with ratings, can force banks and investors to liquidate their positions hurriedly, which might ultimately result in a credit crunch. Rating stability could soften these procyclicality effects.
3. Rating stability supports the reputation of agencies. Rating reversals within a short period have a negative impact on an agency's reputation, even when they reflect true changes in creditworthiness. It is better to be late and right, rather than fast and wrong.
A strong reputation, which underlies the recognition of ratings in financial markets, is in the interest of not only the agencies themselves but also that of authorities, investors, and bond issuing companies. It ensures a broad acceptance of ratings within the financial community, which is essential to lower information costs of investors and fuel the investment flow in corporate bonds.
4. From an organizational perspective, a rating stability objective keeps the costs of internal reviewing and monitoring under control because rating assignments are based on costly and labor-intensive in-depth analyses of companies. An ambitious timeliness objective would require a complete redesign of the rating assignment process. In this case monitoring and reviewing will

have to rely for an important part on automatic processing of statistical data to keep the cost level of rating services at a reasonable level. Nevertheless, because of reputation effects, agencies are likely to use automatic analysis for monitoring purposes.

Rating Stability Enhancement by a Through-the-Cycle Methodology

Agencies aim to balance rating stability and rating timeliness by a through-the-cycle methodology. This methodology has two aspects: first, a focus on the permanent component of default risk, and second, a prudent migration policy.

The first aspect of the through-the-cycle rating methodology refers to its name. By filtering out the temporary component of default risk, only the permanent, long-term, and structural component is measured. The second – less well known – aspect of the through-the-cycle methodology is the enhancement of rating stability by a prudent migration policy. Only substantial changes in the permanent component of default risk lead to rating migrations and, if triggered, ratings are only partially adjusted to the actual level in the permanent component of default risk. Although not officially disclosed by agencies, Moody's provided some insight in January 2002, stating that its migration policy was to be reconsidered: *“Under consideration are more aggressive ratings changes – such as downgrading a rating by several notches immediately in reaction to adverse news rather than slowly reducing the rating over a period of time – as well as shortening the rating review cycle to a period of weeks from the current period of months”*.^c

Precisely how rating agencies put their through-the-cycle methodology into practice is not clear. Treacy and Carey [3] describe the through-the-cycle rating methodology as a rating assessment in a worst-case scenario, at the bottom of a presumed credit quality cycle. Altman and Rijken [4] characterize the migration policy as follows. A rating migration is triggered, when the actual permanent credit risk component exceeds a threshold of 1.25 notch steps, relative to the average “through-the-cycle credit quality” in a particular rating class.^d If triggered, the rating migration only closes 75% of the gap between the current rating level and the target rating level. Although these parameters do not seem to point to an

extreme prudent migration policy, they are sufficient to cause a significant enhancement in rating stability.

In contrast to the through-the-cycle methodology, investors judge the corporate credit quality from a point-in-time perspective, on the basis of the actual position in the corporate credit cycle. Bankers have a point-in-time perspective on corporate credit quality with a time horizon of between 1 and 7 years (*see* Basel Committee [5]; **Potency Estimation**). It is reasonable to assume that this perspective applies to other investors as well. As opposed to through-the-cycle measurement approach, the point-in-time perspective looks at the actual default risk without an attempt to suppress the temporary component in default risk.

Costs of Rating Stability in Terms of Rating Timeliness and Default Prediction Performance

A drawback of the through-the-cycle perspective is its mismatch with the investor's point-in-time perspective. The cost of rating stability enhancement is a deterioration of rating timeliness and default prediction performance. Altman and Rijken [6] have carried out a benchmark study for S&P issuer credit ratings with credit scores to quantify these costs and to separate out the contributions of the two aspects of the through-the-cycle methodology – the long-term perspective and the prudent migration policy – to the rating stability level. Reference point to measure rating stability is the investor's long-term (6-year) point-in-time perspective.

Table 1 lists the quantitative consequences of the agencies' through-the-cycle methodology from a 6-year point-in-time perspective for Moody's issuer credit ratings.^c From a point-in-time perspective on default risk, weighting both permanent and temporary component of default risk, the through-the-cycle methodology lowers the annual rating migration probability by a factor of 2.6 and delays the timing of rating migrations by 8 months on the downgrade side and by 9 months on the upgrade side. As a result, the through-the-cycle rating methodology affects the accuracy of default prediction: from a long-term point-in-time perspective the accuracy ratio drops by 9.9% for 1-year default prediction horizon. As expected, for a 6-year prediction horizon the negative impact of the through-the-cycle methodology is less severe: -3.3%.

The through-the-cycle methodology seriously affects the short-term default prediction performance of ratings. It fully offsets the information advantage agencies naturally have. Rating agencies have resources, capabilities, and access to private information to produce better estimates on credit quality than (outside) investors. In terms of accuracy ratios, their information advantage is about 3% compared to simple credit scores including market information (for example, EDF scores of Moody's KMV; *see* **Credit Scoring via Altman Z-Score**). This information advantage does not seem impressive, but it is extremely difficult to add new information on top of credit score information, including stock price information, which reflects a market opinion consensus.

The Added Value of Rating Outlooks and Rating Reviews

For a complete quantification of the information signaled by agencies, rating outlooks and rating reviews must be taken into account. In addition to their corporate issuer credit ratings, agencies provide rating outlooks and rating reviews. Rating outlooks signal the likely direction of a rating migration in one to two years' time. In response to an event or an abrupt break in a trend, a corporate issuer credit rating is placed on a Watchlist by Moody's, on CreditWatch by S&P, or on a rating Watch by Fitch. In these Rating Review cases, ratings are said to be under review and the outcome is disclosed typically within 90 days.

Although outlooks – rating outlooks and rating reviews – are not meant to be a correction for ratings in the first place, one can use them as a secondary credit risk measure on top of the rating scale. One can add the information of rating outlooks and rating reviews to ratings by adding or subtracting two or three notch steps depending on the type and sign of outlooks. A study by Cantor and Hamilton [7] reveals that the adjustment of ratings significantly improves default prediction performance. The accuracy ratio increases by 4% for a 1-year prediction horizon.

Outlook information is not sufficient to compensate for the negative impact of the through-the-cycle methodology. Potentially, adjusted ratings do have a long-term point-in-time perspective when agencies standardize credit risk information in the outlook scale. As outlooks are not intended to quantify credit risk information explicitly, agencies have not

Table 1 Effects of the through-the-cycle methodology on rating properties from a long-term point-in-time perspective^(a)

	Rating stability Reduction factor in number of rating migrations	Rating timeliness Delay in rating migrations (months)		Default prediction performance Decrease in accuracy ratio (%)	
		For downgrades	For upgrades	1-year prediction horizon	6-year prediction horizon
Effects due to the two aspects of the through-the-cycle methodology from a long-term (6-year) point-in-time perspective					
1. Filtering the temporary credit risk component	×0.75	4.4	3.8	6.9	2.1
2. Prudent migration policy	×0.5	3.4	5.1	3.0	1.2
Total through-the-cycle (TTC) effect	×0.4	7.8	8.9	9.9	3.3

^(a)Results are obtained for Moody's ratings in the period 1982–2005

standardized credit risk information in the outlook assignment process. Credit risk standardization in the outlook scale could improve the default prediction performance of adjusted ratings even further with another 4–6% and could enable to bridge the gap between the agencies' through-the-cycle perspective and the investor's point-in-time perspective. In that case agencies can make full use of their information advantage and also become competitive on rating timeliness and default prediction performance with other credit risk information providers.

A New Balance Between Rating Timeliness and Stability due to More Competition?

In relative transparent markets of US listed companies, corporate issuer ratings have little informational value for investors even when Rating Outlook and Rating Review information is added. The rating stability objective fully offsets the agencies' information advantage. Apparently, agencies have strong incentives to give the rating stability objective priority over a superior timeliness or informational objective. So agencies fulfill most of all the certification needs of investors while investors obtain their up-to-date credit risk information from a wide range of other providers focusing on point-in-time credit risk information.

As the business of credit risk information providers is growing since 2001 and become more sophisticated, agencies can't compromise rating quality too much with their prudent rating through-the-cycle methodology. Credibility is a crucial

condition. Since 2001 the reputation of the major agencies is at stake. Agencies might have pushed the rating stability objective too far. A prudent approach to protect their reputation works well in a relative stable financial environment. However, in volatile capital markets and a more competitive rating industry, agencies must provide more timely information to protect their credibility as rating institutions. So, after the Enron and WorldCom debacles, in 2002 Moody's intended to change ratings more aggressively and update them more frequently, in response to a demand for more rating timeliness. Moody's renounced this intention, after broad consultation with investors, companies, and financial authorities. In their meetings Moody's repeatedly heard that investors value the current level of rating stability and do not want ratings simply to follow market prices. Moody's therefore decided not to change their rating policy, and they continue to produce stable ratings (see Fons *et al.* [8]).

More competition after approval of the Credit Rating Agency Reform Act can put more pressure on the agencies to search for an optimal balance between rating stability and rating timeliness. Although companies are the formal clients of rating agencies, the investor opinion is more important in setting the balance between rating stability and rating timeliness, which will show up in a more competitive rating industry. The NRSRO designation requirement set by regulators, fund trustees, or boards of directors in the portfolio eligibility standards are probably going to be replaced by a list of agencies based on reputation and worldwide coverage. In that case the dominant

market share of Moody's, S&P, and Fitch will not be challenged in the short term. However, in contrast to the past situation, their position in the long run will not be protected by an NRSRO designation alone but also by a track record in rating quality performance. Consequently agencies will have to push their balance between rating stability and rating timeliness more to the edge and serve both the certification needs and information needs of investors.

Presumably only a restricted number of major rating agencies will survive in free market circumstances. Portfolio eligibility standards have to be widely accepted, so investors prefer a restricted number of rating agencies, which limits conflicting rating opinions and safeguards the consistency in the rating scale. Most investors understand the current standards AAA (Aaa)–CCC (Caa) of the major rating agencies. A large number of rating agencies would probably result in rating scale diversity. This could result in a decline of the rating standards and damage the certification function of ratings. A new task of the SEC should be to preserve some general accepted standards in the measurement of credit risk.

The implementation of Basel II, the strong increase in structured products, the strong growth of financial markets in emerging economies, and the disintermediation in financial markets have greatly stimulated the demand for ratings in less transparent markets. This huge increase in demand can probably not be absorbed by a few major agencies, so a large number of smaller agencies will emerge focusing on niche markets. For the success of these new agencies, the key will be whether they can develop sufficient credibility to serve the certification needs of investors. This can be done by building a good track record in rating stability as well as rating timeliness and default prediction performance.

End Notes

^a. In the remainder of this article, ratings refer to the corporate issuer credit ratings of S&P, Moody's, and Fitch, unless stated otherwise.

^b. See, for example, the surveys of the Association for Financial Professionals (AFP) in 2002 and 2004. A similar case has been made earlier by Ellis [9] and Baker and Mansi [10]. The AFP survey reveals that 83% of the investors believe that, most of the time, ratings accurately reflect the issuer's

creditworthiness. On the other hand, all surveys show that investors are not satisfied with the timeliness of ratings. This complaint does not only arise from the particular Enron and WorldCom cases. Halfway the nineties the survey by Ellis showed that 70% of investors believed that ratings should reflect recent changes in default risk, even if they are likely to be reversed within a year.

^c. See *The Financial Times*, 19 January 2002, "Moody's mulls changes to its ratings process".

^d. This threshold level is derived for a monitoring frequency of one year.

^e. Results of the benchmark study for Moody's ratings published in this paper largely agree with the results for S&P ratings published in Altman and Rijken [6].

References

- [1] Cantor, R. (2001). Moody's investor services response to the consultative paper issued by the Basel Committee on Bank Supervision "A new capital adequacy framework", *Journal of Banking and Finance* **25**(1), 171–185.
- [2] Standard & Poor's (2003). *Corporate Ratings Criteria*, <http://www.standardandpoors.com>.
- [3] Treacy, W.F. & Carey, M. (2000). Credit rating systems at large US banks, *Journal of Banking and Finance* **24**(1–2), 167–201.
- [4] Altman, E.I. & Rijken, H.A. (2004). How rating agencies achieve rating stability, *Journal of Banking and Finance* **28**(11), 2679–2714.
- [5] Basel Committee on Banking Supervision (2000). *Range of Practice in Banks' Internal Rating Systems*, in publication series of Bank for International Settlements.
- [6] Altman, E.I. & Rijken, H.A. (2006). The effects of rating trough-the-cycle on rating stability, rating timeliness and default-prediction performance, *Financial Analysts Journal* **62**(1), 54–70.
- [7] Cantor, R. & Hamilton, D. (2004). Rating transition and default rates conditioned on outlooks, *Journal of Fixed Income* **14**, 54–70.
- [8] Fons, J.S., Cantor, R. & Mahoney, C. (2002). *Understanding Moody's Corporate Bond Ratings and Rating Process*, Special Comment, Moody's Investor Services.
- [9] Ellis, D. (1998). Different sides of the same story: investors' and issuers' views of rating agencies, *The Journal of Fixed Income* **7**(4), 35–45.
- [10] Baker, H.K. & Mansi, S.A. (2002). Assessing credit agencies by bond issuers and institutional investors, *Journal of Business Finance and Accounting* **29**(9–10), 1367–1398.

HERBERT A. RIJKEN AND EDWARD I. ALTMAN

Insurability Conditions

What does it mean to say that a particular risk is insurable? This question must be addressed from the vantage point of the potential supplier of insurance who offers coverage against a specific risk at a stated premium. The policyholder is protected against a prespecified set of losses defined in the contract^a.

Two conditions must be met before insurance providers are willing to offer coverage against an uncertain event. Condition 1 is the ability to identify and quantify, or estimate, the chances of the event occurring, and the extent of losses likely to be incurred when providing different levels of coverage. Condition 2 is the ability to set premiums for each potential customer or class of customers. This requires some knowledge of the customer's risk in relation to others in the population of potential policyholders. If Conditions 1 and 2 are both satisfied, a risk is considered to be insurable. But it still may not be profitable. In other words, it may be impossible to specify a rate for which there is sufficient demand and incoming revenue to cover the development, marketing, and claims costs of the insurance and yield a net positive profit. In such cases, the insurer will opt *not* to offer coverage against this risk.

Condition 1: Identifying the Risk

To satisfy this condition, estimates must be made of the frequency at which specific events occur and the magnitude of the resulting loss. Such estimates can use historical data from previous events, such as peril of fire, or scientific analyses of what is likely to occur in the future, such as earthquake peril. New developments in catastrophe risk modeling have made it easier to estimate the risk facing insurers when they are considering how much coverage to provide in hazard-prone areas.

Collecting Data on Risks

Rating agencies typically collect data on all the losses incurred over a period of time for particular risks and exposure units. Suppose the hazard is fire and the exposure unit is a well-defined entity, such as a \$300 000 wood-frame home in California, to be insured for 1 year. The typical measurement is the

pure premium (PP), which is the basis for setting an actual premium for potential customers. The PP is defined as

$$PP = \frac{\text{total losses}}{\text{total number of exposure units}} \quad (1)$$

The PP normally considers loss adjustment expenses for settling a claim. We assume that this is part of total losses. For more details on calculating PP, see Launie *et al.* [2].

Assume that the rating agency has collected data on 100 000 wood-frame homes in that state and has determined that the total annual loss from fires to these structures over the past year is \$20 million. If these data are representative of the expected loss to this class of wood-frame homes in California next year, then, using equation (1), the PP is given by

$$PP = \frac{\$20\,000\,000}{100\,000} = \$200$$

This figure is simply an average. It does not differentiate between wood-frame homes of different values, their locations in the state, the distance of each home from a fire hydrant, or the quality of the fire department serving different communities. All of these factors are often taken into consideration by underwriters who specify a premium that reflects their estimate of the risk to particular structures.

Insurance providers use scientific studies by seismologists, geologists, and structural engineers to estimate the frequency of earthquakes of different magnitudes, as well as the damage that is likely to occur to different structures from such earthquakes.

Condition 2: Setting Premiums for Specific Risks

Once the risk has been identified, the insurer needs to determine what premium it can charge to make a profit while not subjecting itself to an unacceptably high chance of a catastrophic loss. There are a number of factors that play a role in determining what prices companies would like to charge. Owing to problems of ambiguity aversion (for uncertain probabilities), adverse selection, moral hazard, and highly correlated losses, insurers may want to charge premiums that considerably exceed the expected loss.

Insurability Conditions and Demand for Coverage

For some risks the desired premium may be so high that there would be very little demand for coverage at that rate. In such cases, even though an insurer determines that a particular risk meets the two insurability conditions discussed above, it does not invest the time and money to develop the product.

More specifically, the insurer must be convinced that there is sufficient demand to cover the development and marketing costs of the coverage through future premiums received. If state regulations restrict the price insurers can charge for certain types of coverage, then companies may not provide protection against these risks. In addition, if an insurer's portfolio leaves it vulnerable to the possibility of extremely large losses from a given disaster because of adverse selection, moral hazard, and/or high correlation of

risks, then again the insurer may be unwilling to provide coverage against such disasters.

End Notes

^a. This section is based on Freeman and Kunreuther [1, Chapter 4].

References

- [1] Freeman, P.K. & Kunreuther, H. (1997). *Managing Environmental Risks through Insurance*, American Enterprise Institute, Washington, DC; Kluwer: Norwell, MA.
- [2] Launie, J.J., Finley Lee, J. & Baglini, N.A. (1986). *Principles of Property and Liability Underwriting*, 3rd Edition, Insurance Institute of America, Malvern, PA.

HOWARD KUNREUTHER

Insurance Applications of Life Tables

Causes of risk in life insurance (*see Risk Classification/Life*) relate to financial aspects (e.g., investment yield, inflation, etc.), demographical aspects (e.g., lifetimes of policyholders, frequency of disability, lapses and surrenders, etc.), and expenses. In this article, we deal with demographical aspects only, focusing on insureds' lifetimes, which in turn determine the mortality, namely the frequency of death in a life insurance portfolio (or a pension plan).

Mortality constitutes an important cause of risk in a life insurance portfolio. A number of risk factors affect individual mortality. At an individual level, the insured's age has a prominent role. Other risk factors are gender, health status, profession, smoking habits, etc. So, formulae used to calculate premiums and reserves for life insurance and annuity products should allow for various risk factors. In particular, the insured's age enters formulae *via* the age pattern of mortality that is a structure linking survival and death probabilities to the attained age (*see Logistic Regression*).

In this article, we first describe the features of a (period) life table that constitutes the basic tool to represent the age pattern of mortality, and the applications of life tables to the calculation of expected present values (or "actuarial" values) in life insurance. Age-discrete and age-continuous approaches to the representation of the mortality pattern are described. We then introduce the ideas underlying the projected life tables, meant as a tool to represent future mortality. Finally, we focus on randomness in the results of a life insurance portfolio (cash flows, profits, etc.), and the use of stochastic models.

Age as a Risk Factor in Life Insurance

Let x denote the insured's age at policy issue. Consider for example, a 1-year insurance cover providing the beneficiaries with the unitary amount at the end of the year if the insured dies within the year. The expected present value or "actuarial" value, ${}_1A_x$, of the benefit is given by

$${}_1A_x = {}_1q_x v \quad (1)$$

where ${}_1q_x$ (or simply q_x) denotes the probability of a person age x dying within 1 year, and v denotes the discount factor, $v = \frac{1}{1+i}$, where i is the annual rate of interest.

Consider an insurance product (the "pure endowment") providing the policyholder with the unitary amount if she/he is alive at time n (the maturity). The actuarial value of the benefit is as follows:

$${}_nE_x = {}_np_x v^n \quad (2)$$

where ${}_np_x$ denotes the probability of the insured being alive at maturity.

An immediate life annuity provides a sequence of annual amounts (e.g., paid at the end of each year) while the beneficiary is alive. Referring to a unitary annual amount, the actuarial value is given by

$$a_x = {}_1p_x v + {}_2p_x v^2 + \dots \quad (3)$$

Actuarial values, like ${}_1A_x$, ${}_nE_x$, and a_x , depend on age x *via* probabilities of death and survival, like ${}_1q_x$ and ${}_hp_x$, expressing mortality as a function of the current age x (as well as the age at the time of payment).

Other risk factors are usually allowed for in actuarial calculations. As regards the insured's gender, it should be noted that males and females have distinct mortality patterns. Specific patterns owing to the health status or profession are commonly expressed as functions of the "normal" mortality. Recognizing the role of various risk factors in determining the individual mortality pattern means that some heterogeneity is allowed for in an insured population.

Life Tables

The life table is a (finite) decreasing sequence $l_0, l_1, \dots, l_\omega$. The generic item l_x refers to the integer age x , and represents the estimated number of people alive at that age in a properly defined population. The exact meaning of the l_x 's is explained after we discuss two approaches to the calculation of these numbers.

Assume that the sequence $l_0, l_1, \dots, l_\omega$ is provided by statistical evidence, that is, by a "longitudinal" observation of the actual numbers of individuals alive at age 1, 2, $\dots$, ω , out of a given initial "cohort" consisting of l_0 individuals. The age ω is the maximum age (say, $\omega = 110$), i.e., the age such

that $l_\omega > 0$ and $l_{\omega+1} = 0$. The sequence $l_0, l_1, \dots, l_\omega$ is called a *cohort life table*. Clearly, the construction of a cohort table takes $\omega + 1$ years.

Assume, conversely, that the statistical evidence consists of the frequency of death at the various ages, observed throughout a given period, for example, 1 year. Assume the frequency of death at age x (possibly after a graduation with respect to x) as an estimate of the probability q_x .

For $x = 0, 1, \dots, \omega - 1$, define

$$l_{x+1} = l_x(1 - q_x) \quad (4)$$

with l_0 (the radix) assigned (e.g., $l_0 = 100\,000$). Hence, l_x is the expected number of survivors out of a notional cohort initially consisting of l_0 individuals. The sequence $l_0, l_1, \dots, l_\omega$, defined by recursion (4), is called a *period life table*, as it is derived from period observations.

Probabilities of death and survival can be derived from a life table. Referring to a person alive at age x , the probability of being alive at age $x + 1$, ${}_1p_x$ or simply p_x , is clearly given by

$$p_x = 1 - q_x = \frac{l_{x+1}}{l_x} \quad (5)$$

More generally, the probability of being alive at age $x + h$ (with h a positive integer number) is given by

$${}_hp_x = p_x p_{x+1} \dots p_{x+h-1} = \frac{l_{x+h}}{l_x} \quad (6)$$

(with ${}_0p_x = 1$).

The probability of dying before age $x + h$ is given by

$${}_hq_x = 1 - {}_hp_x = \frac{l_x - l_{x+h}}{l_x} \quad (7)$$

(whence ${}_0q_x = 0$). Finally, the probability of dying between ages $x + h$ and $x + h + k$ is as follows:

$${}_{h|k}q_x = {}_hp_x {}_kq_{x+h} = \frac{l_{x+h} - l_{x+h+k}}{l_x} \quad (8)$$

Note that, in all symbols, the right-hand side subscript denotes the age referred to. For example, in equation (8) the probability ${}_kq_{x+h}$ relates to an individual age $x + h$, i.e., assumed alive at that age. Conversely, the left-hand side subscript denotes some duration, whose meaning depends on the specific probability addressed (see equations (6–8)).

Probabilities as defined by equations (5–8), underpin the calculation of actuarial values, which can be used to determine premiums (see **Risk Classification in Nonlife Insurance**) and mathematical reserves (provided that the so-called equivalence principle has been adopted). For example, the single premium, Π , for a n -year term insurance providing a death benefit C at the end of the year of death is given by

$$\Pi = C_n A_x = C \sum_{h=1}^n {}_{h-1|1}q_x v^h \quad (9)$$

The single premium for an immediate life annuity paying the annual amount R is given by

$$\Pi = R a_x = R \sum_{h=1}^{\omega-x} {}_hp_x v^h \quad (10)$$

The mathematical reserve of an insurance contract quantifies the insurer's liability related to the policy itself at a given time t . For example, the mathematical reserve, V_t , of a single-premium term assurance is given by

$$V_t = C_{n-t} A_{x+t} \quad (11)$$

while for an annuity (see **Longevity Risk and Life Annuities**) we have

$$V_t = R a_{x+t} \quad (12)$$

However, in a modern, risk-oriented framework both premium and reserve calculations rely on more complex formulae, possibly allowing for the cost of options and guarantees embedded in the insurance product.

For more information about life tables and their use in actuarial calculations, the reader should refer to actuarial textbooks, for example [1, 2]; see also [3]. The construction of life tables is the specific topic of various textbooks, for example [4–6].

“Population” Tables versus “Market” Tables

Mortality data, and hence life tables, can originate from observations concerning a whole national population, a specific part of a population (e.g., retired workers, disabled people, etc.), an insurer's portfolio, etc.

Life tables constructed on the basis of observations involving a whole national population (split into females and males) are commonly referred to as *population tables*.

Market tables are constructed using mortality data coming from a collection of insurance portfolios and/or pension plans. Usually, distinct tables refer to assurances (i.e., insurance products with a positive sum at risk, for example, term assurances and endowments), annuities individually purchased, and pensions (i.e., annuities paid to the members of a pension plan).

The rationale of distinct market tables lies in the fact that mortality levels may significantly differ on moving from one type of insurance product to another. For example, the beneficiary of a purchased annuity usually is a person with a high life expectancy, whence an “adverse selection” affects the mortality pattern in annuity portfolios. Conversely, pensions are paid to people as a straight consequence of their membership of an occupational pension scheme.

Market tables provide experience-based data for premium and reserve calculations and for the assessment of expected profits. Population tables can provide a starting point when market tables are not available. Moreover, population tables usually reveal mortality levels higher than those expressed by market tables and hence are likely to constitute a prudential (or “conservative”, or “safe-side”) assessment of mortality in assurance portfolios. Thus, population tables can be used when pricing assurances, in order to include a profit margin (or a “safety loading”) into the premiums.

Examples of population tables and market tables can be found in [1] and [5].

Moving to a Time-Continuous Context

In a more formal setting, the probabilities previously defined can be expressed in terms of the remaining random lifetime, T_x , of an individual age x . So, we have

$${}_h p_x = \Pr[T_x > h] \quad (13)$$

$${}_h q_x = \Pr[T_x \leq h] \quad (14)$$

$${}_{h|k} q_x = \Pr[h < T_x \leq k] \quad (15)$$

However, equations (6–8) concern probabilities only involving integer ages x and integer times h

and k . Therefore, if we want to evaluate probabilities as in equations (13–15) when ages and times are real numbers, further tools are needed.

Assume that the function $S(t)$, called the *survival function* (see **Hazard and Hazard Ratio**) and defined for $t \geq 0$ as follows:

$$S(t) = \Pr[T_0 > t] \quad (16)$$

has been assigned. Clearly, T_0 denotes the random lifetime for a newborn.

Consider the probability defined by equation (13). Since

$$\begin{aligned} \Pr[T_x > h] &= \Pr[T_0 > x + h | T_0 > x] \\ &= \frac{\Pr[T_0 > x + h]}{\Pr[T_0 > x]} \end{aligned} \quad (17)$$

we find that

$${}_h p_x = \frac{S(x + h)}{S(x)} \quad (18)$$

For the probability defined by equation (14) we obtain

$${}_h q_x = \frac{S(x) - S(x + h)}{S(x)} \quad (19)$$

The same reasoning leads to

$${}_{h|k} q_x = \frac{S(x + h) - S(x + h + k)}{S(x)} \quad (20)$$

Turning back to the life table, note that since l_x is the expected number of people alive out of a cohort initially consisting of l_0 individuals, we have

$$l_x = l_0 \Pr[T_0 > x] \quad (21)$$

and, in terms of the survival function,

$$l_x = l_0 S(x) \quad (22)$$

(provided that all individuals in the cohort have the same mortality pattern, described by $S(x)$). Thus, the l_x 's are proportional to the values the survival function takes on integer ages x , whence the life table can be interpreted as a tabulation of the survival function.

How does one obtain the survival function for all real ages x ? Relation (22) suggests a practicable approach. First, set $S(x) = l_x/l_0$ for all integers

x using the available life table. Next, for $x = 0, 1, \dots, \omega - 1$ and $0 < f < 1$, define

$$S(x + f) = (1 - f)S(x) + fS(x + 1) \quad (23)$$

and assume $S(x) = 0$ for $x > \omega$, whence the survival function is a piece-wise linear function. Interpolation models other than the linear model used in equation (23) can be adopted. Moreover, the values of $S(x)$ can be fitted using some mathematical formula; however, the use of formulae for representing the mortality pattern can be better placed in the framework described below.

Consider the force of mortality (or instantaneous mortality intensity), μ_x , defined for all $x \geq 0$ as follows:

$$\mu_x = \lim_{t \rightarrow 0} \frac{tq_x}{t} \quad (24)$$

The function μ_x can be estimated, for example, for $x = 0, 1, \dots$, using period mortality observations. The estimated values can then be graduated, in particular, using a mathematical “law”. A number of laws have been proposed in actuarial and demographical literature, and are used in actuarial practice. A famous example is given by the Gompertz–Makeham law:

$$\mu_x = A + b c^x; \quad A \geq 0; b, c > 0 \quad (25)$$

An interesting relation links the survival function to the force of mortality. From definition (24), using equation (19), we obtain

$$\mu_x = -\frac{dS(x)/dx}{S(x)} \quad (26)$$

When μ_x has been assigned, relation (26) is a differential equation. Solving with respect to $S(x)$ (with the obvious condition $S(0) = 1$) leads to

$$S(x) = e^{-\int_0^x \mu_t dt} \quad (27)$$

Thus, the survival function can be obtained, and then all survival and death probabilities can be derived (see equations (18–20), with x, h , and k positive real numbers).

Time-continuous models and mortality laws are described in many actuarial textbooks; see, for example [1, 2]. See also [7] and the references therein.

Static versus Dynamic Mortality

An important hypothesis underlying recursion (4), which defines the l_x 's, should be clearly stated. As the q_x 's are assumed to be estimated from mortality experience in a given period (say, 1 year), the calculation of the l_x 's relies on the assumption that the mortality pattern does not change in the future. The same hypothesis underlies the calculation of $S(x)$ (see equation (27)), as μ_x is assumed to be estimated from period mortality data.

In many countries, however, statistical evidence show that human mortality declined over the twentieth century, and in particular, over its last decades. So, the hypothesis of “static” mortality cannot be assumed in principle, at least when long periods of time are referred to. Hence, in life insurance applications, the use of period life tables should be restricted to products involving short or medium durations (5–10 years, say), like term assurances and endowments, while it should be avoided when dealing with life annuities.

Allowing for trends in mortality implies the use of “projected” life tables, representing future mortality on the basis of the experienced mortality trend.

In a “dynamic” context, let $q_x(t)$ denote the probability of an individual age x in calendar year t dying within 1 year. If t denotes a future year, this probability must be estimated on the basis of past observations and hypotheses about future trends. For example, a projection formula frequently used in actuarial practice is as follows:

$$q_x(t) = q_x(\bar{t}) r_x^{t-\bar{t}} \quad (28)$$

where $q_x(\bar{t})$ refers to calendar year $\bar{t}$, for which a recent period life table is available, while r_x ($0 < r_x < 1$) denotes the annual reduction factor of mortality at age x and is estimated on the basis of past mortality experience. The hypothesis underlying equation (28), which is an extrapolation formula, is that the observed mortality trend will continue in future years.

A number of projection methods are at present available. For more information about forecasting mortality and projected life tables, the reader can refer, for example, to [8, 9], and references therein. A survey on mortality modeling in a dynamic context is provided by Pitacco [10].

From the $q_x(t)$'s, survival probabilities can be derived. For instance, for an individual age x in

calendar year t , the probability of being alive at age $x + h$ (i.e., in calendar year $t + h$) is given by

$${}_h p_x(t) = (1 - q_x(t))(1 - q_{x+1}(t+1)) \dots (1 - q_{x+h-1}(t+h-1)) \quad (29)$$

The actuarial value of a life annuity for a person age x in calendar year t is then given by

$$a_x(t) = {}_1 p_x(t)v + {}_2 p_x(t)v^2 + \dots \quad (30)$$

rather than by expression (3).

Deterministic versus Stochastic Mortality

When the probabilistic description of the mortality pattern is only used to derive expected values, a deterministic model is actually adopted. However, mortality within an insurance portfolio (e.g., number of policyholders dying in a given year, number of policyholders alive at maturity, etc.) is random, and this leads to riskiness in portfolio results (cash flows, profits, etc.). In a deterministic context, riskiness can be (roughly) faced *via* the use of prudential mortality assumptions, i.e., including a safety loading in premiums and reserves. However, the magnitude of the safety loading frequently does not rely on a sound risk assessment.

An appropriate alternative is provided by stochastic models. A stochastic approach to mortality recognizes various risk components. Consider, in particular, the following issues.

1. Randomness in mortality is first originated by random fluctuations of the frequency of death around its expected value. The risk of random fluctuations, or “process risk”, is a well-known component of risk in insurance. Its severity (conveniently assessed) decreases as the portfolio size increases, and this fact partially justifies deterministic models.
2. Mortality experienced throughout time may systematically differ from expected mortality. The risk of systematic deviations can be thought of as an “uncertainty risk”, meaning uncertainty in mortality assessment. In particular, future mortality may differ from projected mortality, originating the so-called longevity risk.
3. Finally, the “catastrophic” component of mortality risk consists in the possibility of a sudden and

short-term rise in mortality for example, because of an epidemic or a natural disaster.

Although the life table is the basic tool even for stochastic mortality, a stochastic model requires further inputs. In particular the following items are needed:

1. correlation assumptions among individual life-times (though in actuarial practice the independence hypothesis is commonly accepted);
2. a model representing systematic deviations (for example, consisting of a set of alternative scenarios for future mortality and a probability distribution over the set itself).

A stochastic model provides outputs such as the probability distributions of the number of deaths and related portfolio results (cash flows, profits, etc.), and in particular, single-figure indexes (expectations, variances, percentiles, etc.). Many applications of stochastic mortality models can be envisaged in the field of insurance risk management. For example:

1. pricing of insurance products (in particular, assessment of the safety loading);
2. choice of reinsurance arrangements;
3. capital allocation for solvency (possibly based on risk measures like *VaR* (namely the *value-at-risk*) and *TailVaR*).

Several features of stochastic models in both life and nonlife insurance are described in [11]. Longevity risk is analyzed, for example, by Olivieri [12] (*see also Stochastic Control for Insurance Companies; Estimation of Mortality Rates from Insurance Data*).

References

- [1] Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A. & Nesbitt, C.J. (1997). *Actuarial Mathematics*, The Society of Actuaries, Schaumburg.
- [2] Gerber, H.U. (1995). *Life Insurance Mathematics*, Springer-Verlag.
- [3] Forfar, D.O. (2004). Life table, in *Encyclopedia of Actuarial Science*, J.L. Teugels & B. Sundt, eds, John Wiley & Sons, Vol. 2, pp. 1005–1009.
- [4] Batten, R.W. (1978). *Mortality Table Construction*, Prentice Hall, Englewood Cliffs.
- [5] Benjamin, B. & Pollard, J.H. (1993). *The Analysis of Mortality and Other Actuarial Statistics*, The Institute of Actuaries, Oxford.

- [6] London, D. (1988). *Survival Models and Their Estimation*, ACTEX Publisher, Winsted.
- [7] Forfar, D.O. (2004). Mortality laws, in *Encyclopedia of Actuarial Science*, J.L. Teugels & B. Sundt, eds, John Wiley & Sons, Vol. 2, pp. 1139–1145.
- [8] Delwarde, A. & Denuit, M. (2006). *Construction de Tables de Mortalité Périodiques et Prospectives*, Economica, Paris.
- [9] Tabeau, E., van den Berg Jeths, A. & Heathcote, C. (eds) (2001). *Forecasting Mortality in Developed Countries*, Kluwer Academic Publishers.
- [10] Pitacco, E. (2004). Survival models in a dynamic context: a survey, *Insurance Mathematics and Economics* **35**, 279–298.
- [11] Daykin, C.D., Pentikäinen, T. & Pesonen, M. (1995). *Practical Risk Theory for Actuaries*, Chapman & Hall, London.
- [12] Olivieri, A. (2001). Uncertainty in mortality projections: an actuarial perspective, *Insurance Mathematics and Economics* **29**, 231–245.

ERMANNNO PITACCO

Insurance Pricing/Nonlife

The fundamental principles concerning insurance pricing in the nonlife area are essentially similar to the principles that are applied to price life insurance policies, but often with different emphasis on the constituent parts (e.g., correlation of risks, temporal longevity of risks, and effects of the capital market). Certain distinct characteristics of nonlife insurance policies suggest modifications of original techniques and have inspired creation of new approaches.

The essentials of pricing both life and nonlife policies can be summarized in the (oversimplified) statement that price equals the expected present value of the loss(es) times a load for expenses, times a load for profit, and times a load for risk bearing (see equation (1)).

An important area in which nonlife and life insurance pricing differ is the significance of regulatory intervention and social influences on pricing. Since certain nonlife policies are mandatory for purchase (e.g., individual automobile insurance, workers compensation), and others are required by contractual obligations (e.g., homeowners insurance on mortgaged homes, comprehensive collision insurance on financed automobiles), there is substantial support theoretically and socially for stronger regulatory control of pricing in nonlife insurance. Most regulatory environments impose the conditions that rates (or premium) cannot be inadequate (to assure solvency), excessive (to prevent excessive profits), or unfairly discriminatory (for social equity), and these considerations must be incorporated into prices.

One other major distinction between life and nonlife pricing is that nonlife risks usually have much shorter contract duration than life or annuity contracts. For example, an important nonlife insurance line, automobile insurance, is usually offered with coverage of 6 months or 1 year. Consequently, the effects of economic factors can differ in nonlife insurance pricing, where insurers often need to account for economic fluctuations much more aggressively than they do in pricing life insurance policies, which may have durations measured in decades.

The shorter duration of nonlife risk exposure means that long-term returns on assets cannot “bail you out” of short-term pricing errors. Thus, in nonlife pricing, financial approaches seem to be more appealing to insurers than some of the traditional supply

side actuarial methods that dominate the life market. Financial approaches are even used in regulatory hearings concerning what constitutes a “fair rate of return” for insurers (*cf.* [1]). Well-known financial pricing models such as the capital asset pricing model (CAPM) (*see Axiomatic Models of Perceived Risk*) and option pricing models have been introduced and tailored for use in pricing nonlife insurance products. As a result, the integration of financial pricing and traditional actuarial pricing has been an ever growing trend in both theory and practice.

In this article, both actuarial approaches and financial pricing methods are discussed for pricing nonlife insurance products. General (actuarial) ratemaking methods are first introduced, followed by a brief summary of important premium principles and the properties of these principles. Credibility theory, one of the core concepts in nonlife actuarial science, is then discussed. Lastly, we describe various financial pricing models and conclude with a comment on the integration of traditional insurance pricing and financial pricing.

General Ratemaking

The general framework for pricing risk in both life and nonlife settings can be illustrated *via* the formula

$$P = E \left[\sum_{i=1}^N L_i V^{T_i} \right] (1 + k_1)(1 + k_2)(1 + k_3) \quad (1)$$

where

L_i : severity (in dollars) of the i th claim,

N : number of claims incurred during the policy period,

$V : \frac{1}{1+r}$, the discount rate based on interest rate r ,

T_i : time until i th claim during policy coverage period,

k_1 : expense load (underwriting, marketing, claims adjusting and administration, commissions, taxes, general operating expenses),

k_2 : profit load (to obtain a “fair” rate of return), and

k_3 : load for risk bearing uncertainty.

Note that in life insurance, $N \equiv 1$, L is fixed at the policy face value, and T (and possibly V) is a random variable. In nonlife insurance, one can have multiple claims of difference sizes on the same policy, so N , L_i , V , and T_i can all be considered as random variables.

Manual Rates

“Manual rates” is the name for rates (or prices) determined using basic rating tables that instruct insurers how to price different classes of risks for a particular insurance product with a certain coverage. From these tables, one obtains a numerical value (say dollars to charge per \$1000 of potential coverage), which can be applied to supply a price for a designated risk within the class of risks covered by the table. Therefore, they are used to price risks in a homogeneous group of exposures, and not to tailor the premium to any specific individual. Two basic methods are used in developing manual rates: the pure premium method and the loss ratio method [2]. The pure premium method determines an indicated rate based on the exposure and expected loss. Mathematically, we have

$$R = \frac{P + F}{1 - V - Q} \quad (2)$$

where R is the indicated rate per unit of exposure, P is the expected loss per unit of exposure (pure premium), F is the fixed expense per exposure, V is the variable expense factor, and Q is a profit and contingency factor. Essentially, this formula says that the rate must be sufficient to pay the expected losses plus fixed expenses after allowing for variable expenses and required profits.

The loss ratio method is a premium adjustment method rather than a premium setting method. It determines the indicated rate changes based on the current rate (R_0) through multiplying R_0 by an adjustment factor (A). This adjustment factor (A) is defined as the ratio of the experience loss ratio (Y) to a prespecified target loss ratio (T). Mathematically, we have

$$R = AR_0 \quad (3)$$

where R_0 is the current rate, $A = \frac{Y}{T}$, $Y = \frac{L}{(R_0)(E)}$, $T = \frac{1-V-Q}{1+M}$, L is the experience losses, E is the earned exposure for the experience period, V is the premium-related expense factor, Q is the profit and contingency factor, and M is the ratio of non-premium-related expenses to losses. It is clear that, by construction, the loss ratio method is not applicable to new lines of business, where a current rate is unavailable. However, it allows the insurer to update last period's rates to obtain new rates for the next

period. Brockett and Witt [3] show that when current prices are set by a regression methodology based on past losses, and the loss ratio method is used to adjust prices, then an autoregressive series is obtained, thus partially explaining insurance pricing cycles. Although the pure premium method and the loss ratio method are constructed from different perspectives, it can be shown that the two methods are identical and would produce the same rates given the same data (for more mathematical details, see [2]).

Individual Risk Rating

Individual risks are usually rated by making changes to manual rates to account for individual characteristics. A basic approach is through the use of a credibility formula, where individual loss experience is incorporated with expected losses for the corresponding risk class to obtain a refined rate (this idea will be detailed in a later subsection). Two basic rating schemes are employed in individual risk rating, namely, prospective rating and retrospective rating [2].

Prospective rating uses experience of past losses to determine the future rate. An important type of prospective rating is experience rating, where the rate for a future period is determined using a weighted combination of experienced losses and expected losses, both from a prior period. Retrospective rating, on the other hand, uses current period experience to obtain the rate for the same period. The rate is also determined by combining individual experience and exposure, in which aspect it works essentially in the same way as experience rating. Thus, with a retrospective rating plan, a certain amount of premium can be rebated back if loss experience during this period is very good, whereas with experience rating, good loss experience affects future premiums. It should be noted that in both of the rating schemes, insurers need to choose an appropriate experience period to balance experience and standard expected rates. Besides the above-mentioned methods, another somewhat different method for individual risk rating is schedule rating. As a type of prospective rating, it makes adjustments to manual rates to reflect characteristics of individual risks that may affect future losses. Unlike other individual risk rating methods, it does not involve individual loss experience to achieve

this goal, but rather directly factors in characteristics known to affect the likelihood of losses (e.g., a sprinkler system in a warehouse may result in a multiplicative factor less than one being applied to the pertinent manual table rates).

Classification

Prices in nonlife insurance are usually set for a group of exposures having similar risk characteristics when individual experience is not credible enough, or it is not cost effective to rate or price the individual risk by itself. Being an integral part of the pricing process, this process (called *classification*) of dividing prospective risks into homogeneous groups is of importance in its own right and is the subject of legal, regulatory, and social constraints. More specifically, by classification, individual risks are grouped together and the same premium is charged to each member of the entire group (the premium will, however, be different for the same coverage for different groups). Grouping is based on multidimensional group characteristics that reflect different insurance costs, including difference in losses, expenses, investment income, and risk (deviation from the expected value). These group characteristics are usually called *rating variables* and they may differ across lines of insurance (e.g., in automobile insurance, the common rating variables are age, gender, and marital status). Various criteria are used to select these rating variables, e.g., actuarial criteria, operational criteria, and social criteria [2]. As competition among insurers becomes more intensive, insurers tend to constantly refine their choice of rating variables and classification to ensure availability of coverage and avoid adverse selection. For example, insurance credit scores based on the credit history of the individual being classified are now used in many automobile and homeowners insurance rating plans, as this variable has been found to be extremely predictive of losses and nonoverlapping with other traditional rating variables (cf. [4]).

Premium Principles

Traditional Premium Principles

(The concepts of premium principles (see **Premium Calculation and Insurance Pricing; Risk Measures**

and Economic Capital for (Re)insurers) are developed in great detail in [5], from which this section is drawn heavily. Interested readers are referred to [5] for more details.)

Certain traditional principles are employed in the process of premium calculation. Among the most commonly used ones are the expected value principle, the maximal loss principle, the variance principle, the standard deviation principle, the semivariance principle, the mean value principle, the zero utility principle, the Swiss premium calculation principle, the Esscher principle, and the Orlicz principle. These principles differ in the rationale they use to assign a (numeric) premium to a given risk. Each of them requires a certain potentially different level of knowledge of the risk distribution and may be appropriate only in certain situations. For the following discussion, denote the risk (random variable) as X and the premium for the risk X as $\Pi(X)$. Table 1 gives a brief description of the above-mentioned premium principles.

Properties of Premium Principles

In designing premium principles, a set of properties are usually desired based on intuitive grounds. We first introduce commonly discussed properties and then examine whether they are satisfied by the premium principles presented in Table 1 (note that many of the properties can be generalized to allow for more complicated considerations).

- Property 1* (Mean value exceeding property): The premium should exceed the expected claim, i.e., $\Pi(X) > E(X)$.
- Property 2.1* (Additivity): For independent risks X_1 and X_2 , a premium principle is additive if $\Pi(X_1 + X_2) = \Pi(X_1) + \Pi(X_2)$.
- Property 2.2* (Subadditivity): For independent risks X_1 and X_2 , a premium principle is subadditive if $\Pi(X_1 + X_2) \leq \Pi(X_1) + \Pi(X_2)$. Note that additivity implies subadditivity.
- Property 3.1* (Translation invariance): A premium principle is translation invariant if $\Pi(X + c) = \Pi(X) + c$ for any constant c .
- Property 3.2* (Subtranslativity): A premium principle is subtranslative if $\Pi(X + c) \leq$

Table 1 Basic premium principles

Name of premium principle	Mathematical definition	Description
Expected value principle	$\Pi(X) = (1 + \lambda)E(X)$	$E(X)$ is the expected value of risk X . λ is a positive safety loading. If $\lambda = 0$, it is called <i>net premium principle</i> .
Maximal loss principle	$\Pi(X) = pE(X) + (1 - p)\text{Max}(X)$	$\text{Max}(X)$ denotes the largest possible value of X . (This value only exists for bounded risks. However, this principle can be generalized.)
Variance principle	$\Pi(X) = E(X) + \lambda\text{Var}(X)$	$\text{Var}(X)$ is the variance of risk X . λ is a positive safety loading.
Standard deviation principle	$\Pi(X) = E(X) + \lambda\text{Stddev}(X)$	$\text{Stddev}(X)$ is the standard deviation of risk X . λ is a positive safety loading.
Semivariance principle	$\Pi(X) = E(X) + \lambda\text{PosVar}(X)$	$\text{Var}(X)$ is decomposed into $\text{PosVar}(X)$ (positive variance component) and $\text{NegVar}(X)$ (negative variance component) to capture the direction of deviation from mean. λ is a positive safety loading.
Mean value principle	$\Pi(X, f)$ solves $f(\Pi) = E(f(X))$	f is a continuous and strictly monotonic function on the domain of X (This is essentially to find the certainty equivalent of X if f is an utility function)
Zero utility principle	$\Pi(X, u)$ solves $E[u(\Pi - X)] = 0$	u is an utility function, which satisfies certain properties: u is continuous, nondecreasing ($u'(x) \geq 0$), and concave ($u''(x) < 0$), $u(0) = 0$ and $u'(0) = 1$. It is called <i>zero utility</i> since the premium is set such that accepting the risk for premium Π gives the insurer the same utility as not accepting the risk.
Swiss premium calculation principle	$\Pi(X, f, z)$ solves $E[f(X - z\Pi)] = f[(1 - z)\Pi]$	f is a continuous and strictly monotonic function. z is in $[0, 1]$. (If $f' \geq 0$, $f'' < 0$, $z = 1$, $f(0) = 0$, $f'(0) = 1$, then it returns a zero utility principle. If $z = 0$ then it returns a mean value principle.)
Esscher principle	$\Pi(X) = \frac{E(Xe^{\alpha X})}{E(e^{\alpha X})}$	$\alpha > 0$
Orlicz principle	$\Pi(X, f)$ solves $E\left[f\left(\frac{X}{\Pi}\right)\right] = f(1)$	$f' > 0$, $f'' > 0$

$\Pi(X) + c$ for any constant c . Note that translativity and subtranslativity are essentially identical.

Property 4 (Iterativity): A premium principle is iterative if $\Pi(X) = \Pi(\Pi(X|Y))$, where Y is the risk parameter (random variable) ($\Pi(X|Y)$ can be considered as a new risk based on Y).

Property 5 (Homogeneity): A premium principle is homogeneous if $\Pi(cX) = c\Pi(X)$ for any constant c . Moreover, it is called *positive homogenous* if the

above relation only holds for $c > 0$. It is called *symmetric* if the above relation only holds for $c = -1$.

Property 6 (Multiplicativity): For independent risks X and Y , a premium principle is multiplicative if $\Pi(XY) = \Pi(X)\Pi(Y)$.

Next, we examine each premium principle for these properties. A summary is presented in Table 2. From Table 2, we can see that some of the properties are more restrictive than others in the sense that

Table 2 Examination of properties of premium principles

Name	Mean value exceeding	Additivity	Subadditivity	Translativity (subtranslativity)	Iterativity	Homogeneity	Multiplicativity
Expected value principle	Yes	Yes	Yes	Yes, only when $\lambda = 0$ (net premium principle) Yes	Yes, only when $\lambda = 0$ (net premium principle) Yes, only when $p = 0$	Yes	Yes
Maximal loss principle	Yes (when $p \neq 1$)	Yes	Yes	Yes	Yes, only when $p = 0$	Yes, when $c > 0$ (positive homogeneous)	Yes, only when $p = 1$ (net premium principle) No
Variance principle	Yes	Yes	Yes	Yes	No	Yes, only when $\lambda = 0$ (net premium principle) Yes, when $c > 0$	No
Standard deviation principle	Yes	No	Yes	Yes	No	(positive homogeneous)	No
Semivariance principle	Yes	No	No	Yes	No	Yes, only when $\lambda = 0$ (net premium principle)	No
Mean value principle	Yes (when f is continuous, strictly increasing, and convex)	Yes, only when $f(X) = X$ (net premium principle) or $f(X) = \exp(\alpha X)$ (exponential principle)	Yes for positive subadditivity ($X, Y > 0$), only when f is continuous and strictly increasing, f' is concave	Yes, only when $f(X) = X$ (net premium principle) or $f(X) = \exp(\alpha X)$ (exponential principle)	Yes	Yes, under certain conditions on $f(x)$ (see [5] for details)	Yes, only when $f(X) = X$ (net premium principle)
Zero utility principle	Yes	Yes, only when $u(X) = X$ (net premium principle) or $u(X) = \exp(\alpha X)$ (exponential principle)	Yes, only when $u(X) = X$ (net premium principle) or $u(X) = \exp(\alpha X)$ (exponential principle)	Yes	Yes, only when $u(X)$ is linear or exponential	Yes, when $c > 0$ (positive homogeneous) under certain conditions on $u(x)$ (see [5] for details)	Yes, only when $u(X) = X$ (net premium principle)

(continued overleaf)

Table 2 (continued)

Name	Mean value exceeding	Additivity	Subadditivity	Translativity (subtranslativity)	Iterativity	Homogeneity	Multiplicativity
Swiss premium calculation principle	Yes (when f is continuous, strictly increasing, and convex)	Yes, only when $f(X) = X$ (net premium principle) or $f(X) = \exp(\alpha X)$ (exponential principle)	Yes, for positive subadditivity only when reduced to mean value principle and satisfy conditions specified for mean value principle	Yes, only when $f(X) = X$ (net premium principle) or $f(X) = \exp(\alpha X)$ (exponential principle)	Yes, only when it returns to mean value principle or zero utility principle with linear or exponential $u(X)$	Yes, under certain conditions on $f(x)$ (see [5] for details)	Yes, only when $f(X) = X$ (net premium principle)
Esscher principle	Yes, when $\alpha > 0$	Yes	Yes	Yes	Yes, only when $\alpha = 0$ (net premium principle)	Yes, when $c > 0$ (positive homogeneous) and when $\alpha = 0$ (net premium principle)	Yes, only when $\alpha = 0$ (net premium principle)
Orlicz principle	Yes	Yes, only when f is linear (net premium principle)	Yes, only when f is linear (net premium principle)	Yes, only when f is linear (net premium principle)	Yes, only when $f(x) = x^{1+\alpha}$, f'' is continuous	Yes, when $c > 0$ (positive homogeneous)	Yes, only when f is linear (net premium principle) (certain conditions of f)

fewer principles satisfy these properties. In practice, however, it is not necessary for the premium principle adopted to possess all the listed properties. The choice of premium principle will depend on the specific rating scenario.

Wang's Class of Premium Principles

Premium principles introduced in the last subsection are mostly motivated by the net premium concept ($\Pi(X) = E(X)$) or utility theory (see **Optimal Risk-Sharing and Deductibles in Insurance; Risk Measures and Economic Capital for (Re-)insurers; Mathematics of Risk and Reliability; A Select History; Clinical Dose–Response Assessment**). Another way to motivate the design of premium principles is to first delineate the desired properties and then search for the class of rules that satisfy these properties. An increasingly important class of premium principles named *Wang's class of premium principles* are motivated in this fashion. Wang [6] makes use of the survival function $S(x)$ ($S(x) = \Pr(X > x) = 1 - F(x)$, where $F(x)$ is the distribution function of X) to describe the layer premiums. Premium principles are desired to preserve the ordering of risks in the first order and the second order (or, “stochastic dominance preservative” (see **Risk Attitude**)) and to be additive for comonotonic risks (or, “comonotonic additive” (see **Credit Risk Models**)). Comonotonic risks refer to risks X_1 and X_2 that can both be represented as nondecreasing functions of risk Z , or mathematically, $X_1 = f_1(Z)$ and $X_2 = f_2(Z)$, where f_1 and f_2 are nondecreasing functions. The desired class of premium principles that possess the above two properties are obtained by transforming the survival function using a distortion function g , i.e., $g[S(x)]$, where g satisfies certain conditions as specified in [6] (e.g., g can be a utility function with $g(0) = 0$ and $g(1) = 1$). The premium principle is then defined as

$$\Pi(X) = \int_0^\infty g[S(x)] dx \quad (4)$$

This class of premium principles satisfies basic properties (e.g., mean value exceeding, homogeneity, and additivity) as well as the two particular properties emphasized above. The notion of a distortion function underlying the derivation of this class of premium principles has also been employed extensively in

developing risk measures and in other insurance applications.

Credibility of Data in Determining Rates

The idea behind credibility theory dates back to Mowbray in 1914 [7]. It is said to be “the casualty actuaries’ (see **Reinsurance**) most important and enduring contribution to casualty actuarial science” [8, p. 3]. The goal is to construct a method that incorporates individual level data together with standard or industry level data in a manner that allows for the individual data not to be unduly influential on the final price unless the amount of individual data is large enough for it to become “credible”. It is the quantification of this idea that constitutes the area of credibility theory.

Credibility theory has seen wide application in insurance pricing areas such as experience rating and ratemaking for new risks. It recognizes both the reliability of some form of class estimate and the additional value of heterogeneous individual risk experience, both of which together constitute an estimate for expected losses of the individual risk. Under credibility theory, a credibility factor z is assigned to the individual data, while a factor $(1 - z)$ is assigned to the class estimate. Mathematically, a credibility formula can be written as

$$C = z\hat{m} + (1 - z)M \quad (5)$$

where C is the premium or the best estimate of expected future loss for the individual risk, $\hat{m}$ represents the individual loss experience, M is the class or industry level estimate of loss or a prior estimate of loss, and z is the credibility factor. A higher credibility factor indicates that the particular individual data examined has a higher degree of accuracy in predicting future losses and thus is more important in premium calculation, i.e., more “credible”. For example, insurers are usually more confident in using experience of an individual risk with a large exposure to derive its premium than in the contrary case.

This rating method gives rise to the issue of estimating the credibility factor in order to come up with a final combined estimate for the individual risk. A natural way to decide credibility is to associate it with the variation of individual risk experience, i.e., individual data is more valuable when risk experience is relatively stable. To formalize the

theoretical construction and practice of credibility theory, two major branches of research were in place, namely, the limited fluctuation approach first proposed by Mowbray [7] and the greatest accuracy approach by Bühlmann [9, 10].

Recently, credibility theory has been developed for multidimensional cases and allows for nonlinear structures [11]. Moreover, the underlying connection between credibility formulae and empirical Bayesian (*see Risk in Credit Granting and Lending Decisions: Credit Scoring; Repairable Systems Reliability; Microarray Analysis*) approaches has been established. Advancement in modern Bayesian theory and computation will further facilitate the development and utilization of credibility theory (*cf.* [12, 13]).

Limited Fluctuation Approach

The main focus of limited fluctuation approach is to restrict the deviation of individual loss experience from the true underlying individual risk to grant full credibility to the individual risk data. More specifically, under certain assumptions on the losses, this method arrives at a required number of observations (e.g., exposure units) from the particular risk. Assume that we have, for the particular individual risk, independently identically distributed (i.i.d.) annual claim amounts $X_i, i = 1, \dots, N$. X_i 's have mean $\mu(\theta)$ and variance $\sigma^2(\theta)$, where θ is a (nonstochastic) parameter characterizing this particular individual risk. Let the individual estimate in the credibility formula be the average of the claims, i.e., $\hat{m}(X) = \frac{1}{N} \sum_{i=1}^N X_i$. Under limited fluctuation approach, we would like the difference between $\hat{m}$ (the estimate from individual data) and μ (the true mean of this individual risk) to be less than a chosen percentage (k) of μ with a large probability, or mathematically,

$$\Pr[|\hat{m}(X) - \mu(\theta)| \leq k\mu(\theta)] \geq p \quad (6)$$

where k is a positive small value and $0 < p < 1$, p is close to 1. Under Normal approximation, we derive

$$N \geq \frac{z_{(1+p)/2}^2 \sigma^2}{k^2 \mu^2} \quad (7)$$

where $z_{(1+p)/2}$ is the $(1+p)/2$ fractile of a standard normal distribution. The percentage k and confidence level p are parameters to be determined by the insurer. Formula (7) says that full credibility is

granted to the individual risk data when the condition on N is satisfied. Partial credibility can also be calculated by requiring the variance of component $z\hat{m}$ in credibility formula (5) to be less than the variance of individual estimate $\hat{m}$ when full credibility is warranted, through adjusting down the credibility factor z from 1.

The above procedure shows that the limited fluctuation approach is easy to implement and flexible in the sense that insurers can freely select parameters based on the specific rating scenario. However, this approach is subject to several limitations. Firstly, the case of full credibility is stressed by determining a large enough number of previous claims N . However, this approach does not provide solid theoretical justification of the decision of partial credibility factor, making it less useful in many situations when full credibility cannot be reasonably obtained [11]. Secondly, this approach fails to fully account for the quality of the class estimate M (treated as a constant). When the class estimate cannot accurately capture the individual risk, higher credibility should naturally be assigned to individual data, even when N is relatively small [2].

Greatest Accuracy Approach

Greatest accuracy approach (least-squares approach) was proposed initially by Bailey [14] and gained popularity through Bühlmann's work [9, 10]. Bühlmann suggested a nonparametric approach to come up with the credibility formula [9]. Suppose we have conditionally independent and identically distributed (i.i.d.) (given θ , the individual risk parameter, which is drawn from the parameter set Θ) annual claim amounts $X_1, \dots, X_N$ with mean $\mu(\theta)$ and variance $\sigma^2(\theta)$. The individual estimator $\hat{m}(X)$ is taken to be the average of the claims $\hat{m}(X) = \frac{1}{N} \sum_{i=1}^N X_i$. The idea of the (linear) least-squares approach is to find an estimator in the form of a linear function of $\hat{m}(X)$, which minimizes the expected squared error in estimating the premium for the individual risk. We arrive at the credibility formula (5) with

$$M = E\mu(\theta) \quad (8)$$

$$\hat{m} = \frac{1}{N} \sum_{i=1}^N X_i \quad (9)$$

and

$$z = \frac{N}{N + K} \quad (10)$$

$$K = \frac{E\sigma^2(\theta)}{\text{var}\mu(\theta)} \quad (11)$$

where M is the overall mean claim and $\hat{m}$ is the mean of the individual claim data for this particular risk. We can see that the credibility factor z as in equation (10) depends on the variation of the underlying risk premium $\text{var}\mu(\theta)$, variation of the individual claim observations $E\sigma^2(\theta)$, and the number of observations (N). Intuitively, when class estimate is more volatile or we have a larger amount of individual loss information, higher credibility will be assigned to the individual loss experience. When individual loss experience is more volatile, less credibility is assigned to it. In this way, the greatest accuracy approach provides both a theoretically justified and an intuitive way to combine individual loss data and the class estimate, taking into account their respective quality.

A further development of this approach considers different amount of risk exposure. In this case, we consider X_i 's as loss ratios (total amount of claims for year i divided by expected amount of claims (exposure) for year i) and the variance of X_i 's is allowed to depend on the risk exposure p_i for year i . The credibility formula takes the form of equation (5) with

$$M = E\mu(\theta) \quad (12)$$

$$\hat{m} = \frac{\sum_{i=1}^N p_i X_i}{\sum_{i=1}^N p_i} \quad (13)$$

and

$$z = \frac{\sum_{i=1}^N p_i}{\sum_{i=1}^N p_i + K} \quad (14)$$

$$K = \frac{E\sigma^2(\theta)}{\text{var}\mu(\theta)} \quad (15)$$

where p_i is the exposure (or size of the risk) for year i . We can see that the credibility factor z as defined in equation (14) depends on the total exposure for the risk. Larger exposure warrants higher credibility. This model is known as the *Bühlmann–Straub model* [15].

Financial Pricing of Insurance

As described in the introduction, financial pricing methods are widely used in nonlife insurance. In essence, insurance products can be viewed as a contingent future performance contract, where the future performance involves financial obligations with contingent claim payments. The trigger for the contingent claims is a qualified loss on the underlying risk event, such as loss owing to natural disasters, liability, etc. Thus, insurance pricing models can be constructed using pricing models for contingent claim contracts from the finance literature. These models are detailed below. (For a more complete summary of this topic, see [16].)

Insurance CAPM

The insurance CAPM views an insurance policy as a standard financial contract (e.g., a debt contract) and follows the rationale of the well-known CAPM. Papers by Fairley [17] and Hill [18] are among the first to discuss using the CAPM to address investment income and determine fair profit rates for property-liability insurers. The idea of the insurance CAPM is illustrated in a simple model that follows (notations here basically follow [19]). The insurer's net income is modeled as the sum of investment income and underwriting profit.

$$Y = I + \pi_p = r_A A + r_p P \quad (16)$$

where Y is net income, I is investment income, π_p is underwriting profit, r_A is (stochastic) return on assets, A is assets, r_p is (stochastic) return on underwriting, and P is insurance premium. On the basis of this setup, we can obtain a representation of the return on equity (ROE) as

$$r_E \left(= \frac{Y}{E} \right) = r_A + s(r_A k + r_p) \quad (17)$$

where r_E is the ROE, E is the equity, $s = \frac{P}{E}$ is the premium to surplus ratio, and $k = \frac{L}{P}$ is the funds generating factor (L is the liabilities and $L + E = A$).

Following the standard CAPM, expected return on asset i is modeled as risk-free rate plus expected individual risk premium for asset i , and expected individual risk premium is obtained by applying the beta coefficient for asset i to the expected market risk premium. Mathematically, we have

$$E(r_i) = r_f + \beta_i(E(r_m) - r_f) \quad (18)$$

where r_i is the (stochastic) return on asset i , r_f is the risk-free rate, r_m is the (stochastic) market return, and $\beta_i = \frac{\text{COV}(r_i, r_m)}{\text{var}(r_m)}$ is the beta coefficient of asset i .

To derive the insurance CAPM (i.e., the formula for calculating expected underwriting return), we first derive the underwriting beta as $\beta_p = \frac{\text{COV}(r_p, r_m)}{\text{var}(r_m)}$. Then by multiplying both sides of equation (17) with r_m and taking the covariance, we obtain, according to the formula for beta coefficients,

$$\beta_E = (1 + ks)\beta_A + s\beta_p \quad (19)$$

where β_E is the equity beta coefficient and β_A is the asset beta coefficient. Note that expected ROE, $E(r_E)$, can be derived by the CAPM equation (18) or by taking the expectation of equation (17). The resulting expressions from the above two approaches have to be equal. Thus by equating the two, we derive the insurance CAPM as

$$E(r_p) = -kr_f + \beta_p(E(r_m) - r_f) \quad (20)$$

The above calculation shows how an insurance policy is priced in the traditional CAPM framework. It is worth noting that instead of a positive risk-free return component (r_f) as in standard CAPM, the corresponding term in the insurance CAPM equation (20) has a negative coefficient, which stands for the part of rate rewarded back to the policyholders since the premium they paid at the beginning was used by the insurer to generate investment return for approximately k periods.

Despite its parsimony, the insurance CAPM is not likely to be a good candidate for use in practice. Besides general criticism that the CAPM has received (such as the insufficiency of using beta coefficients to explain returns [20]), other concerns have been raised for the insurance CAPM. These concerns include the use of fund generating factor, assumption of no bankruptcy, assumption of a constant interest rate, and estimation accuracy of underwriting beta [19].

Discounted Cash Flow Models

An alternative to the CAPM is the widely used discounted cash flow model (see **Credit Risk Models; Solvency**), which models price (present value) as a stream of future cash flows discounted at an appropriate interest rate. A simple version (in the discrete case) of this model in the context of insurance is presented below to illustrate the idea [19]. Assume a two period model with $t = 0, 1$. Premium is paid at $t = 0$ and loss payment is made at $t = 1$. Thus, according to the discounted cash flow model, premium is calculated as loss payment discounted by nominal interest rate

$$\Pi = \frac{L}{(1+r)} = \frac{L_0(1+\pi)}{(1+i)(1+r_r)} \quad (21)$$

where Π is premium, L is the value of promised loss payment at $t = 1$, L_0 is the present value of promised loss payment, π is the insurance inflation rate, r is the discount rate which can be decomposed into i , general inflation rate, and r_r , real interest rate. More realistic models taking into account other important cash flows of insurance companies (e.g., investment income, equity capital, tax payment) are built in a similar but much more complicated fashion (for example, see [21]). Discounted cash flow models have also been developed in the continuous-time setting. One class of these models are based on arbitrage pricing theory (APT), which is also a major asset pricing method in finance (for example, see [22]).

Option Pricing Models

The contingent nature of insurance claim payments suggests viewing insurance policies in the context of financial derivative contracts, e.g., options. The no-arbitrage pricing framework originally established for pricing derivatives can be adopted for nonlife insurance pricing. A simple description of the method (in the continuous case) is sketched below. Under this method, assets and liabilities are assumed to follow stochastic processes (e.g., geometric Brownian motion (see **Risk-Neutral Pricing; Importance and Relevance; Default Correlation**)). Thus, the basic mathematical representation for the asset process is

$$dA_t = r_A A_t dt + \sigma_A A_t dW_{tA} \quad (22)$$

and for the liability process is

$$dL_t = r_L L_t dt + \sigma_L L_t dW_{tL} \quad (23)$$

where A_t is assets, L_t is liabilities, r_A and r_L are drifts for assets and liabilities respectively, σ_A and σ_L are volatilities for assets and liabilities respectively, and W_{tA} and W_{tL} are standard Brownian motions. Built upon the underlying stochastic processes, characteristics of various contingent claims on the insurance companies can be captured by modeling them as put–call options, and premium (or other value of interest) can then be found by standard mathematical finance techniques (e.g., Black–Scholes formula [23] (*see Equity-Linked Life Insurance; Statistical Arbitrage; Asset–Liability Management for Life Insurers; Risk-Neutral Pricing: Importance and Relevance; Credit Scoring via Altman Z-Score; Statistical Arbitrage*)).

Direct use of the above model is limited in practice owing to the imposed restrictions, e.g., the use of European options. A few special cases where this model is suitable, such as premium calculation for the guaranty fund, are listed in Cummins [19]. However, this baseline model can be extended in many dimensions. For example, multiple liability processes can be considered, since many insurers write multiple lines of business rather than just offer a single product. Exotic options rather than standard European options can be used to model the underlying insurance processes.

Option pricing methods are considered especially useful in pricing catastrophic risks (e.g., earthquake, hurricane). Because of the unique nature of catastrophic risks (e.g., extremely large losses, high loss correlations), insurers seek capital market to finance their products. As a result, catastrophic bonds, futures, and options (e.g., PCS catastrophic insurance options (*see Actuary; Operational Risk Modeling*)) introduced by Chicago Board of Trade (CBOT)) are created to handle these risks. The potential value of using option pricing methods in this particular area is thus self-evident (*cf.* [24]).

In summary, the introduction of financial pricing methods have undoubtedly provided insurers with a much larger and potentially more effective toolkit for pricing nonlife insurance risks. However, limitations do exist in the current insurance financial methods and many areas in finance are yet to be explored.

Integration of Financial Pricing and Insurance Pricing

In a broad sense, insurance pricing and financial pricing are integrated in a no-arbitrage pricing framework. As discussed extensively in mathematical finance theory and applications, no-arbitrage pricing finds the expectation of the discounted value of the contingent claim (e.g., loss payments) under the risk neutral measure, i.e., the probability measure that makes the underlying stochastic process a martingale. Under this risk neutral measure, adverse events are assigned higher weights than under the original physical measure. Thus, the risk neutral measure incorporates the idea of safety loading in insurance pricing [25]. In traditional financial economics, this type of pricing scheme works well since the financial market is often considered as a complete market and thus a replicating portfolio can be formed to follow the rationale of Black and Scholes [23].

When this idea is applied to the field of insurance, however, a major challenge emerges: the insurance market is oftentimes illiquid and incomplete, since many of the underlying risks are nontradable and indivisible. As a result, it is often impossible to find a unique risk neutral measure. A method proposed for pricing in an incomplete market is “indifference pricing” [26]. The idea is to maximize expected utilities with and without writing the risk at premium Π , through which a value of Π can be obtained such that the insurer is indifferent between writing the policy and not doing so. This process incorporates investor’s risk preference (captured by the utility function), which cannot be eliminated as in the traditional setting owing to market incompleteness.

However, when we are able to find appropriate risk neutral measures and price insurance claims in the no-arbitrage framework, we see that traditional premium principles (e.g., expected value principle and Esscher principle) derived from actuarial approaches can be recovered by the proper choice of risk neutral measures, which validates the interrelation between these two approaches [25].

The interactions between the areas of insurance and finance have greatly advanced the development of pricing techniques in insurance. Inventions of new instrument products in this process may potentially complete the market and thus make standard finance models more accessible and applicable in insurance pricing. Continuous unification of insurance pricing

and financial pricing is an inevitable trend that will bring new insights and provide benefits to both fields.

References

- [1] Cummins, J.D. & Harrington, S.A. (eds) (1987). *Fair Rate of Return in Property-Liability Insurance*, Kluwer Academic Publishers, Hingham.
- [2] Casualty Actuarial Society (1990). *Foundations of Casualty Actuarial Science*, New York.
- [3] Brockett, P.L. & Witt, R.C. (1982). The underwriting risk and return paradox revisited, *Journal of Risk and Insurance* **49**, 621–627.
- [4] Brockett, P.L. & Golden, L.L. (2007). Biological and psychobehavioral correlates of credit scores and automobile insurance losses, toward an explication of why credit scoring works, *Journal of Risk and Insurance* **74**, 23–63.
- [5] Goovaerts, M.J., de Vylder, F. & Haezendonck, J. (1984). *Insurance Premiums, Theory and Applications*, North-Holland, Amsterdam.
- [6] Wang, S. (1996). Premium calculation by transforming the premium layer density, *ASTIN Bulletin* **26**, 71–92.
- [7] Mowbray, A.H. (1914). How extensive a payroll exposure is necessary to give a dependable pure premium, *Proceedings of the Casualty Actuarial Society* **1**, 24–30.
- [8] Rodermund, M. (1990). Introduction, in *Foundations of Casualty Actuarial Science*, Casualty Actuarial Society, New York.
- [9] Bühlmann, H. (1967). Experience rating and credibility, *ASTIN Bulletin* **4**, 199–207.
- [10] Bühlmann, H. (1969). Experience rating and credibility, *ASTIN Bulletin* **5**, 157–165.
- [11] Norberg, R. (2004). Credibility theory, in *Encyclopedia of Actuarial Science*, J.L. Teugels & B. Sundt, eds, John Wiley & Sons, Chichester.
- [12] Makov, U.E., Smith, A.F.M. & Liu, Y.-H. (1996). Bayesian methods in actuarial science, *The Statistician* **45**, 503–515.
- [13] Klugman, S.A. (1982). *Bayesian Statistics in Actuarial Science with Emphasis on Credibility*, Kluwer Academic Publishers, Norwell.
- [14] Bailey, A.L. (1945). A generalized theory of credibility, *Proceedings of the Casualty Actuarial Society* **32**, 13–20.
- [15] Bühlmann, H. & Straub, E. (1970). Glaubwürdigkeit für Schadensätze, *Mitteilungen der Vereinigung Schweizerischer Versicherungsmathematiker* **70**, 111–133.
- [16] Cummins, J.D. & Phillips, R.D. (2000). Applications of financial pricing models in property-liability insurance, in *The Handbook of Insurance Economics*, G. Dionne, ed, Kluwer Academic Publishers, Boston.
- [17] Fairley, W. (1979). Investment income and profit margins in property-liability insurance, theory and empirical tests, *The Bell Journal of Economics* **10**, 192–210.
- [18] Hill, R. (1979). Profit regulation in property-liability insurance, *The Bell Journal of Economics* **10**, 172–191.
- [19] Cummins, J.D. (1991). Statistical and financial models of insurance pricing and the insurance firm, *Journal of Risk and Insurance* **58**, 261–302.
- [20] Fama, E. & French, K. (1992). The cross-section of expected stock returns, *Journal of Finance* **47**, 427–465.
- [21] Taylor, G. (1994). Fair premium rating methods and the relations between them, *Journal of Risk and Insurance* **61**, 592–615.
- [22] Kraus, A. & Ross, S. (1982). The determination of fair profits for the property-liability insurance firm, *Journal of Finance* **33**, 1015–1028.
- [23] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.
- [24] Froot, K. (ed) (1999). *The Financing of Catastrophic Risk*, The University of Chicago Press, Chicago.
- [25] Embrechts, P. (2000). Actuarial versus financial pricing of insurance, *The Journal of Risk Finance* **1**, 17–26.
- [26] Musiela, M. & Zariphopoulou, T. (2004). An example of indifference prices under exponential preferences, *Finance and Stochastics* **8**, 229–239.

JING AI AND PATRICK L. BROCKETT

Integration of Risk Types

Traditionally, financial institutions manage and therefore, also measure their risk in different departments, usually the credit, marketing, and operations department. The corresponding risks are credit risk, market risk, and operational risk (OR). Some institutions also measure other risks like business risk, liquidity risk, private equity risk and environmental risk, among others. On the other hand, these institutions need an aggregate unified view of their risk positions. Therefore, some procedures to integrate those risks are in place. The simplest one is, assuming that in each type the risk is finally measured in some monetary unit, like economic capital, to add all these figures to obtain an overall monetary risk measure. However, this totally ignores the concept of diversification across risk types. To obtain a more accurate picture of the integrated risk, more sophisticated methods are necessary. The two basic approaches are the so-called bottom-up and top-down [1–3]. Both are possible since risk measurement follows the same basic route in all risk types. This will be explained in the following section on the common features of risk measurement, whereas in sections titled “Top-Down Approach” and “Bottom-Up Approach” the two integration approaches are described in greater detail. Since the bottom-up approach is much more challenging and usually only achieved for specific departments of a bank, it is only briefly commented upon. Our focus is not on the risk models in the different risk types. Hence, the credit, market, or operational risk models are not covered in detail, but emphasis is more on their potential integration. For the different risks, we the reader may refer to more specific articles in this encyclopedia or in the cited literature.

Common Features of Risk Measurement

The basic common features can be described by means of the loss variable denoted by $L_{\text{type}}(\omega)$, the loss in risk type, $\text{type} \in \{\text{market}, \text{credit}, \text{operational} \dots\}$ (cf. [4, 5]) in state of the world $\omega \in \Omega$. It is modeled on an underlying probability space $(\Omega, \mathcal{A}, P)$. Then we usually have

$$L_{\text{type}}(\omega) = f_{\text{type}}(\vec{X}(\omega)) \quad (1)$$

Here the loss variable L can refer to a single transaction, a subunit of the bank, or the bank’s entire loss distribution in risk type “type”. Here, we assume that the vector $\vec{X}$ consists of all factors used in any risk calculation and $f_{\text{type}}(\vec{x})$ gives the functional form of the change in value in risk type “type” in state of the factor $\vec{x}$. f can also be thought of as a valuation function; it only remains to specify the joint distribution of the vector $\vec{X}$.

Example

1. The so-called delta–gamma approach in market risk (as in [4, 6]) assumes a lognormal distributed X and takes a second-order approximation of the value function F , i.e., the loss is approximated by

$$L_{\text{market}}(\omega) = \Delta \cdot \vec{X}(\omega) + \vec{X}^T(\omega) \Gamma \cdot \vec{X}(\omega) \quad (2)$$

Here, Δ is the gradient of the value function of the position with respect to the underlying risk factors, i.e., the i th component

$$\Delta_i = \sum_{k=1}^K \Delta_{i,k} \quad (3)$$

where $\Delta_{i,k}$ is the sensitivity of transaction k to risk factor i . The matrix Γ denotes the second derivatives of the transactions with respect to the risk factors. The factors $\vec{X}$ are usually equities, Foreign Exchange (FX), interest rates, implied volatility, and other market variables.

2. In credit risk in a simple default mode (cf. [7, 8] or [9]), one would take

$$f_{\text{credit}}(\vec{x}) = \sum_{i=1}^m \mathbf{1}_{\{x_i \leq C_i\}} \quad (4)$$

and therefore

$$L_{\text{credit}}(\omega) = f_{\text{credit}}(\vec{X}(\omega)) = \sum_{i=1}^m \mathbf{1}_{\{X_i(\omega) \leq C_i\}} \quad (5)$$

A typical modeling assumption here is that the vector $\vec{X}$ represents the asset values of the counterparties.

3. OR (*cf.* [4, 5]) usually assumes an insurance-type loss distribution (*see* **Large Insurance Losses Distributions**)

$$L_{\text{OR}} = \sum_{i=1}^N S_i \quad (6)$$

where N is the number of events in a certain OR type and S stands for the severity of an event. There are many attempts to define and calibrate risk factors, often called *risk indicators*, $\vec{X}$ such that $N = N(\vec{X})$ and $S_i = S_i(\vec{X})$. However, this development is currently in a very preliminary stage and usually integration with OR is done by the copula version of the top-down approach.

Integration can now be done at different levels. A full integration means that the entire loss distribution is modeled as a function f_{total} of the entire vector of the risk factors. In this most challenging approach, the entire set of risk factors is integrated, as well as the valuation function f of the portfolio on this vector, i.e.,

$$L_{\text{total}}(\omega) = f_{\text{total}}(\vec{X}(\omega)) \quad (7)$$

Since the portfolio loss is always the sum of the losses in the individual transactions

$$L_{\text{type}} = \sum_{\text{all transactions}} L_{\text{type}}^{\text{transaction}} \quad (8)$$

this approach already requires an integration of the loss function on transaction level

$$L_{\text{total}}^{\text{transaction}}(\omega) = f_{\text{total}}^{\text{transaction}}(\vec{X}(\omega)) \quad (9)$$

Therefore, it is usually called a *bottom-up integration*. Top-down integration means that integration is carried out at the level of portfolio loss variables. Here the basic assumption is that the loss variables are additive.

Top-Down Approach

In the top-down approach, the financial institution measures the different risk types separately in different departments and systems. As a result, the loss distribution L_{type} is specified at the institution-wide level, but separately for each risk type. The total loss

distribution is then the distribution of the random variable

$$L_{\text{total}} = \sum_{\text{all types}} L_{\text{type}} \quad (10)$$

The associated function f_{total} is as such the sum of the f_{type} 's. An important feature of this approach is that the loss variable is additive in the risk types. Intuitively, this means that the total loss is always the sum of the losses due to credit risk, market risk, and other risks. One can therefore also refer to this as “additivity of risk types”. In the bottom-up approach, this additivity does not hold since the loss or change in value is a possible nonlinear function of the entire vector of risk drivers. Separation into credit and market risk is not possible in such a case.

Factor Approach

If the bank can identify the entire factor space $\vec{X}$ driving all risk types and specify the joint distribution, then the entire loss distribution is given by

$$L_{\text{total}}(\omega) = \sum_{\text{all types}} f_{\text{type}}(\vec{X}(\omega)) \quad (11)$$

In some cases, there might be risk factors that drive, for example, credit and market risks. In many cases, the factor spaces are modeled separately and the joint distribution of $\vec{X}$ is specified. Then often a copula approach (*cf.* [5, 10–12]) at the level of the factors is implemented. An obvious challenge in this approach is to minimize the number of factors.

Copula Approach

If the underlying factor vector $\vec{X}$ cannot be specified and only the marginal distributions of L_{type} for all risk types are known, the joint distribution of the vector

$$\vec{L} = (L_{\text{type}})_{\text{type} \in \{\text{all types}\}} \quad (12)$$

can be specified by the copula approach:

$$\vec{F}(\vec{l}) = C(F^{(1)}(l_1), \dots, F^{(m)}(l_m)) \quad (13)$$

where $\vec{F}$ is the distribution function of $\vec{L}$, C is the copula function, m is the number of risk types, and

$F^{(j)}$ is the distribution function for losses in risk type j , $j = 1, \dots, m$.

Example

Let us exhibit a concrete, realistic example from [1] for the integration of credit and market risks. The current mark-to-market of the bank equals 100 and the current credit exposure equals 10 000. The assumption on market risk is that the percentage change over 1 year is normally distributed with an annual volatility of 20% and mean zero. The 99% value-at-risk turns out to be 46.6. As a plausible credit loss distribution, the Vasicek distribution, which can be found in [13] or [14], with $p = EL = 0.003$ (30 basis points annual expected loss) and an asset correlation of $\rho = 12\%$ is used. The 99% percentile of the corresponding credit loss distribution minus the expected loss, that is, the economic capital for credit risk, is then 162.5.

In the traditional nonintegrated calculation, the overall “integrated” risk is just the sum of the two (i.e., 100% correlation is assumed). With a correlation assumption of 50% for the normal copula, there is around 10% reduction in economic capital. This is only a small benefit owing to diversification across risk types, since credit risk dominates in any case. If one changes the Mark-to-Market (MtM), the percentile, and also the correlation, one obtains the figures in Table 1. Here $\rho_{M,C}$ is the correlation in the normal copula between credit and market risks. Diversification benefit is the decrease in integrated risk compared to the simple addition of credit and market risks. Compare [1]: the largest benefit (28%) can be observed, if we have roughly equal risk capital for market and credit risks, market risk = 255, and credit risk = 206, and a small correlation $\rho_{M,C} = 0.3$.

Table 1 Examples of top-down integration

Percentile	$\rho_{M,C}$	MtM	Market risk	Diversification benefit (%)
0.99	0.5	100	46	10
0.99	0.3	500	230	25
0.995	0.5	100	51	7
0.995	0.3	500	255	28

Remarks

1. Traditionally, market risk in the trading area is measured at a 1- or 10-day value-at-risk (see **Equity-Linked Life Insurance; Credit Scoring via Altman Z-Score**) and credit-value-at-risk (see **Extreme Value Theory in Finance**), and OR usually on a 1-year horizon. For integration purposes, there are different ways to solve the problem with different time horizons. Since it is very difficult to consider credit risk at a shorter horizon, one usually tries to measure market risk on a 1-year horizon. As a compromise a 3-month horizon can also be taken for the integration.
2. The effect of dynamic portfolios whose modeling is a great challenge is currently ignored in the risk measurement models.

Bottom-Up Approach

Here, the change in value T_i of a transaction i is a function of the risk factors denoted by $\vec{x}$, the entire universe of risk factors. As a typical example, consider what is called *wrong-way exposure* in credit risk management (cf. [15]). Assume e.g., that the bank has bought a put on the Nikkei from a large Japanese bank, which carries a large weight in the Nikkei. As the Japanese bank deteriorates in credit quality, the Nikkei will drop in value, making the put more valuable. But at the same time, the probability that the Japanese bank will not fulfill its contingent payment (see **Actuary; Asset-Liability Management for Life Insurers; Equity-Linked Life Insurance**) in the put will also increase. The effect on the overall credit adjusted price of this defaultable option is not clear and in no way additive. Only an integrated analysis of the effect will reveal the proper price. Since this is more a valuation problem rather than a risk measurement topic, it is not elaborated further, but the reader may refer to e.g., [1, 2, 16].

References

- [1] Overbeck, L. (2006). Integration of credit and market risk, in *Risk Management: A Modern Perspective*, M. Ong, ed, Academic Press.
- [2] Jobst, N., Mitra, G., Stavros, A. & Zenios, A.S. (2006). Integrating market and credit risk: a simulation and optimisation perspective, *Journal of Banking and Finance* 30(2) 717–742.

- [3] Kiesel, R., Liebermann, T. & Stahl, G. (2006). Mathematical framework for integrating credit and market risk, in *Risk Management: A modern Perspective*, M. Ong, ed, Academic Press.
- [4] Crouhy, M., Galai, D. & Mark, R. (2000). *Risk Management*, McGraw-Hill.
- [5] McNeil, A., Frey, R. & Embrechts, P. (2006). *Quantitative Risk Management*, Princeton University Press.
- [6] Risk Metrics (1996). *Technical Document*, <http://www.riskmetrics.com>.
- [7] Bluhm, C., Overbeck, L. & Wagner, C. (2003). *An Introduction to Credit Risk Modeling*, Chapman & Hall/CRC.
- [8] Merton, R. (1974). On the pricing of corporate debt: the risk structure of interest rates, *The Journal of Finance* **29**, 449–470.
- [9] Credit Metrics (1997). *Technical Document*, J.P. Morgan.
- [10] Nelsen, R. (1999). *An Introduction to Copulas*, Springer, New York.
- [11] Cherubini, U., Luciano, E. & Vecchiato, W. (2004). *Copula Methods in Finance*, John Wiley & Sons.
- [12] Frey, R. & Nyfeler, M. (2001). Copulas and credit models, *Risk* **14**(10), 111–114.
- [13] Vasicek, O. (2002). Loan Portfolio Value, *Risk* **15**(12), 160–162.
- [14] Gordy, M.B. (2003). A risk-factor model foundation for ratings-based bank capital rules, *Journal of Financial Intermediation* **12**, 199–232.
- [15] Fingers, C. (2000). *Toward a Better Estimation of Wrong-Way Credit Exposure*, Risk Metrics Group, <http://www.riskmetrics.com>.
- [16] Deventer, D.R., Kenji Imai, K. & Mesler, M. (2004). *Advanced Financial Risk Management: Tools and Techniques for Integrated Credit Risk and Interest Rate Risk Management*, John Wiley & Sons.

Related Articles

Actuary

Asset–Liability Management for Life Insurers

Equity-Linked Life Insurance

LUDGER OVERBECK

Intensity Modeling: The Cox Process

A Cox process can be thought of as a Poisson process with a random intensity where the intensity is exogenous to the jump process, i.e., the occurrence of jumps is affected by the level of the intensity but the occurrence of jumps does not affect the intensity. This is in contrast, for example, with the self-exciting point processes, or Hawkes process, in which the occurrence of jumps leads to increased intensity. Here we focus on the one-dimensional case in which the process can be thought of as a simple counting process, but applications of the same construction are useful for more general point processes, for example, in modeling spatial patterns (*see Managing Infrastructure Reliability, Safety, and Security; Global Warming; Vulnerability Analysis for Environmental Hazards; Digital Governance, Hotspot Detection, and Homeland Security*).

Construction and Examples

The mathematical construction of a Cox process is most easily understood as a random time change of a standard Poisson process. Let λ denote a nonnegative stochastic process and define the nondecreasing process Λ as $\Lambda_t = \int_0^t \lambda(s) ds$. Assume that N is a Poisson process with intensity 1 that is independent of λ . Now define

$$Y(t) = N(\Lambda(t)) \quad (1)$$

We then say that Y is a Cox process with intensity process λ . We could have started directly with a nondecreasing process Λ and not insisted that it should be pathwise absolutely continuous with respect to Lebesgue measure, but here we focus on the intensity-based formulation, which is by far the most important for applications.

Some important special cases of Cox processes are as follows:

1. A Poisson processes with a deterministic, time-varying intensity. This corresponds to the special case where λ is a deterministic function of time. In queuing theory, (*see Large Insurance Losses*

Distributions) for example, the arrival rate of customers to a service stations often follow time patterns that can be described deterministically.

2. A mixed Poisson process. This corresponds to the case where each path of λ is constant but where the constant level is random. Here we can think of λ as a random variable. This is typically applied as a way of modeling arrivals of events in populations with heterogeneity.
3. A Markov-modulated Poisson process. In this case, the intensity process λ is controlled by a continuous-time Markov chain X , which for analytical tractability is typically assumed to have a finite state space. With a slight abuse of notation we write the intensity as a function of the level of the Markov chain as $\lambda_t = \lambda(X_t)$ so that $\Lambda(t) = \int_0^t \lambda(X(s)) ds$. This modeling framework has applications in a variety of fields. In queuing theory, the intensity of service may fluctuate owing to random breakdowns in components, in credit risk modeling the intensity of defaults may fluctuate owing to changes in business cycles, etc.
4. Intensity modulated by a diffusion or jump diffusion. This is similar to the Markov-modulated case except that the intensity is now modeled as a diffusion process or a diffusion with jumps.

Moments and Overdispersion

The assumption of independence between the intensity and the underlying Poisson process facilitates computation greatly since, after conditioning in the intensity, we can use properties of the Poisson process. For example, the mean and variance of Y are easily computed through conditioning:

$$\begin{aligned} EY(t) &= E(N(\Lambda(t))) = E(E(N(\Lambda(t))|\Lambda(t))) \\ &= E\Lambda(t) \end{aligned} \quad (2)$$

$$\begin{aligned} VY(t) &= E(V(N(\Lambda(t))|\Lambda(t))) \\ &\quad + V(E(N(\Lambda(t))|\Lambda(t))) \\ &= E\Lambda(t) + V\Lambda(t) \end{aligned} \quad (3)$$

where we have used the fact that the mean and variance of the unit rate Poisson process $N(t)$ are both equal to t . We note that the ratio of the variance

to the mean for the Cox process is greater than 1, which is the ratio of a Poisson process, regardless of its intensity. Consequently, the Cox process is said to display overdispersion relative to the Poisson process. This overdispersion merely reflects the fact that, in periods with low intensity, there are relatively few occurrences, whereas when the intensity is high there are many occurrences. Consistent with this observation, we note that the degree of overdispersion increases when the variability of λ , and hence of Λ , increases. Similar calculations show that the covariance between the number of occurrences in disjoint intervals is positive and the correlation increases with the variability in λ . In general, the moments of the Cox process are characterized by the moment properties of the directing random measure specified by the directing intensity. For more on this, see for example Daley and Vere-Jones [1].

Inference and State Estimation

The theory of inference and state estimation for point processes (*see* **Extreme Value Theory in Finance; Reliability Growth Testing**) developed strongly in the early 1970s. Typical problems involve the estimation of the directing intensity from observation of the jumps, for example, with the purpose of detecting change points. The dynamics of the underlying intensity may be known or may be parametric and parameters estimated as part of the exercise. The literature is very large and we refer to books by Snyder [2], and for the martingale methods refer to Brémaud [3], which contains an introduction and a comprehensive list of references.

A nice example of state estimation from observed occurrences in which solutions can be worked out explicitly is the case with a Markov-modulated intensity as studied in Rudemo [4].

An interesting aside is that one cannot always tell from the observed jumps of a point process whether the underlying structure is actually a Cox process structure. An example of this is attributed to Kendall (for a textbook treatment, see Resnick [5]), who showed that the linear birth process in which the intensity of new births increases linearly with the number of individuals in the population (i.e., a case where the intensity is not independent of the occurrence of jumps) can in fact be represented by a mixed inhomogeneous Poisson process where

the mixing variable is exponentially distributed and mixes between deterministic, exponentially increasing intensities. Other examples exist in which heterogeneity and contagion cannot be separated merely from observation of the Poisson process. A survey of some of the early contributions in this area can be found in Taibleson [6].

Example: Credit Risk Modeling

In this section we present a brief outline of an application (*see* **Credit Migration Matrices; Default Risk; Mathematical Models of Credit Risk; Copulas and Other Measures of Dependency**) of Cox processes in the area of financial economics and risk management. The largest source of risk for a typical commercial bank is the risk that borrowers default on loans granted by the bank. For a bank involved in trading of financial derivatives (*see* **Enterprise Risk Management (ERM); Actuary; Numerical Schemes for Stochastic Differential Equation Models; Default Correlation**), a similar important source of risk is the risk of default of a counterparty to a derivative contract with positive value to the bank. Quantifying such risks is a typical example of what is studied under the area broadly referred to as *credit risk modeling*. The fundamental problem in credit risk modeling is to model the price of corporate bonds, i.e., bonds issued by corporations. A corporate bond typically promises to pay coupons and repay principal in a manner similar to a government bond, but (in countries where we can think of a government bond as safe) the value of the bond is lower than that of the government bond. How much lower it is depends mainly on two quantities: the default probability of the firms and the recovery in the event of default, i.e., the fraction of the promised payment that is received in the event of default. In this presentation, we ignore the modeling of recovery and for simplicity treat it as zero, i.e., in the event of default on a corporate bond, the holder of the bond receives nothing. We focus now on why Cox processes are a convenient tool for modeling defaultable corporate bonds, emphasizing the point that the Cox process formulation allows us to apply tools developed for modeling the pricing and dynamic evolution of government bonds.

A fundamental relationship in the modeling of government bonds is the relationship between the

short rate of interest and the prices of zero-coupon bonds (see **Equity-Linked Life Insurance; Credit Migration Matrices; Mathematical Models of Credit Risk**). Zero-coupon bonds are bonds that pay no coupons before maturity but repay the principal at maturity. They constitute a basic building block for modeling all types of government bonds. The short rate is associated with a traded financial instrument, referred to as the *money market account*, which transforms a unit of account invested at time t into $\exp(\int_t^u r_s ds)$ units of account at time u . A fundamental result in financial economics states that for an economy with zero-coupon bonds trading (having maturities less than some terminal date T) under certain conditions there exists a probability measure Q such that for all pairs of dates $t, u < T$, the price at date t of a zero-coupon bond maturing at date u is given by

$$p(t, u) = E_t^Q \exp \left(- \int_t^u r(s) ds \right) \quad (4)$$

where E^Q denotes expectation taken under the measure Q and the subscript refers to information available at time t , which could be the sigma field generated by the factors driving the short rate up to time t .

Now turning to corporate zero-coupon bonds with zero recovery, we let the default time τ of the corporate bond issuer be modeled as the first jump of a Cox process with intensity process λ , i.e., τ is modeled as

$$\tau = \inf \left\{ t : \int_0^t \lambda(s) ds \geq E_1 \right\} \quad (5)$$

where E_1 is an exponential random variable with mean 1 (under the measure Q). To see how Cox processes bridge the link from corporate bonds to government bonds, consider the price $v(0, t)$ of a zero-coupon bond maturing at t with zero recovery in default and issued by a company with default intensity λ under the risk-neutral measure Q :

$$\begin{aligned} v(0, t) &= E^Q \left[\exp \left(- \int_0^t r(s) ds \right) 1_{\{\tau > t\}} \right] \\ &= E^Q \left[\exp \left(- \int_0^t (r + \lambda)(s) ds \right) \right] \end{aligned} \quad (6)$$

The functional expression for the defaultable bond now has the same form as that of a government bond but with an “adjusted” short rate $r + \lambda$ replacing the short rate r . This in turn means that explicit pricing formulas that exist for government bonds can easily be extended to corporate bonds; see for example Lando [7]. The explicit formulas can be obtained when r and λ belong to the class of so-called affine models, which are jump-diffusion processes in which the drift term, the squared diffusion term, and the jump intensity are all affine in the state variables. For a precise statement explaining this and elaborating on the computational advantages, see Duffie *et al.* [8]. Note that all these computational results pertain to computation of survival probabilities of the form $E \exp(-\int_0^t \lambda(s) ds)$ as well and therefore greatly improve our ability to perform calculations in diffusion and jump-diffusion (see **Lévy Processes in Asset Pricing; Default Risk**) driven Cox process models.

Cox-type regression models may be used to estimate the dependence of default intensities on covariates including company-specific key ratios (such as the amount of debt issued compared to the value of assets) and macroeconomic indicators (such as growth in gross domestic product (GDP)). The techniques are similar to those of survival analysis in biostatistics. However, for the models to be useful for pricing and for dynamic risk management, a full modeling of the stochastic behavior of the covariates is necessary as well. The Cox process structure facilitates the analysis by splitting the likelihood function for default events into a component describing the evolution of the covariates and a component describing the occurrence of jumps conditional on the covariates.

An important use of the Cox setting for modeling defaultable securities proceeds by treating the intensity process as an unobserved latent variable, which drives the price of a corporate bond, or a credit default swap, which is a contract insuring against the default of a particular bond issuer. In this case, the estimation of the underlying intensity process proceeds as a filtering exercise using the frequently observed price data as measurements. The driving intensity is then estimated using the Kalman filter (see **Nonlife Loss Reserving**); see Duffee [9] or using Markov chain Monte Carlo (MCMC) methods (see **Reliability Demonstration; Bayesian Statistics in Quantitative Risk Assessment**). This mainly works for corporate bonds with a fairly liquid market. For

valuing less liquid bonds or loans for which no market exists, one often has to rely on option-theoretic models or try to map the characteristics (such as industry and accounting ratios) of the issuer onto issuers of more liquidly traded bonds.

Recently, the market for credit derivatives has seen a rapid growth in so-called collateralized debt obligations (CDOs), which introduce products whose prices depend strongly on the correlation between default events of different issuers. A Cox process approach to modeling prices of CDOs can be found in Duffie and Gârleanu [10], where the default intensities of individual issuers depend on a “factor” stochastic intensity process affecting all issuers and idiosyncratic intensities that are specific to each issuer and conditionally independent given the factor intensity.

References

- [1] Daley, D. & Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*, Springer, New York.
- [2] Snyder, D. (1975). *Random Point Processes*, John Wiley & Sons, New York.
- [3] Brémaud, P. (1981). *Point Processes and Queues-Martingale Dynamics*, Springer, New York.
- [4] Rudemo, M. (1973). Point processes generated by transitions of Markov chains, *Advances in Applied Probability* **5**, 262–286.
- [5] Resnick, S.I. (1992). *Adventures in Stochastic Processes*, Birkhäuser, Boston.
- [6] Taibleson, M. (1974). Distinguishing between contagion, heterogeneity and randomness in stochastic models, *American Sociological Review* **39**, 877–880.
- [7] Lando, D. (1998). On Cox processes and credit risky securities, *Review of Derivatives Research* **2**, 99–120.
- [8] Duffie, D., Pan, J. & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions, *Econometrica* **68**, 1343–1376.
- [9] Duffee, G. (1999). Estimating the price of default risk, *Financial Studies* **12**, 197–226.
- [10] Duffie, D. & Gârleanu, N. (2001). Risk and valuation of collateralized debt obligations, *Financial Analysts Journal* **57**, 41–59.

DAVID LANDO

Intent-to-Treat Principle

In a randomized clinical trial (RCT) (*see* **Randomized Controlled Trials**), subjects or patients are randomized to one or more study treatment or intervention arms. Ideally, randomization should occur after the investigators have verified that the patients have the disease under study and that they have met the eligibility criteria for entry into the trial. All patients who are randomized into a clinical trial comprise what is commonly referred to as the “intent-to-treat” population. In other words, all patients who are randomized are included in the analysis of the study results in the group to which they were assigned regardless of whether or not they received the assigned treatment [1–4]. The implications of this are several:

- If the patient is found not to have the disease under study, the patient remains in the analysis of results. This can occur, for example, when final verification of disease status is based on the results of special tests that are completed after randomization or on a central review of patient eligibility.
- If the patient never receives a single dose of the study drug, the patient remains in the analysis of results.
- If the patient does not comply with the treatment regimen or does not complete the course of treatment, the patient remains in the analysis of results.
- If the patient withdraws from the study for any reason, the patient remains in the analysis of results.

The intent-to-treat analysis based on the randomized treatment assignments allows us to conduct an unbiased test of the null hypothesis of no treatment difference (although an estimate of treatment effect may still be biased) and to attribute an observed difference to the treatment with a causal link [5]. The major issue for analysis, however, is how to handle those patients who did not receive the randomized treatment as planned owing to lack of compliance, side effects of the treatment, or other, possibly treatment related, reasons (*see* **Compliance with Treatment Allocation**). The proposed alternatives to intent-to-treat analysis include as-treated and

per-protocol analyses [1, 3]. Such analyses will result in distortions of the type I error rate as well as potentially biased estimates of the treatment effect [4, 6].

The interpretation of the results of any trial depends on the conduct of the trial as well as the analysis. The risks of reaching incorrect conclusions regarding the efficacy of a new treatment are increased with increased deviations from the planned design and randomization. For example, if there are differences in the eligibility rates, compliance, losses to follow-up, or withdrawal rates, which may be related to the treatment received, the overall estimated difference between the two arms may differ from the actual difference in efficacy. By using the intent-to-treat analysis, we estimate the effectiveness of the planned treatment strategies rather than efficacy directly. For example, if 90% of patients who are randomized to the new treatment refuse to take the treatment, even if the 10% of patients who do take the treatment have a 100% success rate (efficacy), the actual observed rate for the intent-to-treat analysis is 10%, which may be lower than the standard treatment rate.

In the case of a superiority trial, that is, an RCT designed to demonstrate that one treatment is superior to one or more other treatments, the intent-to-treat analysis is generally conservative. All patients who do not satisfy the conditions of the study at randomization or during the study are effectively treatment failures. If such patients are distributed equally among the several treatment arms, then any differences among the arms are generally reduced, and the effective study sample size is reduced. In such cases, if the results of the trial still indicate superiority, then while the estimate of effectiveness may be reduced, the overall conclusion would hold. If, however, any of these effects are differential between the treatment groups, the differences may be inflated in the intent-to-treat analysis [7].

In the case of noninferiority and equivalence trials (*see* **Inferiority and Superiority Trials**), the intent-to-treat analysis can bias the results in the direction of noninferiority or equivalence since the reduction in effective sample size produced by the inclusion of patients who were ineligible, noncompliant, etc., decreases the difference between the treatment groups.

When the intent-to-treat and the per-protocol or as-treated populations do not differ in any major ways, then the risks of incorrectly attributing efficacy (or

lack of efficacy) to a new treatment are reduced. Thus, it is critical that investigators design a study so that the eligibility criteria are clear, the treatment plan is well defined, and compliance issues are addressed.

Alternative approaches to statistical analysis have been proposed. These include longitudinal methods to handle dropouts [8], Bayesian methods to incorporate additional treatment information (such as rescue medications for treatment failures) using data augmentation algorithms [9], selection models that allow formal consideration of potential outcomes and pattern mixture models that model associations between observed exposures and outcomes [10], and the use of causal effect models for realistic treatment rules when the expected treatment assignment (ETA) is violated [11]. Hybrid intent-to-treat/per-protocol analyses that exclude noncompliant patients and incorporate the impact of missing data have been proposed for non-inferiority studies [12].

Supportive analyses always include analyses of patients who met the eligibility criteria; patients who met the eligibility criteria and received at least one dose of study medication; patients who met the eligibility criteria and completed the course of treatment (known as *per-protocol* population); among others. If all of the secondary and supportive analyses yield results that suggest the same conclusions as the results of the intent-to-treat analyses, then the trial is consistent. On the other hand, if the results are inconsistent in the various secondary and supportive analyses for the same endpoint, then it is the responsibility of the investigator to identify the reasons for the apparent differences, and the conduct of the trial may be an issue.

We note that any analyses of safety should be conducted on an as-treated basis.

Summary

In general, the intent-to-treat approach to the statistical analysis of RCTs is preferred to alternatives. However, investigators must provide a careful accounting of all patients randomized, all patients treated, and all postrandomization exclusions for lack of efficacy, lack of compliance, or lack of safety [4, 13, 14]. The intent-to-treat model provides a paradigm for the conduct of randomized trials that focuses on reducing any biases in patient assignment or evaluation of

outcome. However, the realities of clinical trial conduct can sometimes necessitate consideration of the deviations from the ideal model in the analysis.

References

- [1] Piantadosi, S. (1997). *Clinical Trials: A Methodologic Perspective*, Wiley-Interscience, New York.
- [2] Friedman, L.M., Furberg, C.D. & DeMets, D.L. (1998). *Fundamentals of Clinical Trials*, 3rd Edition, Springer-Verlag, New York.
- [3] Ellenberg, J. (1996). Intent-to-treat analysis versus as-treated analysis, *Drug Information Journal* **30**, 535–544.
- [4] DeMets, D.L. (2004). Statistical issues in interpreting clinical trials, *Journal of Internal Medicine* **255**, 529–537.
- [5] Harrington, D.B. (2000). The randomized clinical trial, *Journal of the American Statistical Association* **95**, 312–315.
- [6] Lee, Y.J., Ellenberg, J.H., Hirtz, D.G. & Nelson, K.E. (2001). Analysis of clinical trials by treatment actually received: is it really an option? *Statistics in Medicine* **10**, 1595–1605.
- [7] Goldberg, J.D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table, *Journal of the American Statistical Association* **70**, 561–567.
- [8] Hogan, J.W., Roy, J. & Korkontzelou, C. (2004). Tutorial in biostatistics: handling drop-out in longitudinal studies, *Statistics in Medicine* **23**, 1455–1497.
- [9] Shaffer, M. & Chinchilli, V. (2004). Bayesian inferences for randomized clinical trials with treatment failures, *Statistics in Medicine* **23**, 1215–1228.
- [10] Goetghebeur, E. & Loeys, T. (2002). Beyond intention to treat, *Epidemiologic Reviews* **24**, 85–90.
- [11] Van der Laan, M.J. & Petersen, M.L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules, *The International Journal of Biostatistics* **3**(1), Article 3 at <http://www.bepress.com/ijb/vol3/iss1/3>.
- [12] Sanchez, M.M. & Chen, X. (2006). Choosing the analysis population in non-inferiority studies: per-protocol or intent-to-treat, *Statistics in Medicine* **25**, 1169–1181.
- [13] Begg, C.B. (2000). Commentary: ruminations on the intent-to-treat principle, *Controlled Clinical Trials* **21**, 241–243.
- [14] Lachin, J.M. (2000). Statistical considerations in the intent-to-treat principle, *Controlled Clinical Trials* **21**, 167–189.

Related Articles

Causality/Causation

Repeated Measures Analyses

JUDITH D. GOLDBERG AND ILANA
BELITSKAYA-LEVY

Interpretations of Probability

Explicit and Tacit Probabilities

Except for a few scholars interested in mathematical esoterica, the familiar rules of probability, axiomized by Kolmogorov [1], are well established and generally agreed upon by researchers and practitioners. Despite this nearly universal agreement on its laws, people differ on the source and meaning of probability. Three dominant schools of thought are frequency probability, subjective probability (*see Subjective Probability*), and logical probability. All three traditions have a long history in mathematics and philosophy, and many pages have been written in support of and in refutation of these schools. What makes these continuing debates remarkable is that the quantification of uncertainty has been a tremendous achievement and benefit to society [2]. It appears in many diverse areas and forms the basis of empirical science, of legal evidence (*see Epidemiology as Legal Evidence*), of industries such as insurance and financial investments, and of artificial intelligence, to name just a few. Yet, we are none the closer to a consensus opinion about the meaning of “probability”. These differing interpretations of probability are more than moot academic arguments and can affect how insurance and risk professional approach problems. We often do not realize that we are taking sides in the debate because the interpretations are encoded in statistical software and standard methods.

Many practitioners, authors, and teachers, who are very adept with probability calculus, find it difficult to give a succinct interpretation without recourse to informal references to chance and uncertainty. Community standards may play a more important role in the interpretation of probability than, perhaps, mathematicians would care to admit. We develop our probabilistic intuitions through reference to various enumeration exercises and chance mechanisms. These insights do not come naturally, but require repetition and guidance by experts. The unnaturalness of probability and the need to transmit its meaning through formal education and socialization may be one reason that its development as a quantitative discipline traces only to the seventeenth century correspondence between Fermat and Pascal, while

gambling devices are found in prehistoric settlements, and qualitative expressions of chance and uncertainty appear in the earliest myths [3–5]. Using Polanyi’s [6] taxonomy of knowledge, the laws of probability are explicit knowledge: they are written down, accessible to everyone, and exist independently of the observer, while the interpretations of probability have a stronger resemblance to tacit knowledge, which is developed with experience, resides within individuals, and is transmitted through socialization. This tacit property may be the reason for several interpretations.

Relative Frequency

Most students start their probability and statistics training with probability being the same as relative frequency or proportion. A typical introduction will have illustrations similar to the following: “According to the Centers for Disease Control (CDC) (<http://www.cdc.gov>), there were 38 553 new diagnoses of HIV/AIDS in 2004 in 35 areas of United States with confidential name-based reporting, of which 10 510 were female. Consequently, the probability that a new diagnosis is female is 27%.” This approach provides a useful pedagogical device for illustrating the laws of probability to a general audience: all of the laws can be demonstrated with various cross tabulations. Not only are new HIV/AIDS diagnoses men and women, but they also can be classified by means of transmission. Then one could derive the probability of transmission through heterosexual contact to be 78% for women and 16% for men, and so on.

Relative frequency is an easy device to motivate the calculus of probability, but should proportions be evaluated to the stature of probability? To do so, teachers and authors add the tag line, “randomly selected”, as in “The probability that a randomly selected new HIV/AIDS diagnosis is a woman is 27%.” Sometimes, a student will ask, “Either the selected person is a man or women; what does 27% mean?” Teachers frequently invoke repeated sampling: if the sampling was repeated a large number of times (with replacement of course), then 27% of all resulting samples of size one will consist of a woman. Almost always, “randomly selected” is left to students’ imagination. Later, students learn that random number generators can be used for random selection. When asked

about the origins of random numbers, they are informed that computer programmers have conveniently worked them out. On deeper investigation, they discover that these sequences of numbers are deterministic and can be perfectly predicted if one knows the algorithm and its initialization: the correct phase diagram morphs these chaotic digits into patterns [7].

This vignette demonstrates the problem of revealing the meaning of probability. Relative proportions – 27% of new HIV/AIDS diagnoses are women – become probabilities – 27% chance of randomly selecting a woman – through a randomization mechanism. The randomization mechanism is not at all random, so probability becomes operationally defined by means of a deterministic process. Of course, one could roll a “38 553 sided die” or use some other randomization device, but then one needs to explain the meaning of probability for the “38 553 sided die”.

Not all is lost. One attempt to salvage the situation is the argument that “random selection” is an idealization that can only be approximated in reality. Calculations based on pseudorandom digits conform, to a high degree, to those derived from a random sample from a uniform distribution. The relative frequency that elements of a long series of digits belong to a subinterval converges in probability to the length of the subinterval; the average and variance of the digits will be close to $\frac{1}{2}$ and $\frac{1}{12}$; the autocorrelations of nonzero lag will be nearly zero; and so on. However, properties of the two do not perfectly match: the pseudorandom numbers will eventually repeat.

Every so often, a student will make the embarrassing observation that the original “27% probability of randomly selecting a woman” was an attempt to define probability, and the teacher merely transferred the definition of probability to an even more remote source – an idealized randomization device. The pseudorandom number generator is then certified by recourse to empirical evidence based on ill-defined inferential procedures that are inherently probabilistic: round and round the roulette wheel of circular interpretation spins.

A different attack is to use symmetry, the lynchpin of classical probability [8]. Though it is true that the random numbers are not actually random, there is nothing in their use that distinguishes one HIV/AIDS case over the other. Therefore, the random selection

process is indifferent to the case actually selected, and all outcomes are equally likely. Thus, the probability of randomly selecting a women patient *must* be evaluated at 27%. Symmetry arguments also apply to other phenomenon. Without further information about the physical process that flips a coin or rolls a die, a sensible observer would view all outcomes as being equally likely; hence one-half for a head and one-sixth for each side of a die.

The “must” injunction of symmetry arguments turns out to be more nuanced than it first appears. “Probability of randomly selecting a woman is 27%” may be a fair assessment if the person is selected with pseudorandom numbers. What if the randomization mechanism selects patients relative to their height? Then symmetry arguments would imply that the probability of a new HIV/AIDS diagnosis being a woman is less than 27% because the male distribution for height stochastically dominates the female distribution.

Subjective Probability

In using symmetry, our assignment of probabilities change depending on the assumptions we make about the population, sampling mechanism, or experiment. In other words, our probability assessments depend on our state of knowledge about the phenomenon. This marks a subtle and important shift in the interpretation of probability: the subjective one (*see Subjective Probability*). In the frequency interpretation, probability is an innate characteristic. Flipped coins come up heads 50% of the time because that is the property of coins. In subjective probability, flipped coins come up heads 50% of the time because the observer does not have information that would lead him or her to believe that one side is preferred over the other. The locus of uncertainty is not the coin and its environment, but resides within the observer. If the observer knew more about the physical laws of coin flips and the initial conditions, then he or she could refine his or her probability assessment.

The subjective interpretation provides advantages and challenges. By reporting internal states of uncertainty with probability, it frees practitioners to consider phenomenon that do not fit the repeated sampling framework. Underwriting insurance policies for large risk groups is well informed by relative frequencies, while they are less useful in certifying the safety of a new aircraft design. The relative

frequency approach would use accident rates of previous aircraft designs as the probability for the new design. A subjectivist could combine this historical data with engineering knowledge about the design and test results to arrive at an assessment of the design's safety. Even in setting a premium for an individual's life insurance, obtaining more information, including genetic data, about a policyholder makes it harder to identify the relevant reference population on which to base the life table. Structural breaks, such as the introduction of HIV/AIDS to the United States in 1981 and the subsequent treatment regimens and behavioral modification, can invalidate historical relative frequencies. In contrast, a subjectivist can freely combine historical rates with other information to arrive at an updated probability evaluation.

In addition to the flexibility of subjective probability, it neatly separates the interpretation of probability from inference while at the same time providing a unified method for inference and prediction through Bayes' theorem (*see* **Bayes' Theorem and Updating of Belief**). The frequentist reports the probability of obtaining a heads as $1/2$ if 50% of the tosses in a long series resulted in heads. This definition entwines the definition of probability with inference for that probability. A subjectivist reports the probability of heads as a distribution for possible values to reflect his or her knowledge, or lack thereof, about the phenomenon. If the subjectivist reports the probability θ of a head as a constant, such as $1/2$, then he or she is being very dogmatic in his or her beliefs and does not allow the possibility to learn from experimentation. If the subjectivist were uncertain about the exact value of θ , then he or she would give a prior distribution for θ . As outcomes of the flips are observed, Bayes' theorem is used to update the distribution for θ . The subjectivists incorporate relative frequency information from observations into a formal learning process that revises his or her beliefs about the probability of obtaining a head.

Making a Science out of Personal Beliefs

Challenges for subjectivists are twofold. First, why use probability for expressing beliefs? There is a large gap between the qualitative expressions, such as, "It is fairly likely that the new aircraft design is safe," to quantification: "The probability that the new design

will have an accident due to design flaws is 0.001%." An even larger gap exists between a quantification of one event to mandating that subjective beliefs for any and all collections of events *must* conform to the laws of probability. Second, how should one make a science out of a concept that is dependent on an observers' internal state?

Decision Theory

von Neumann and Morgenstern's [9] economic theory of rational choice provides a mechanism to show that a person's knowledge about uncertain events should be stated as probabilities by deriving them from the person's preferences. For example, the firm that designs a new airplane could choose either to manufacture and sell the new airplane or to cancel the program. If the firm sells an unsafe airplane, then it will incur a large financial and reputation loss, not to mention the human suffering it could cause. If it cancels a safe design, then the firm has an opportunity costs. Embedded within the decision lurks the firm's assessment of the safety of the design and the attractiveness of various gains and losses.

The primary object of the theory of choice is the preferences for actions available to the decision maker. If these actions (build the new design or cancel the project) satisfy a number of axioms, called the *axioms of rationality*, then a mathematician can derive a utility function on the space of consequences (financial gains and losses, opportunity costs, death, pain, and suffering) and a probability function on the space of events (safe or unsafe design) such that the ordering of actions based on preferences is the same as the ordering of actions based on expected utility (*see* **Subjective Expected Utility; Utility Function**). A mathematician can transform purely qualitative information about choices to numerical assessments of both utility and probability. The theory does not assume that people have probability distributions and utility functions and choose according to expected utility, but if their choices conform to the axioms or rationality, then their behavior is consistent with expected utility theory for some utility function and probability distribution. The converse is also true: if one orders actions according to expected utility, how every one arrives at utility and probability, one is guaranteed to be "rational". Thus, expected utility is the basis of decision theory (*see* **Decision Analysis; Decision Modeling**), and a person's belief

about uncertain events should be stated as probabilities.

von Neumann and Morgenstern assumed that there exists an unspecified, randomization device by which the mathematician is able to calibrate the subject's subjective probability for events. Their system is not useful as a definition of probability because it requires (frequency?) probability to define subjective probability. Savage [10] extended von Neumann and Morgenstern's axioms so that subjective probability can be derived internally to the subject's preferences, without reference to an external randomization device. The von Neumann–Morgenstern–Savage (NMS) model for rational choice not only requires that opinions about events conform to the laws of probability, but also provides an operational method to elicit those probabilities, along with utility functions, from observed preferences.

The NMS model depends on the context of the decision problem. Returning to the new HIV/AIDS cases, it may not be meaningful to ask for the probability of selecting a woman because the problem is devoid of a decision context: there are neither actions nor consequences. One way to provide a context is to give subjects a gamble, such as, "Suppose you will win \$1 if a randomly selected new HIV/AIDS case is a woman. What is the most that you be willing to play the game?" The modal answer is often well below 27 cents, which partly reflects risk aversion (see **Risk Attitude**), but more than likely for such small sums, merely reflects that the scenario is not designed to elicit the true or fair amount for playing the game.

Gambles and Coherence

Ramsey [11] and de Finetti [12] devised gambling situations that are less elaborate than the NMS axioms of rationality though special cases of rationality. To elicit honest responses, the subject is put into the scenario of a bookmaker (bookie) that states odds on events and must accept any sequence of bets, both positive or negative. The game goes as follows. The bookie (our subject) posts odds $P(A)$, $P(B)$, ... on the events $A, B, \dots$. A gambler, the bookie's customer, puts a stake S_A , which can be negative, on event A . It costs the gambler $S_A P(A)$ to play the game. If A occurs, the bookie pays the customer S_A , and the customer's total winnings are $S_A[1 - P(A)]$.

If A does not occur, the bookie pays the customer nothing, and the customer's total winnings are $-S_A P(A)$. Because the stakes on any event can be positive or negative, the bookie must judiciously select P to avoid becoming a sure loser regardless of the outcome. Being a sure loser is sometimes called *Dutch book*, and modern finance calls it *arbitrage*, which plays a central role in financial theory. de Finetti termed the specification of P that does not lead to arbitrage or Dutch book as being "coherent" or "coherently" stating one's beliefs. He shows that P is coherent if and only if it follows the laws of probability.

de Finetti's coherence theorem is very compelling for a bookie stating odds or a market maker quoting buy and sell prices. But not everyone faced with uncertainty wants to enter a casino, in which case they are free to quantify their beliefs however they choose. However, those of us who do not always conform to the strict dictates of rationality or coherence should be aware of the consequences of our departures, which, perhaps, points toward the role of education and socialization in understanding probability.

Social Determinates of Science

The second major obstacle for subjective probability is how to make a science from a concept that is focused on an individual's belief. An individual's belief about a gamble is a personal decision that only affects the gambler. What of an engineer's belief about the safety of a new aircraft design; an actuary's computation of premiums; or a medical researcher's belief about the efficacy of a new drug? The bookie and gambler paradigm may no longer apply because more is at stake than self-interest, and the consequence of these personal beliefs extend beyond the individual who holds them (see **Group Decision**). At the very least, will people examining the same data arrive at the same probability judgments? There is nothing in the theory of subjective probability that compels them to do so.

This nonuniqueness is a long-standing reason that frequentists reject subjective probability as a fruitful approach to science. The subjectivists' rejoinder is that objectivity is a chimera and that the best that one can do is to state underlying assumptions as accurately as possible and to present clearly the

evidence for or against a theory, so that other scientists can evaluate the results and update their probability assessments. Subjective probability fits better with science as a social endeavor where researchers marshal evidence to persuade colleagues about the suitability of a theory, which is then modified or discarded, than as a solitary pursuit of objectively knowable truth. Nevertheless, the frequentists' critique is true in an absolute sense: subjectivists believe in individual liberty to choose the probabilities that best fits their beliefs.

Inference and Prediction

These debates are far from inconsequential and affect how one approaches inference. Consider the model where $X_1, \dots, X_n$ is a random sample from a distribution with density f given an unknown parameter θ . The information in the data about θ is the likelihood function: $L(\theta) = \prod_{i=1}^n f(x_i|\theta)$. Members of both schools agree that the likelihood function summarizes the information in the data about θ . A subjective Bayesian adds his or her personal information about θ in the form of a prior distribution with density $p(\theta)$. The Bayesian then updates the prior opinions after observing the data $\underline{x}_0$ to obtain the posterior distribution with density $p(\theta|\underline{x}_0) = L(\theta)p(\theta)f(\underline{x}_0)$ where $f(\underline{x}_0) = \int L(\theta)p(\theta)d\theta$ (see **Bayesian Statistics in Quantitative Risk Assessment**).

Repeated Sampling versus Bayesian Learning

Frequentists make probability statements by integrating functionals of the data over possible values of X or the sample space, given the parameters. For example, for the parameter θ , Neyman's [13] 95% confidence interval has the form $LB(\underline{x})$ to $UB(\underline{x})$ based on the sample random sample $\underline{x} = (x_1, \dots, x_n)$. The bounds are constructed so that 95% of all possible samples of $\underline{x}$ will result in intervals such that the interval covers the population parameter: $P[LB(\underline{x}) < \theta < UB(\underline{x})|\theta] = 0.95$, which conditions on θ and treats the data as random. The inductive leap is that for the observed sample $\underline{x}_0$, we have 95% "confidence" that $LB(\underline{x}_0)$ to $UB(\underline{x}_0)$ contains θ . All Bayesian inferential statements are based on the posterior distribution $p(\theta|\underline{x}_0)$, which conditions on the sample information $\underline{x}_0$ and treats θ as random. A 95% highest posterior credible interval $LB(\underline{x}_0)$ to

$UB(\underline{x}_0)$ for θ is an interval such that given the data, the posterior probability that θ is in the interval is 95%: $P[LB(\underline{x}_0) < \theta < UB(\underline{x}_0)|\underline{x}_0] = 0.95$.

Neyman's repeated sampling framework created a paradigm shift in the scientific method and is still the dominant mode of inference. It provides an answer to the question of how sensitive the conclusions of a study are to random sampling. Inconsequential theories receive scant scrutiny, while success invites criticism; the subjectivists' critiques of the frequentists' inference is a case in point. Lindley [14] maintains that confidence intervals are a probability statement about the interval $LB(\underline{x})$ to $UB(\underline{x})$, given the parameter θ , while the proper focus ought to be the probability that θ is in the observed interval $LB(\underline{x}_0)$ to $UB(\underline{x}_0)$, given the data. Similarly, significance levels in hypothesis-testing report the probability of some aspect of the data, given the hypothesis, while science requires the probability of the hypothesis, given the data. This role reversal of random and fixed quantities in frequentist inference can lead to a number of anomalies owing to incoherency of not using a proper probability specification.

Exchangeability, Prediction, and Invariance

de Finetti's [12] exchangeability theorems provide a representation for subjective probability where the prior and likelihood function are jointly determined from a person's beliefs about observables. (See [15] for references that pre-date de Finetti.) Exchangeability means that the order of observations does not matter in probability assessments: $f(x_1, \dots, x_n) = f(x_{\pi_1}, \dots, x_{\pi_n})$ for any permutation $(\pi_1, \dots, \pi_n)$ in the order of the observations. All random samples are exchangeable, but not all exchangeable sequences are random samples. For infinitely exchangeable sequences the joint distribution of $X_1, \dots, X_n$ can be decomposed into a prior and likelihood: $f(x_1, \dots, x_n) = \int \prod_{i=1}^n f(x_i|\theta)p(\theta)d\theta$. One direction is trivial. If the observations are a random sample from $f(\bullet|\theta)$ given θ and if the prior for θ is $p(\theta)$, then the observations are exchangeable with marginal density $f(x_1, \dots, x_n)$. The other direction is of great importance. de Finetti argues that people have beliefs $f(x_1, \dots, x_n)$ about observables, not beliefs about

unobserved parameters. A mathematician can decompose an individual's specification for the joint distribution into a likelihood and a prior distribution, which is analogous to the theory of rational choice that deduces probability and utility from preferences.

Prediction becomes a simple task using the exchangeability representation. Given the data $X_1, \dots, X_n$, the predictive distribution for the next m observations, $X_{n+1}, \dots, X_{n+m}$ is:

$$\begin{aligned} f(x_{n+1}, \dots, x_{n+m} | x_1, \dots, x_n) \\ &= \frac{f(x_1, \dots, x_{n+m})}{f(x_1, \dots, x_n)} = \frac{\int \prod_{i=1}^{n+m} f(x_i | \theta) p(\theta) d\theta}{\int \prod_{i=1}^n f(x_i | \theta) p(\theta) d\theta} \\ &= \int \prod_{j=1}^m f(x_{n+j} | \theta) p(\theta | x_1, \dots, x_n) d\theta \quad (1) \end{aligned}$$

The predictive distribution depends on the data through the posterior distribution, which summarizes both the sample and prior information about θ . The prior, likelihood, and posterior are mathematically convenient devices to obtain the predictive distribution for future observations. In the HIV/AIDS example, the 10 510 female cases out of 38 553 new diagnoses are of interest only to the extent that they inform us about the predictive distribution of future cases, which de Finetti's bookies will use in posting betting odds.

Invariance representation theorems extend de Finetti's exchangeability representation theorems by identifying the form of the likelihood. For example, Freedman [16] and Kingman [17] show that spherical symmetry results in scale mixtures of independent normal distributions. The mixing distribution for the scale parameter could be viewed as a prior distribution, while the normal densities could be considered the likelihood function. Invariance theorems convert qualitative information about outcomes into formal, statistical models where "must" must hold. In subjective probability, the individual is free to assume or to reject invariance; however, if he or she accepts it, then the invariance theorems identify the appropriate families of distributions.

Logical Probability

The third interpretation of probability is logical or epistemic, proposed by Keynes [18] and elaborated by Carnap [19]. The logical probability of a hypothesis H given evidence E arises from the semantic relations among the statements that describe both hypothesis and evidence. In the sampling of women HIV/AIDS cases, we would assign a probability to the statement, $H =$ "There is a 27% probability of randomly selecting a woman", based on the strength of evidence contained in the statement: $E_1 =$ "10 510 women out of 39 553 new cases". Since we are not told about the sampling mechanism, we may assign a wishy-washy $P(H|E_1) = 0.5$, where $P(H|E_1)$ is the objective degree of logical support for H provided by E_1 . If we are informed about random selection *via* $E_2 =$ "Using a random number generator to select the case", then $P(H|E_1 \text{ and } E_2)$ is (nearly) one. On the other hand, if random selection is $E_3 =$ "Select the case proportion to the subject's height", and we know $E_4 =$ "The distribution of women heights are stochastically dominated by the distribution of men heights", then $P(H|E_1 \text{ and } E_3 \text{ and } E_4)$ drops to around zero.

Logical probability requires the specification of the joint probability of all relevant statements under consideration. Once one has this specification, it is possible to use the calculus of probability to deduce the degree of confirmation for any hypothesis based on any evidential statements. Specifying the joint distribution for logical probabilities is similar to specifying a rational preference structure for all possible actions in the theory of choice or the joint distribution for exchangeable sequences: all three are exceedingly difficult to do in practice.

Logical probability, as posit by Keynes, fell into disfavor, in part, because this joint probability for a collection of statements should be objectively evident to all who discern them, sometimes termed "necessary" probability. If one drops the "must" part of the probability assignments, most subjectivists do not have a problem with logical probability. In fact, Franklin [20] credibly argues that frequency and subjective probability are special instances of logical probability.

Logical probability has had negligible impact on empirical sciences; unlike frequentist or Bayesian inference, it has not developed a collection of models and methods for inference, hypothesis testing,

and forecasting. Nevertheless, the core ideas form the basis of artificial intelligence using Bayes nets, introduced by Pearl [21] (*see Expert Judgment; Influence Diagrams*). The nodes of a Bayes net are statements, and the arcs are probability assessments. Bayes nets express the semantic relations among statements, along with probability assessments. For instance, the nodes of a Bayes net for aircraft safety could include engineering specifications, flight conditions, and failure modalities. Bayes nets use the laws of probability to compute the probability of hypotheses, given the evidence. Many of us have unwittingly used Bayes nets; Microsoft Office's automatic help function is based on a Bayes net.

A Multifaceted Future

All human conflicts eventually come to an end, if not because there is a clear victor, then due to exhaustion. Academics in early and middle part of the twentieth century held strong beliefs about the interpretation of probability, and careers were made or lost depending on one's allegiances. Statisticians and researchers in the twenty-first century are more sanguine and flexible. Bayarri and Berger [22] note that the distinctions between the schools are less obvious, at least in practice. Researchers freely adopt the interpretation that best fits their needs, and few are interested in rehashing the battles of the twentieth century. Knowledge engineers who design Bayesian nets do not make a distinction between frequency, subjective, or logical probabilities: they use any variety that is available and reliable. Researchers and risk professionals readily flip between frequency and Bayesian methods. Hybrid Bayesian use subjective priors and frequentist likelihoods – likelihoods determined from the observed variation of the data and not based on prior beliefs. Computational Bayesian employ frequentist concepts when analyzing the output of Markov chain Monte Carlo simulations, and frequentist often treat p values as the probability of the null hypothesis and interpret confidence intervals as subjectivists. No longer are priors reflexively shunned in scientific journals; they are critiqued as one aspect of the research design. The advancement of science requires the intelligent and sensible construction of models, appropriate use of methodology and measurement, and trust that researchers accurately portray their findings. The dogmatic adherence to schools of interpretation

will not compensate for ill-conceived measurements, poorly designed and executed experiments, inadequate models, or unethical researchers. The interpretation of probability will continue to evolve over time, and education and social processes, enabled by user-friendly software, are bound to blur further the distinctions among the different interpretations of probability.

Acknowledgment

I wish to thank Bruce Hill for his insight and knowledge about subjective probability, and Simon French and Michel Wedel for commenting on earlier drafts.

References

- [1] Kolmogorov, A.N. (1933). *Grundbegriffe der Wahrscheinlichkeitrechnung*, Ergebnisse Der Mathematik; (translated as) *Foundations of Probability*, Chelsea Publishing Company, 1950.
- [2] Bernstein, P.L. (1996). *Against the Gods: The Remarkable Story of Risk*, John Wiley & Sons, New York.
- [3] David, F.M. (1969). *Games, Gods and Gambling; the Origins and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era*, Hafner, New York.
- [4] Hald, A. (1990). *A History of Probability and Statistics and their Applications before 1750*, John Wiley & Sons, New York.
- [5] Stigler, S.K. (1986). *The History of Statistics: the Measurement of Uncertainty before 1900*, Belknap Press of Harvard University Press, Cambridge.
- [6] Polanyi, M. (1966). *The Tacit Dimension*, Doubleday, Garden City.
- [7] Ripley, B.D. (1987). *Stochastic Simulation*, John Wiley & Sons, New York.
- [8] Laplace, P.S. (1951). *A Philosophical Essay on Probabilities*, Dover Publications, New York; originally published in 1814.
- [9] von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*, Princeton University Press, Princeton; John Wiley & Sons, New York.
- [10] Savage, L.J. (1954). *The Foundations of Statistics*, Republished in 1972 by Dover, New York.
- [11] Ramsey F.P. (1931). Truth and probability, in *Foundations of Mathematics and Other Essays*, R.B. Braithwaite, ed, Routledge & P. Kegan, 156–198; reprinted in *Studies in Subjective Probability*, 2nd Edition, Robert E. Krieger Publishing Company, 1980, pp. 23–52; reprinted in *Philosophical Papers*, D.H. Mellor, ed, Cambridge University Press, Cambridge, 1990.
- [12] de Finetti, B. (1937). La Prévision: ses lois logiques, ses sources subjectives, *Annales De L Institut Henri*

- Poincare 7, 1–68; translated as (1980). Foresight. its logical laws, its subjective sources, in *Studies in Subjective Probability*, H.E. Kyburg Jr & H.E. Smokler, eds, Robert E. Krieger Publishing Company.
- [13] Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability, *Philosophical Transactions of the Royal Society of London, Series A* **236**, 333–380.
- [14] Lindley, D.V. (2000). The philosophy of statistics, *The Statistician* **49**(3), 293–337.
- [15] Hewitt, E. & Savage, L.J. (1955). Symmetric measures on Cartesian products, *Transactions of the American Mathematical Society* **80**, 470–501.
- [16] Freedman, D.A. (1963). Invariants under mixing which generalize the de Finetti's theorem, *Annals of Mathematical Statistics* **34**, 1194–1216.
- [17] Kingman, J.F.C. (1972). On random sequences with spherical symmetry, *Biometrika* **59**, 492–494.
- [18] Keynes, J.M. (1921). *Treatise on Probability*, Macmillan Publishers, London.
- [19] Carnap, R. (1950). *Logical Foundations of Probability*, University of Chicago Press, Chicago.
- [20] Franklin, J. (2001). Resurrecting logical probability, *Erkenntnis* **55**, 277–305.
- [21] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann.
- [22] Bayarri, M.J. & Berger, J.O. (2004). The interplay of Bayesian and frequentist analysis, *Statistical Science* **19**, 58–80.
- [23] von Mises, R. (1957). *Probability, Statistics and Truth*, revised English edition, Macmillan, New York.
- [24] Feller, W. (1957). *An Introduction to Probability Theory and Its Application*, John Wiley & Sons, New York, Vol. 1.
- [25] Feller, W. (1966). *An Introduction to Probability Theory and Its Application*, John Wiley & Sons, New York, Vol. 2.
- [26] Breiman, L. (1968). *Probability*, Addison-Wesley, Reading.
- [27] Cox, D.R. & Hinkley, D.V. (1975). *Theoretical Statistics*, Chapman & Hall, London.
- [28] Bickel, P.J. & Doksum, K.A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco.
- [29] Jeffreys, H.J. (1961). *Theory of Probability*, Clarendon Press, Oxford.
- [30] de Finetti, B. (1974). *Theory of Probability*, translated be A. Machi & A. Smith, eds, John Wiley & Sons, Chichester, Vol. 1.
- [31] de Finetti, B. (1975). *Theory of Probability*, translated be A. Machi & A. Smith, eds, John Wiley & Sons, Chichester, Vol. 2.
- [32] Kyburg, H.E. & Smokler, H.E. (1980). *Studies in Subjective Probability*, 2nd Edition, Robert E. Krieger Publishing Company, Huntington.
- [33] Fishburn, P.C. (1986). The axioms of subjective probability, *Statistical Science* **1**(3), 335–345.
- [34] Box, G.E.P. & Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading.
- [35] DeGroot, M.H. (1970). *Optimal Statistical Decisions*, McGraw-Hill, New York.
- [36] Ferguson, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press, New York.
- [37] Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, John Wiley & Sons, New York.
- [38] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer-Verlag New York, New York.
- [39] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*, John Wiley & Sons, Chichester.
- [40] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman & Hall/CRC, New York.
- [41] Congdon, P. (2001). *Bayesian Statistical Modelling*, John Wiley & Sons, New York.
- [42] Cowell, R.G., Dawid, A.P., Lauritzen, S.L. & Spiegelhalter, D.J. (1999). *Probabilistic Networks and Expert Systems*, Springer-Verlag, New York.
- [43] Jensen, F.V. (2001). *Bayesian Networks and Decision Graphs*, Springer.
- [44] Hájek, A. Interpretations of probability, in *The Stanford Encyclopedia of Philosophy (Summer 2003 Edition)*, E.N. Zalta, ed, The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford, <http://www.plato.stanford.edu/archives/sum2003/entries/probability-interpret/>.
- [45] Fonseca, G.L. *Choice under Risk and Uncertainty*, Department of Economics, New School for Social Research, New York, <http://www.cepa.newschool.edu/het/essays/uncert/choicecont.htm>.
- [46] Lenk, P. (2001). *Bayesian Inference and Markov Chain Monte Carlo*, <http://webuser.bus.umich.edu/plenk/downloads.htm>.

Further Reading

There are many wonderful sources, too numerous to mention here, on probability and inference, in addition to those already cited. von Mises [23] attempts to establish a mathematical foundation for frequency probability with infinite sequences or “collectives”. Classical texts on mathematical probability are those by Feller [24, 25] and Breiman [26]. References [27, 28] are standard introductions to frequentist statistical inference. The books by Jeffreys [29] and de Finetti [30, 31] are landmark texts for subjective Bayesian probability. Kyburg *et al.* [32] provide a collection of historical, influential articles on subjective probability. Fishburn [33] provides a brief overview of the axioms of subjective probability. Early texts on Bayesian inference are by Box *et al.* [34], DeGroot [35], Ferguson [36], and Zellner [37]. For theoretically inclined readers, two of the finest, modern texts on Bayesian inference are by Berger [38] and Bernardo *et al.* [39], and the work of Gelman *et al.* [40] and Congdon [41] should appeal to applied researchers. Cowell

et al. [42] and Jensen [43] provide good introductions to Bayes nets.

The internet holds a wealth of material. In particular, Hájek [44] provides a scholarly account, and Fonseca [45] summarizes the axioms of rationality. Lenk [46] provides mathematical notes on Bayesian inference using Markov chain Monte Carlo. Free statistical software, mostly frequentist, but also

Bayesian, is available through <http://www.r-project.org/>; free Bayesian software can be obtained at <http://www.mrc-bsu.cam.ac.uk/bugs/>, where one can also join an online community of Bayesians.

PETER LENK

Kriging

Kriging is the name given to a collection of spatial interpolation methods that are derived from model-based optimality criteria and provides model-based estimates of interpolation error. Kriging, as is usually practiced, proceeds in two stages. In the first stage the available spatial data are used to fit a model of spatial correlation; in the second stage the fitted model is used to optimize the interpolator and to provide error estimates. The spatial variables may be continuous spatial properties such as temperature, multivariate properties such as composition, or discrete properties such as land use. The spatial domains are typically continuous two-dimensional or three-dimensional geographic regions, but could include domains on other spatial scales, such as rock thin sections or human cells and organ scans. Kriging has been generalized to space–time interpolation problems. The data used for modeling spatial dependence and for local interpolation control are commonly measurements made at discrete locations within the spatial domain of interest, such as the data provided by a survey, by monitoring networks, and by probes of various kinds.

In its simplest form, a kriging interpolator is an optimized linear combination of control data at observation locations. The coefficients of the linear kriging interpolator and the associated error measure both depend on the spatial configuration of the control data relative to the interpolation location, as well as the fitted model of spatial correlation.

Kriging is named after D. Krige whose early interest in issues of spatial scale in mining applications [1] provided the stimulus. The kriging method has close links to Wiener optimal linear filtering in the theory of random functions, Gandin objective analysis in meteorology [2], spatial splines, and generalized least-squares estimation in a spatial context. Kriging methods have been studied and applied extensively since 1970 and have since been adapted, extended, and generalized. For example, kriging has been generalized to classes of nonlinear functions of the observations, modified to increase robustness, extended to take advantage of covariate information, adapted for fields whose statistical properties are spatially evolving, and placed into a formal Bayesian framework (*see Bayesian Statistics in Quantitative Risk Assessment*).

Matheron [3] presented an influential early systematic exposition of kriging. Among book length accounts published in the decade or so since 1990 are, in order of date, Cressie [4], Kitanidis [5], Goovaerts [6], Wackernagel [7], Armstrong [8], Olea [9], Stein [10], and Chiles and Delfiner [11]. In addition, commercial and public domain software to implement kriging is widespread and includes components written for the public domain statistical programming language R.

Model-Based Interpolation Errors

Kriging interpolators are optimized linear combinations of the control data used for interpolation. Kriging interpolators minimize the “expected” squared interpolation error computed in the context of a specified probabilistic model for the spatial field of interest, which is here called the *a priori probability model*, and the expected squared interpolation error under the *a priori* model is called the *performance* of the interpolator. The optimization under the *a priori* model does not involve conditioning on the available data, and the performance measure treats data at observed locations as though they were still random variables.

The optimization of interpolator performance, subject to unbiasedness constraints described below, yields the kriging coefficients that are applied to the control data for interpolation. These optimized coefficients depend only on the *a priori* model’s covariances between values of the spatial field at location pairs. The optimized interpolating coefficients do not depend on the actual values of the data, except insofar as the data are used to estimate the *a priori* model’s pairwise covariances.

In general, linear kriging interpolators, i.e., those that are optimized for the *a priori* model, will on average have larger squared errors than nonlinear interpolators that minimize the *conditional* expected squared error, given the available data. The advantage of kriging interpolators is that the expected squared error depends on the model only through the *a priori* covariances between values at different spatial locations, thus requiring a rather minimal model specification. In contrast, conditional expected value interpolators depend on the *a priori* model in more complicated ways that require complete model specifications.

However, when the *a priori* model is a Gaussian spatial field then the expected squared interpolation error in the kriging context coincides with the expected squared error conditioned on the data. Furthermore, the kriging interpolator will coincide with the conditional expected value at the interpolation location given the data, in this special case.

Kriging interpolation procedures are optimized pointwise, i.e., interpolation performance is separately optimized at each interpolation location. No constraints are placed on the collective properties of interpolated values and the performance measure does not consider the collective properties of interpolated values. Indeed, pointwise kriging optimization results in an interpolated spatial field whose spatial properties are not consistent with the *a priori* spatial probability model. Interpolated fields have less spatial variability and show a stronger spatial continuity than what is implied by the *a priori* model. Therefore, it is inappropriate to use a map of interpolated values obtained by kriging to infer variability or continuity properties of the spatial field.

Notation and Data

$Z(x)$ denotes the property (value) of the spatial field at location x , where x is any point (site, location) in the geographic region of interest. For example, $Z(x)$ is the surface concentration of a pollutant at location x in a specified air basin. We may generalize the spatial property $Z(x)$ to be multivariate, such as a suite of multiple-pollutant concentrations and weather variables, defined at every site x in the air basin.

$Z_0 = Z(x_0)$ is the unknown value of the spatial field at an interpolation location x_0 .

$x_i, i = 1 \dots n$, denotes n discrete data locations used for interpolating at location x_0 . These are called *the control data locations*. Control data locations are usually chosen in a neighborhood of the interpolation location x_0 .

$\mathbf{Z}_n$ is the vector of control data, i.e., the n vector of observations $Z(x_1) \dots Z(x_n)$.

Observations are not always in the form of discrete point observations. Measuring devices may produce locally averaged values, typical of remote sensing data. Or data may be measured continuously along one-dimensional trajectories such as data from moving tracking devices. Kriging methods can be generalized to accommodate such data.

The probabilistic model treats the spatial field $Z(x)$ as a random spatial function. The control data $\mathbf{Z}_n$ and the unobserved values of the field Z_0 will have a joint probability distribution that is derived from the model specification. The notation for model quantities is:

$\mu(x)$ is the *a priori* model expected value of the random variable $Z(x)$ at a location x . $\boldsymbol{\mu}_n$ is the n vector, $\mu(x_i), i = 1 \dots n$, of expected values at the n control data locations. μ_0 is the expected value, $\mu(x_0)$, at the unobserved interpolation location x_0 .

$c(x, x')$ is the model covariance between the pair of random variables $Z(x)$ and $Z(x')$ at the location pair x and x' . $\mathbf{c}_n$ is the $n \times n$ matrix, $(c(x_i, x_j))$, of variances and covariances between pairs of random variables at data locations x_i and x_j . $\mathbf{c}_0$ is the n vector of model covariances, $(c(x_i, x_0))$, between the random variable $Z(x_0)$ at the interpolation location x_0 and the random variables $Z(x_i)$ at each of the n observed locations, and $c_{0,0}$ is the model variance for the random variable at the unobserved location x_0 .

The model variogram, $\gamma(x, x')$, is defined as the expected value of the squared difference between the pair of residuals $Z(x) - \mu(x)$ and $Z(x') - \mu(x')$ for locations x and x' . The variogram may be expressed in terms of the model covariances: $\gamma(x, x') = c(x, x) + c(x', x') - 2c(x, x')$, but the variogram is usually easier to estimate directly from the available data than is the covariance function. $\boldsymbol{\gamma}_n$ is the $n \times n$ variogram matrix whose components are the variogram values for pairs of observed control locations. $\boldsymbol{\gamma}_0$ is the n vector whose components are the variogram values for the data locations paired with the interpolation location x_0 .

The Spatial Mean Function and Kriging with Unbiasedness Constraints

A linear kriging estimator, denoted Z_0^* , interpolates the value of the spatial field at an unobserved site x_0 using a linear combination of the control data values $\mathbf{Z}_n$. We write $Z_0^* = \mathbf{Z}_n' \boldsymbol{\lambda}_{n,0}$ where $\boldsymbol{\lambda}_{n,0}$ is an n vector of coefficients that multiply the values at the control locations. The notation emphasizes that kriging coefficients depend on the estimation location x_0 . The coefficients $\boldsymbol{\lambda}_{n,0}$ are chosen to minimize the expected value of the squared interpolation error $(Z_0^* - Z_0)^2$, under the *a priori* probability model.

The minimizing coefficients $\boldsymbol{\lambda}_{n,0}$ are constrained so that the interpolator is “unbiased”. Unbiasedness,

as used here, requires that the unknown Z_0 has the same expected value under the probability model as does the kriging interpolator, Z_0^* , derived from the control data $\mathbf{Z}_n$. Thus, this unbiasedness requirement regards the control data as n random variables, irrespective of their observed values. This unbiasedness constraint for kriging depends only on the *a priori*-modeled mean function, $\mu(x)$, expressed as a function of location x . The constraints on the kriging coefficients are thus given by $\mu_0 = \boldsymbol{\mu}_n' \boldsymbol{\lambda}_{n,0}$.

If the mean function $\mu(x)$ of the *a priori* spatial model is completely specified for every location x , then the *a priori* model can be reduced to a zero-mean model, i.e., we use kriging to interpolate the zero-mean residuals after subtracting the specified mean function. Then $\mu(x)$ is redefined to be zero everywhere for kriging purposes. This is called *simple kriging*. The unbiasedness constraint is trivially satisfied for simple kriging, i.e., there is no effective constraint on the kriging coefficients.

When the mean function $\mu(x)$ of the *a priori* spatial model is not completely specified then kriging typically uses a parameterized mean function that is linear in the unspecified parameters: the *a priori* mean at location x is an unspecified linear combination of “covariates” that are specified at every location x . We can write $\mu(x) = \mathbf{b}'\mathbf{f}(x)$, where $\mathbf{b}$ is an unspecified p vector consisting of mean function parameters, and $\mathbf{f}(x)$ is a specified p vector of covariate values. For example, a covariate might be the known elevation above sea level at every location, or it might simply be the latitude at location x that reflects a north–south linear trend in the mean function.

Set $\mathbf{f}_0 = \mathbf{f}(x_0)$ as the p vector of specified covariate values at the interpolation location x_0 , and $\mathbf{f}_n$ as the specified $p \times n$ matrix of covariate values at the n data locations. This parameterization of the mean function implies that the unbiasedness constraint for kriging is equivalent to $\mathbf{f}_0 = \mathbf{f}_n' \boldsymbol{\lambda}_{n,0}$. This relation represents p separate linear constraints on the kriging coefficients $\boldsymbol{\lambda}_{n,0}$. Optimized linear estimation with this kind of parametric mean function and unbiasedness constraint has been called *universal kriging*. Unbiased universal kriging does not require explicit estimation of the *a priori* mean function parameter vector $\mathbf{b}$, nor is this the goal; the goal remains interpolation of the spatial field $Z(x)$ at unobserved locations.

The unbiasedness constraints on the kriging coefficients involve covariate values only within the control data neighborhood of the interpolation location. Indeed, the implicit unspecified spatial mean parameter, $\mathbf{b}$, can in principle be different in different regions of the spatial domain, thus the spatial mean function viewed over the whole spatial domain becomes effectively nonparametric and possibly quite complex.

The simplest case is a single parameter mean, $p = 1$, with the scalar covariate $f(x)$ having the same value at all observed and unobserved locations x in the estimation neighborhood. This is a mean-stationary *a priori* model. Without loss of generality we can take $f(x) = 1$ and the scalar parameter becomes the unspecified common mean value. It follows that the unbiasedness constraint on the kriging interpolation coefficients is that their sum is unity. Optimized linear kriging estimation in the context of an unspecified constant-mean *a priori* model is called *ordinary kriging*. More general parametric mean functions with $p > 1$ typically include the constant covariate $f(x) = 1$, so that the unit sum of kriging coefficients remains one of the constraints. This inclusion simplifies expressions of the kriging performance measure. For spatial mean functions that are completely specified *a priori*, i.e., the simple kriging situation with a zero-mean function described above, there are no kriging coefficient constraints, including the unit sum constraint.

Commonly, geographic coordinates are themselves used as spatial covariates for the *a priori* spatial mean function specification like for example, a locally linear function of the spatial coordinates. For a two-dimensional region, the spatial mean is then described by an unspecified planar surface over the neighborhood that includes the unobserved interpolation location and the associated observed control locations. The local planar mean surface would vary as we vary the interpolation location and its associated control data locations, creating the potential for mean functions that are globally complex.

Imposing an unbiasedness constraint on the kriging interpolator necessarily degrades its performance, if only modestly. On the other hand, the constraint has the advantage that the expected squared error performance measure is free of the unspecified parameters of the *a priori* model spatial mean function and these

unspecified model mean parameters do not enter into the kriging interpolator optimization.

Optimization of Kriging Interpolation Coefficients

The kriging interpolation coefficients $\lambda_{n,0}$ optimize the performance measure, i.e., they minimize the *a priori* model expected value of the squared interpolation error, subject to the unbiasedness constraints on the kriging coefficients. Because of the unbiasedness constraints, minimization of expected squared error is equivalent to minimization of the error variance. In the case where there is a unit sum constraint on the kriging coefficients the error variance can be expressed in terms of the *a priori* model variogram, viz.,

$$\text{Var}\{Z_0^* - Z_0\} = 2\lambda_{n,0}'\gamma_0 - \lambda_{n,0}'\gamma_n\lambda_{n,0} \quad (1)$$

The n components of the vector γ_0 are the respective expected squared differences between the values at the n control data locations and the interpolation location, as derived from the *a priori* model variogram described earlier. The entries of the $n \times n$ matrix γ_n are likewise obtained from the variogram for the control data location pairs. Model error estimation variance involves the variogram function only for interpoint separations on the neighborhood scale of the control data locations.

The optimized kriging interpolation coefficients reflect the geometry of the control data locations in relation to the interpolation location x_0 . Typically, data closer to x_0 will have larger coefficients. Data that occur in spatial clusters will tend to each have smaller coefficients reflecting their redundancy because of their proximity, thus kriging is said to decluster data. Kriging coefficients can be negative.

The parameterization of the *a priori* model mean, as a function of spatial location, has two important implications. First, the parameterized mean function constrains the permissible unbiased interpolators. Second, the specification of the spatial mean function influences the spatial variogram. Roughly, a more complex mean function will imply that the spatial field has smaller residuals, hence a smaller variogram. The effect of a more complex mean parameterization on kriging interpolation errors is not always clear: smaller errors are associated with a smaller variogram, but larger errors are associated with more

restrictive unbiasedness constraints that accompany a more complex spatial mean parameterization.

Modeling Spatial Variograms

The performance of the unbiased linear kriging interpolator can be expressed in terms of the *a priori* variogram model that gives the expected squared difference between values of the spatial residual field at every pair of spatial locations. Typically, the variogram model is selected from a parametric class of models that are necessarily negative definite. In kriging practice, estimated parameters of the variogram model are treated as though they are specified, unlike the case of a parametric spatial mean function.

For example, the variogram may be specified as a function of the geographic vector separation between spatial locations. In this case the model is said to be second-order stationary, i.e., invariant to geographic shifting of a location pair. Further, if the variogram model depends on the geographic separation only through geographic distance, then the variogram model is said to be isotropic, i.e., invariant to geographic rotation. An example is a variogram of the form $v_1 + v_2(1 - c^d)$, where d is geographic distance between location pairs and v_1, v_2, c are positive-valued parameters with $c < 1$. This variogram model is typical in that it increases with increasing separation toward an asymptote.

Stationarity implies that, on average, residual variability is the same in all parts of the spatial field of interest. In particular, model-calculated kriging interpolation performance will depend only on the relative spatial configuration of the control data locations with respect to the interpolation location, but not on where this configuration occurs in the region of interest, nor on the observed values at the control data locations. Stationarity is imposed mainly as a model of convenience because it gives a compact variogram model, that is, a function only of separation and makes it convenient to estimate adjustable variogram parameters from the observed data. On the other hand, nonstationary variogram models could be used to assign lower kriging interpolation error variance in those parts of the region that have less spatial variability and/or greater spatial continuity.

Parameters of a parametric spatial covariance or variogram model are typically inferred from the

available field observations. If the spatial field is considered to have a constant-mean function then a plot of squared differences between data values as a function of their spatial separation can form the basis of variogram parameter estimation. A better approach is to use squared linear *contrasts* of the observations for this purpose. A contrast is a linear combination of the observations having zero expected value under the assumed spatial mean model. If there are n observation locations and p linear constraints on the mean function, then there will be $n - p$ linearly independent contrasts that can be used for variogram estimation. These linear contrasts form the basis of residual maximum-likelihood (REML) approaches to variogram estimation for Gaussian spatial models as reported by Kitanidis [5].

The variogram needs to be specified only for spatial lags that are on scales comparable to the separation among local control data values used for interpolation. For this reason, estimation of variogram models should focus on the appropriate spatial scale and should use data contrasts for estimation that involve data locations with comparable separation.

Model-derived error variances for kriging interpolators are typically computed from the fitted variogram model, treating the fitted model as though it were known. However, the fitted variogram parameters actually depend on the available data so that the model interpolation error variance, computed from the fitted variogram, is itself a data-dependent estimate of the interpolation error variance. The statistical properties of the interpolation error variance estimate require a more complete specification of the random field model.

Nugget Effect and Measurement Error

Models for spatial variograms can have different spatial scales. The short scale is smaller than the separation between data locations. The intermediate scale is that comparable to the separation among control data locations used for interpolation, called *the neighborhood scale*. The third scale is a regional scale comparable to separation between nonproximate observations. Short-scale variability is commonly inferred by extrapolation of neighborhood scale variability, although a better method is to have the sampling design incorporate closely spaced observations. When extrapolation of variograms to

short separations suggests a discontinuity at zero separation, this discontinuity is called *the nugget effect*. The nugget effect is a contribution to variability without spatial continuity. Parametric variogram models commonly include a parameter for the discontinuity at zero separation.

Measurement error, if appreciable, can be regarded as a spatial field superposed on the underlying field of interest. Then the data at the observed locations include the contributions of measurement error. Frequently, measurement errors are taken to be spatially uncorrelated on all spatial scales. In this case, the measurement error variance acts statistically like an addition to the short-scale nugget effect of the underlying field. Both the nugget effect and the zero-mean measurement error have no effect on unbiasedness constraints for kriging coefficients. However, these short-scale contributions to spatial variability can substantially affect the optimization of kriging coefficients and the model-based performance of interpolators, including kriging interpolators. For example, the optimized interpolator coefficients associated with the control data become more nearly equal and less dependent on their proximity to the interpolation location when the nugget effect is larger. The separation of the nugget effect into short-scale variability and measurement error plays no role in optimizing the kriging interpolator, although the model-computed interpolation error variance will be too large without such separation.

Cokriging and Multivariate Spatial Estimation

A spatial field is a multivariate field if, at each location x , $\mathbf{Z}(x)$ is a vector of attributes. For example, $\mathbf{Z}(x)$ could be a suite of pollutant concentrations, surface temperature, humidity, and wind speed. One might consider each of these spatial attributes as a separate field with separate modeling as described above. However, interpolation of ozone concentration, say, could have a smaller error variance if the available ozone data were supplemented by the available surface air temperature data. The ozone interpolator can be a linear function of both the ozone data and the temperature data at the control locations. The control data locations for the two variables need not be the same; indeed, temperature data may be more widely available, even at the ozone interpolation location (see **Air Pollution Risk**).

Table 1 Optimized kriging weights and interpolation errors for five variogram models

				$r = 0.1$			$r = 0.4$		$r = 0.8$
				$v = 10\%$	$v = 50\%$		$v = 10\%$	$v = 50\%$	$v = 10\%$
+	+	+	+	inner wt	18%	15%	24%	13%	21%
				outer wt	4%	5%	1%	6%	2%
+	+	+	+	E	0.95	0.97	0.71	0.79	0.43
				E^*	0.79	0.89	0.57	0.77	0.38

The expected squared error of the interpolator is now computed under an *a priori* model that expresses the joint statistical properties of both attributes. There will be model unbiasedness constraints for both the ozone and temperature coefficients. Such constrained optimized linear interpolation is called *cokriging*. For each of the variables, unbiasedness constraints will depend on the specified *a priori* spatial mean structures. For example, if each variable is assigned a common *a priori* mean value at all control data locations and the interpolation location, then unbiasedness constrains the coefficients that multiply the ozone observations to sum to unity, and the coefficients that multiply the covariate temperature observations to sum to zero. The optimization of the linear cokriging estimate requires modeling of the cross-covariance between variable pairs at location pairs.

If only the ozone randomness figures in the *a priori* probability model, while the covariate temperature information is considered fixed, then the optimization and coefficient constraints are different. For example, suppose that the mean ozone at location x , conditional on the covariate field, is taken to be an unspecified linear function of the temperature covariate at location x . Then we are in the universal kriging modeling situation described above. With this approach there is no need to model cross-covariances of ozone and temperature.

Illustration

The illustration is based on gridded data that are typical of reconnaissance surveys and satellite measurements. For definiteness suppose that radioactivity is measured at the grid locations and that we wish to quantify the uncertainty associated with kriging interpolation of the data to unmeasured locations. A variogram model, fit to the data, expresses

how radioactivity differences at two locations are related, on average, to their spatial separation d . This illustration uses a variogram model that is proportional to $2v + 2(1 - v)(1 - c^d)$, where v is the nugget variance (measurement error and short-scale variability) as a proportion of the total variance, and $r = (1 - v)c$ is the correlation between measurements at adjacent grid locations. The root-mean-squared interpolation error E is computed under the specified variogram model.

Suppose one interpolates the center of a measurement grid square using the 12 nearest measurements as control data for interpolation, as shown in Table 1, comprising an inner ring of four control data values and an outer ring of eight control data values.

Table 1 shows the following: the approximate optimized kriging weights assigned to inner-ring data values and outer-ring data values, respectively; the corresponding relative values of the model-based interpolation error E ; the reduced error E^* that would result from a fourfold increase in data density with half the original grid spacing. Five model scenarios are described with different levels of short-scale variability and correlation as represented by the variogram parameters v and r described above.

References

- [1] Krige, D.G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand, *Journal of the Chemical, Metallurgical and Mining Society of South Africa* **52**(6), 119–139.
- [2] Gandin, L.S. (1965). *Objective Analysis of Meteorological Fields*, Israel Program for Scientific Translations, Jerusalem.
- [3] Matheron, G. (1965). *Les Variables Regionalisees et leur Estimation*, Masson, Paris.
- [4] Cressie, N. (1991). *Statistics for Spatial Data*, John Wiley & Sons, New York.

-
- [5] Kitanidis, P.K. (1997). *Introduction to Geostatistics: Applications to Hydrogeology*, Cambridge University Press.
- [6] Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*, Oxford University Press.
- [7] Wackernagel, H. (1998). *Multivariate Geostatistics*, 2nd Edition, Springer.
- [8] Armstrong, M. (1998). *Basic Linear Geostatistics*, Springer.
- [9] Olea, R. (1999). *Geostatistics for Engineers and Earth Scientists*, Kluwer Academic Publishers.
- [10] Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*, Springer.
- [11] Chiles, J.-P. & Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons, New York.

Related Articles

Disease Mapping

Spatial Risk Assessment

Spatiotemporal Risk Analysis

PAUL SWITZER

Large Insurance Losses Distributions

Subexponential Distribution Functions

Subexponential distribution functions (DFs) are a special class of heavy-tailed DFs. The name arises from one of their properties, that their right tail decreases more slowly than any exponential tail; see equation (3). This implies that large values can occur in a sample with nonnegligible probability, which proposes the subexponential DFs as natural candidates for situations where extremely large values occur in a sample compared to the mean size of the data. Such a pattern is often seen in insurance data, for instance in fire, wind–storm, or flood insurance (collectively known as *catastrophe insurance*), and also in financial and environmental data. Subexponential claims can account for large fluctuations in the risk process of a company. Moreover, the subexponential concept has just the right level of generality for risk measurement in insurance and finance models. Various review papers have appeared on subexponentials; see e.g., [1–3]; textbook accounts are in [4–8]. We present two defining properties of subexponential DFs.

Definition 1 (Subexponential distribution functions) Let $(X_i)_{i \in \mathbb{N}}$ be i.i.d. positive random variables with DF F such that $F(x) < 1$ for all $x > 0$. Denote by $\bar{F}(x) = 1 - F(x)$ for $x \geq 0$, the tail of F , and for $n \in \mathbb{N}$,

$$\begin{aligned} \bar{F}^{n*}(x) &= 1 - F^{n*}(x) \\ &= \mathbb{P}(X_1 + \dots + X_n > x), \quad x \geq 0 \end{aligned} \quad (1)$$

the tail of the n -fold convolution of F . F is a *subexponential DF* ($F \in \mathcal{S}$) if one of the following equivalent conditions holds:

- (a) $\lim_{x \rightarrow \infty} \frac{\bar{F}^{n*}(x)}{\bar{F}(x)} = n$ for some (all) $n \geq 2$
- (b) $\lim_{x \rightarrow \infty} \frac{\mathbb{P}(X_1 + \dots + X_n > x)}{\mathbb{P}(\max(X_1, \dots, X_n) > x)} = 1$ for some (all) $n \geq 2$

Remark 1 (a) A proof of the equivalence is based on ($\sim$ means that the quotient of the left hand side

and the right hand side tends to 1 as $x \rightarrow \infty$)

$$\begin{aligned} \frac{\mathbb{P}(X_1 + \dots + X_n > x)}{\mathbb{P}(\max(X_1, \dots, X_n) > x)} &\sim \frac{\bar{F}^{n*}(x)}{n\bar{F}(x)} \xrightarrow{x \rightarrow \infty} 1 \\ &\iff F \in \mathcal{S} \end{aligned} \quad (2)$$

(b) In much of the present discussion, we deal only with the right tail of a DF. This concept can be formalized by denoting two DFs F and G with support unbounded to the right *tail-equivalent* (see **Premium Calculation and Insurance Pricing; Risk Measures and Economic Capital for (Re-)insurers**) if $\lim_{x \rightarrow \infty} \bar{F}(x)/\bar{G}(x) = c \in (0, \infty)$. From Definition 1 and the fact that $\mathcal{S}$ is closed with respect to tail-equivalence, it follows that $\mathcal{S}$ is closed with respect to taking sums and maxima of i.i.d. random variables. Subexponential DFs can also be defined on $\mathbb{R}$ by requiring that F restricted to $(0, \infty)$ is subexponential; see [9–11].

(c) The heavy-tailedness of $F \in \mathcal{S}$ is demonstrated by the implications

$$\begin{aligned} F \in \mathcal{S} &\implies \lim_{x \rightarrow \infty} \frac{\bar{F}(x-y)}{\bar{F}(x)} = 1 \quad \forall y \in \mathbb{R} \\ &\implies \frac{\bar{F}(x)}{e^{-\varepsilon x}} \xrightarrow{x \rightarrow \infty} \infty \quad \forall \varepsilon > 0 \end{aligned} \quad (3)$$

A famous subclass of $\mathcal{S}$ is the class of DFs with regularly varying tails; see [12, 13].

Example 1 (Distribution functions with regularly varying tails) For a positive measurable function f we write $f \in \mathcal{R}(\alpha)$ for $\alpha \in \mathbb{R}$ (f is *regularly varying with index* α) if

$$\lim_{x \rightarrow \infty} \frac{f(tx)}{f(x)} = t^\alpha \quad \forall t > 0 \quad (4)$$

Let $\bar{F} \in \mathcal{R}(-\alpha)$ for $\alpha > 0$, then it has the representation

$$\bar{F}(x) = x^{-\alpha} \ell(x), \quad x > 0 \quad (5)$$

for some $\ell \in \mathcal{R}(0)$. To check Definition 1(a) for $n = 2$, split the convolution integral and use partial integration to obtain

$$\begin{aligned} \frac{\bar{F}^{2*}(x)}{\bar{F}(x)} &= 2 \int_0^{x/2} \frac{\bar{F}(x-y)}{\bar{F}(x)} dF(y) + \frac{(\bar{F}(x/2))^2}{\bar{F}(x)}, \\ &\quad x > 0 \end{aligned} \quad (6)$$

Immediately, by equation (4), the last term tends to 0. The integrand satisfies $\bar{F}(x-y)/\bar{F}(x) \leq \bar{F}(x/2)/\bar{F}(x)$ for $0 \leq y \leq x/2$; hence, Lebesgue dominated convergence applies and, since F satisfies equation (3), the integral on the right hand side tends to 1 as $x \rightarrow \infty$. Examples of DFs with regularly varying tail include the Pareto, Burr, transformed beta (also called generalized F), log-gamma, and stable DFs (see [6]).

Example 2 (Further subexponential distributions)

Apart from the DFs in Example 1, the lognormal, the two Benktander families and the heavy-tailed Weibull (shape parameter less than 1) also belong to $\mathcal{S}$. The integrated tail DFs (see equation (11)) of all of these DFs are also in $\mathcal{S}$ provided they have a finite mean; see Table 1.2.6 in [6].

Insurance Risk Models

The Cramér–Lundberg Model

The classical insurance risk model is the *Cramér–Lundberg model* (cf. [6–8]), where the *claim times* constitute a Poisson process, i.e., the *interclaim times* $(T_n)_{n \in \mathbb{N}}$ are i.i.d. exponential random variables with parameter $\lambda > 0$. Furthermore, the claim sizes $(X_k)_{k \in \mathbb{N}}$ (independent of the claims arrival process) are i.i.d. positive random variables with DF F and $\mathbb{E}(X_1) = \mu < \infty$. The *risk process* for initial reserve $u \geq 0$ and *premium rate* $c > 0$ is defined as

$$R(t) = u + ct - \sum_{k=1}^{N(t)} X_k, \quad t \geq 0 \quad (7)$$

where (with the convention $\sum_{i=1}^0 a_i := 0$) $N(0) = 0$ and $N(t) = \sup \{k > 0 : \sum_{i=1}^k T_i \leq t\}$ for $t > 0$. We denote the *ruin time* by

$$\tau(u) = \inf \left\{ t > 0 : u + ct - \sum_{k=1}^{N(t)} X_k < 0 \right\}, \quad u \geq 0 \quad (8)$$

The *ruin probability in infinite time* (see **Asset–Liability Management for Nonlife Insurers; Longevity Risk and Life Annuities; Reinsurance; Solvency**;

Ruin Probabilities: Computational Aspects) is defined as

$$\begin{aligned} \psi(u) &= \mathbb{P}(R(t) < 0 \text{ for some } 0 \leq t < \infty \mid R(0) = u) \\ &= \mathbb{P}(\tau(u) < \infty), \quad u \geq 0 \end{aligned} \quad (9)$$

By definition of the risk process, ruin can occur only at the claim times (see also Figure 1), hence for $u \geq 0$,

$$\begin{aligned} \psi(u) &= \mathbb{P}(R(t) < 0 \text{ for some } t \geq 0 \mid R(0) = u) \\ &= \mathbb{P}\left(u + \sum_{k=1}^n (cT_k - X_k) < 0 \text{ for some } n \in \mathbb{N}\right) \end{aligned} \quad (10)$$

Provided that $\mathbb{E}(cT_1 - X_1) = c/\lambda - \mu > 0$, then $(R(t))_{t \geq 0}$ has a positive drift and hence, $R(t) \rightarrow +\infty$ a.s. as $t \rightarrow \infty$.

A ladder height analysis shows that the *integrated tail distribution function* (see **Extreme Value Theory in Finance**)

$$F_I(x) = \frac{1}{\mu} \int_0^x \bar{F}(y) dy, \quad x \geq 0 \quad (11)$$

of F is the DF of the first undershoot of $R(t) - u$ below 0, under the condition that $R(t)$ falls below u in finite time. Setting $\rho = \lambda\mu/c < 1$, the number $\tilde{N}$ of times, where $R(t)$ achieves a new local minimum in finite time plus 1, is geometrically distributed with parameter $(1 - \rho)$, i.e. $\mathbb{P}(\tilde{N} = n) = (1 - \rho)\rho^n$, $n \in \mathbb{N}_0$. As $(R(t))_{t \geq 0}$ is a Markov process, its sample path splits into i.i.d. cycles, each starting with a new

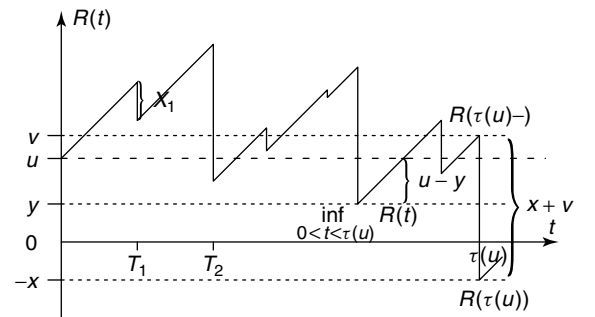


Figure 1 Sample path of a risk process in the Cramér–Lundberg model

local minimum of the risk process. Combining these findings, we obtain the following representation of the ruin probability.

Theorem 1 (Pollaczek–Khinchine formula) *Consider the risk process 7 in the Cramér–Lundberg model with $\rho = \lambda\mu/c < 1$. Let F_I be as in equation (11). Then the ruin probability is given by*

$$\psi(u) = (1 - \rho) \sum_{n=1}^{\infty} \rho^n \overline{F_I^{n*}}(u), \quad u \geq 0 \quad (12)$$

We are interested in the ruin event represented by a probabilistic description of the last cycle before ruin, i.e., the following quantities:

- $R(\tau(u))$ the level of the risk process at ruin,
- $R(\tau(u)-)$ the level of the risk process
 just before ruin,
- $\inf_{0 \leq t < \tau(u)} R(t)$ the infimum of the risk process
 before ruin

The following result is part of the quintuple law presented in [14] with references to previous work; see also Theorem 7.7 of [15]. Its proof is based on a careful ladder height analysis of the risk process.

Theorem 2 *Consider the risk process 7 in the Cramér–Lundberg model with $\rho = \lambda\mu/c < 1$. The renewal measure of the descending ladder height process of $(ct - \sum_{k=1}^{N(t)} X_k)_{t \geq 0}$ is given by*

$$V(u) = \frac{1}{c} \sum_{n=0}^{\infty} \rho^n F_I^{n*}(u), \quad u \geq 0 \quad (13)$$

where F_I is as in equation (11). Then for $x > 0$, $v \geq y$ and $y \in [0, u]$,

$$\begin{aligned} \mathbb{P}(-R(\tau(u)) \in dx, R(\tau(u)-) \in dv, \inf_{0 \leq t < \tau(u)} R(t) \in dy) \\ = \lambda F(dx + v) V(u - dy) dv \end{aligned} \quad (14)$$

The Lévy Risk Model

A natural and useful generalization of the Cramér–Lundberg model as in equation (7) is to replace

the compound Poisson process $(\sum_{k=1}^{N(t)} X_k)_{t \geq 0}$ by a general subordinator $(S(t))_{t \geq 0}$. A *subordinator* is an increasing Lévy process, i.e., it starts at 0, has independent and stationary increments, and has a.s. increasing sample paths. A subordinator can be characterized by its *Lévy–Khinchine representation* $\mathbb{E}(\exp(ivS(t))) = \exp(t\varphi(v))$ for $t \geq 0$, $v \in \mathbb{R}$ with

$$\varphi(v) = \int_{\mathbb{R}_+} (e^{ivx} - 1) v(dx) \quad (15)$$

where v is a measure on $\mathbb{R}_+ := (0, \infty)$, called *Lévy measure*, satisfying $\int_{\mathbb{R}_+} (1 \wedge |x|) v(dx) < \infty$; see the monographs [15–18] for more details on Lévy processes. The upward movement of the subordinator represents the total claim amount process, extending the compound Poisson model by modeling many small claims between more severe claims at Poisson times.

On top of this already rather general model, uncertainty in the aggregated claims and in the premium income can be modeled by a Brownian motion $(B(t))_{t \geq 0}$ giving the risk process

$$R(t) = u + ct - S(t) + \sigma B(t), \quad t \geq 0 \quad (16)$$

where u and c are as in the Cramér–Lundberg model and $\sigma \geq 0$. More general Lévy insurance models are investigated in [14, 19–21]; see also the monograph [15]. More about the Cramér–Lundberg model perturbed by a Brownian motion can be found in [22–26] and perturbed by an α -stable Lévy motion in [27, 28]. The class of Γ -subordinators with and without perturbation by a Brownian motion have been investigated in [29, 30].

In this model, the ruin time is defined as

$$\begin{aligned} \tau(u) = \inf\{t > 0 : u + ct - S(t) + \sigma B(t) < 0\}, \\ u \geq 0 \end{aligned} \quad (17)$$

and the ruin probability by $\psi(u) = \mathbb{P}(\tau(u) < \infty)$.

Again, we have to assume that $R(t) \rightarrow +\infty$ a.s. as $t \rightarrow \infty$, which is implied by $\rho := \mathbb{E}(S(1))/c < 1$, i.e., the premium income outweighs the expected total claims.

Analogously to equation (11), we define the integrated tail DF of v as

$$v_I(x) = \frac{1}{\mathbb{E}(S(1))} \int_0^x v(y, \infty) dy, \quad x \geq 0 \quad (18)$$

In the Cramér–Lundberg model, $v(y, \infty) = \lambda \bar{F}(y)$ and $\mathbb{E}(S(1)) = \lambda \mu$, hence we get back $v_I = F_I$ and $\rho = \lambda \mu / c$.

To present the results in this general setup, we also require the exponential DF

$$G(x) = 1 - \exp\left(\frac{-2cx}{\sigma^2}\right) \mathbf{1}_{\{x > 0\}}, \quad x \geq 0 \quad (19)$$

which is the DF of $(\sup_{t \geq 0} \{-ct - \sigma B(t)\})_{t \geq 0}$. Finally, we have to build convolutions of v_I with itself and with convolution powers of G . As an explaining example, we recall that

$$\begin{aligned} G * v_I(x) &= \int_0^x G(x-y) dv_I(y) \\ &= \frac{1}{\mathbb{E}(S(1))} \int_0^x G(x-y) v(y, \infty) dy \\ &\quad \text{for } x > 0 \end{aligned} \quad (20)$$

Then the ruin probability satisfies the following equation (see [19, Theorem 3.1]).

Theorem 3 (Pollaczek–Khinchine Formula) *Consider the risk process (equation (16)) in the Lévy risk model with $\rho = \mathbb{E}(S(1))/c < 1$. Let v_I be as in equation (18) and G as in equation (19). Then the ruin probability can be represented as*

$$\psi(u) = (1 - \rho) \sum_{n=1}^{\infty} \rho^n \overline{G^{(n+1)*} * v_I^{n*}}(u), \quad u \geq 0 \quad (21)$$

For this model, the result corresponding to Theorem 2 reads as follows.

Theorem 4 *Consider the risk process (equation (16)) in the Lévy risk model with $\rho = \mathbb{E}(S(1))/c < 1$. The renewal measure of the descending ladder height process of $(ct - S(t) + \sigma B(t))_{t \geq 0}$ is given by*

$$V(u) = \frac{1}{c} \sum_{n=0}^{\infty} \rho^n (G^{(n+1)*} * v_I^{n*})(u), \quad u \geq 0 \quad (22)$$

where v_I is as in equation (18) and G as in (19). Then for $x > 0$, $v \geq y$ and $y \in [0, u]$,

$$\begin{aligned} \mathbb{P}(-R(\tau(u)) \in dx, R(\tau(u)-) \in dv, \\ \inf_{0 \leq t < \tau(u)} R(t) \in dy) \\ = v(dx + v)V(u - dy) dv \end{aligned} \quad (23)$$

Risk Theory in the Presence of Heavy Tails

Asymptotic Ruin Theory

In this section we investigate the occurrence of large, possibly ruinous claims modeled by $v_I \in \mathcal{S}$, which translates in the Cramér–Lundberg model to $F_I \in \mathcal{S}$. We refer to [20, 21] for more details on the results of this section; these papers also include more general Lévy insurance risk models. The results for the Cramér–Lundberg model can also be found in [4].

The following part (a) of Proposition 1 is found in [31], Proposition 1. The result (b) can partly be found already in [5]. Its present form goes back to [31–33].

Proposition 1 (a) *Let $H = G * F$ be the convolution of two DFs on $(0, \infty)$. If $F \in \mathcal{S}$ and $\bar{G}(x) = o(\bar{F}(x))$ as $x \rightarrow \infty$, then $H \in \mathcal{S}$.*

(b) *Suppose $(p_n)_{n \geq 0}$ defines a probability measure on $\mathbb{N}_0$ such that $\sum_{n=0}^{\infty} p_n (1 + \varepsilon)^n < \infty$ for some $\varepsilon > 0$ and $p_k > 0$ for some $k \geq 2$. Let*

$$K(x) = \sum_{n=0}^{\infty} p_n H^{n*}(x), \quad x \geq 0 \quad (24)$$

Then

$$\begin{aligned} H \in \mathcal{S} &\iff \lim_{x \rightarrow \infty} \frac{\bar{K}(x)}{\bar{H}(x)} = \sum_{n=1}^{\infty} n p_n \\ &\iff K \in \mathcal{S} \quad \text{and} \quad \bar{H}(x) \neq o(\bar{K}(x)) \\ &\quad \text{as } x \rightarrow \infty \end{aligned} \quad (25)$$

For the ruin probability, this Proposition along with the Pollaczek–Khinchine formula (Theorem 3) implies the following result.

Theorem 5 (Ruin Probability) *Consider the risk process (equation (16)) in the Lévy risk model with $\rho = \mathbb{E}(S(1))/c < 1$. If $v_I \in \mathcal{S}$, then as $u \rightarrow \infty$*

$$\begin{aligned} \psi(u) &\sim \frac{\rho}{1 - \rho} \bar{v}_I(u) \\ &= \frac{1}{c - \mathbb{E}(S(1))} \int_u^{\infty} v(y, \infty) dy \end{aligned} \quad (26)$$

Sample Path Leading to Ruin

Not surprisingly, the asymptotic behavior of the quantities in Theorem 4 are consequences of extreme

value theory; see [6] or any other book on extreme value theory for background. As shown in [34], under weak regularity conditions, a subexponential DF F belongs to the *maximum domain of attraction* of an extreme value DF G_α , $\alpha \in (0, \infty]$, (we write $F \in MDA(G_\alpha)$), where

$$G_\alpha(x) = \Phi_\alpha(x) = \exp(-x^{-\alpha})\mathbf{1}_{\{x>0\}},$$

$$\text{for } \alpha < \infty \quad \text{and} \quad (27)$$

$$G_\infty(x) = \Lambda(x) = \exp(-e^{-x}) \quad (28)$$

The meaning of $F \in MDA(G_\alpha)$ is that there exist sequences of constants $a_n > 0$, $b_n \in \mathbb{R}$ such that

$$\lim_{n \rightarrow \infty} n\bar{F}(a_n x + b_n) = -\log G_\alpha(x)$$

$$\forall x \in \begin{cases} (0, \infty) & \text{if } \alpha \in (0, \infty) \\ \mathbb{R} & \text{if } \alpha = \infty \end{cases} \quad (29)$$

The following result describes the behavior of the process at ruin, and the upcrossing event itself; see [21].

Theorem 6 (Sample Path Leading to Ruin) *Consider the risk process 16 in the Lévy risk model with $\rho = \mathbb{E}(S(1))/c < 1$. Define $\tilde{v}(x) := v(1, x)/v(1, \infty)$ for $x > 1$ and assume that $\tilde{v} \in MDA(G_{\alpha+1})$ for $\alpha \in (0, \infty]$. Define $a(\cdot) = v_I(\cdot, \infty)/v(\cdot, \infty)$. Then, in $\mathbb{P}(\cdot \mid \tau(u) < \infty)$ distribution,*

$$\left(\frac{R(\tau(u)-) - u}{a(u)}, \frac{-R(\tau(u))}{a(u)} \right) \longrightarrow (V_\alpha, T_\alpha),$$

$$\text{as } u \rightarrow \infty \quad (30)$$

V_α and T_α are positive random variables with DF satisfying for $x, y > 0$

$$\mathbb{P}(V_\alpha > x, T_\alpha > y) = \begin{cases} \left(1 + \frac{x+y}{\alpha}\right)^{-\alpha} & \text{if } \alpha \in (0, \infty) \\ e^{-(x+y)} & \text{if } \alpha = \infty \end{cases} \quad (31)$$

Remark 2 Extreme value theory (see **Individual Risk Models; Extreme Value Theory in Finance; Extreme Values in Reliability; Mathematics of Risk and Reliability: A Select History; Copulas and Other Measures of Dependency**) and Theorem 4 form the basis of this result: recall first that $\tilde{v} \in MDA(\Phi_{\alpha+1})$ is equivalent to $v(\cdot, \infty) \in \mathcal{R}(-(\alpha+1))$ and, hence, to $v_I \in MDA(\Phi_\alpha)$ by Karamata's

theorem. The normalizing function $a(u)$ tends in the subexponential case to infinity as $u \rightarrow \infty$. For $v(\cdot, \infty) \in \mathcal{R}(-(\alpha+1))$ Karamata's theorem gives $a(u) \sim u/\alpha$ as $u \rightarrow \infty$.

Insurance Risk Models with Investment

The following extension of the insurance risk process has attracted attention over recent years. It is based on the simple fact that an insurance company not only deals with the risk coming from insurance claims, but also invests capital on a large scale in financial markets. To keep the level of sophistication moderate, we assume that the insurance risk process $(R(t))_{t \geq 0}$ is the Cramér–Lundberg model as defined in equation (7). We suppose that the insurance company invests its reserve into a Black–Scholes type market (see **Default Risk**) consisting of a *riskless bond* and a *risky stock* modeled by an exponential Lévy process (see **Lévy Processes in Asset Pricing**). Their price processes follow the equations

$$P_0(t) = e^{\delta t} \quad \text{and} \quad (32)$$

$$P_1(t) = e^{L(t)}, \quad t \geq 0 \quad (33)$$

The constant $\delta > 0$ is the *riskless interest rate*, $(L(t))_{t \geq 0}$ denotes a Lévy process characterized by its Lévy–Khintchine representation $\mathbb{E}(\exp(ivL(t))) = \exp(t\varphi(v))$ for $t \geq 0$, $v \in \mathbb{R}$, with

$$\varphi(v) = ivm - \frac{1}{2}v^2\sigma^2$$

$$+ \int_{\mathbb{R}} (e^{ivx} - 1 - ivx\mathbf{1}_{[-1,1]}(x)) v(dx) \quad (34)$$

The quantities (m, σ^2, ν) are called the *generating triplet* of the Lévy process L . Here, $m \in \mathbb{R}$, $\sigma^2 \geq 0$, and ν is a Lévy measure satisfying $\nu(\{0\}) = 0$ and $\int_{\mathbb{R}} (1 \wedge |x|^2) \nu(dx) < \infty$. We assume that

$$0 < \mathbb{E}(L(1)) < \infty \quad \text{and} \quad \text{either}$$

$$\sigma > 0 \quad \text{or} \quad \nu(-\infty, 0) > 0 \quad (35)$$

such that $L(t)$ is negative with positive probability and hence, $P_1(t)$ is less than one with positive probability.

We denote by $\theta \in (0, 1]$ the *investment strategy*, which is the fraction of the reserve invested into the risky asset. For details and more background on this

model, we refer to [35]. Then the *investment process* is the solution of the stochastic differential equation

$$dP_\theta(t) = P_\theta(t-) d((1 - \theta)\delta t + \theta\widehat{L}(t)), \quad t \geq 0 \quad (36)$$

where $(\widehat{L}(t))_{t \geq 0}$ is a Lévy process satisfying $\mathcal{E}(\widehat{L}(t)) = \exp(L(t))$, and $\mathcal{E}$ denotes the stochastic exponential (cf. [36]). The solution of equation (36) is given by

$$P_\theta(t) = e^{L_\theta(t)}, \quad t \geq 0 \quad (37)$$

where L_θ is such that

$$\mathcal{E}((1 - \theta)\delta t + \theta\widehat{L}(t)) = \exp(L_\theta(t)) \quad (38)$$

The *integrated risk process* of the insurance company is given by

$$I_\theta(t) = e^{L_\theta(t)} \left(u + \int_0^t e^{-L_\theta(v)} (c \, dv - dS(v)) \right), \quad t \geq 0 \quad (39)$$

In this model with L and R independent, we are interested in the ruin probability

$$\psi_\theta(u) = \mathbb{P}(I_\theta(t) < 0 \text{ for some } t \geq 0 \mid I_\theta(0) = u), \quad u \geq 0 \quad (40)$$

To find the asymptotic ruin probability for such models, there exist two approaches. The first one is analytic and derives an integro-differential equation for the ruin probability, whose asymptotic solution can, in certain cases, be found. This works in particular for the case where the investment process is a geometric Brownian motion with drift, and has been considered in [37–40]; see also [41] for an overview.

This method breaks down for general exponential Lévy investment processes. However, equation (39) can be viewed as a continuous time random recurrence equation, often called a *generalized Ornstein–Uhlenbeck process*, cf. [42] and references therein. Splitting up the integral in an appropriate way, asymptotic theory for discrete random recurrence equations can be applied. The theoretical basis for this approach can be found in [43] and is based on the seminal paper [44]. This approach is applied in [45, 46].

The common starting point is the process $(I_\theta(t))_{t \geq 0}$ as defined in equation (39), which is the same as the process (9) in [45]. We present Theorem

5 (a) of [45] below and formulate the necessary conditions in our terminology. The Laplace exponent of L_θ is denoted by $\phi_\theta(v) = \log \mathbb{E}(e^{-vL_\theta(1)})$. If we assume that

$$\mathcal{V}_\infty := \{v \geq 0 : \mathbb{E}(e^{-vL_1}) < \infty\} \text{ is right open} \quad (41)$$

then (cf. e.g., Lemma 4.1 of [35]) there exists a unique $\kappa(\theta)$ such that

$$\phi_\theta(\kappa(\theta)) = 0 \quad (42)$$

Theorem 7 (Ruin probability) *We consider the integrated risk process $(I_\theta(t))_{t \geq 0}$ as in equation (39) satisfying equations (35) and (41). Furthermore, we assume for the claim size variable X that $\mathbb{E}(X^{\kappa(\theta)+\epsilon}) < \infty$ for some $\epsilon > 0$, where $\kappa(\theta)$ is given in (42), and we assume that $\phi_\theta(2) < \infty$. Let U be uniformly distributed on $[0, 1]$ and independent of $(L_\theta(t))_{t \geq 0}$. Assume that the DF of $L_\theta(U)$ has an absolutely continuous component. Then there exists a constant $C_1 > 0$ such that*

$$\psi_\theta(u) \sim C_1 u^{-\kappa(\theta)}, \quad \text{as } u \rightarrow \infty \quad (43)$$

Remark 2 If the insurance company invests its money in a classical Black–Scholes model, where $L(t) = \gamma t + \sigma^2 W(t)$, $t \geq 0$, is a Brownian motion with drift $\gamma > 0$, variance $\sigma^2 > 0$, and $(W(t))_{t \geq 0}$ is a standard Brownian motion, then $(L_\theta(t))_{t \geq 0}$ is again a Brownian motion with drift $\gamma_\theta = \theta\gamma + (1 - \theta)(\delta + \frac{\sigma^2}{2}\theta)$ and variance $\sigma_\theta^2 = \theta^2\sigma^2$. Since

$$\phi_\theta(s) = -\gamma_\theta s + \frac{\sigma_\theta^2}{2} s^2, \quad s \geq 0 \quad (44)$$

we obtain $\kappa(\theta) = 2\gamma_\theta/\sigma_\theta^2$. Hence, if $\mathbb{E}(X^{\max\{1, \kappa(\theta)+\epsilon\}}) < \infty$ for some $\epsilon > 0$, Theorem 7 applies.

Whereas the ruin probability is indeed the classical insurance risk measure, other risk measures have been suggested for investment risk. The *discounted net loss process*

$$V_\theta(t) = \int_0^t e^{-L_\theta(v)} (dS(v) - c \, dv), \quad t \geq 0 \quad (45)$$

converges a.s. to V_θ^* when $\phi_\theta(1) < \lambda$, and measures the risk of the insurance company, since, in particular,

the relation

$$\mathbb{P}(I_\theta(t) < 0 \mid I_\theta(0) = u) = \mathbb{P}(V_\theta(t) > u),$$

$$t, u \geq 0 \quad (46)$$

holds. This fact has been exploited in [35] to review this model from the point of view of the investing risk manager. Invoking the value-at-risk as alternative risk measure to the ruin probability, the tail behavior of V_θ^* is derived. For heavy tailed as well as for light tailed claim sizes V_θ^* has right Pareto tail, indicating again the riskiness of investment. See also [47, 48] for further insight into this problem.

Acknowledgment

The paper was written when the first author visited the Department of Operations Research and Industrial Engineering at Cornell University. She takes pleasure in thanking the colleagues there for their hospitality. Financial support by the Deutsche Forschungsgemeinschaft through a research grant is gratefully acknowledged.

References

- [1] Goldie, C.M. & Klüppelberg, C. (1998). Subexponential distributions, in *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, R.J. Adler & R.E. Feldman, eds, Birkhäuser, Boston, pp. 435–459.
- [2] Sigman, K. (1999). A primer on heavy-tailed distributions, *Queueing Systems* **33**, 125–152.
- [3] Klüppelberg, C. (2004). Subexponential distributions, in *Encyclopedia of Actuarial Science*, B. Sundt & J. Teugels, eds, John Wiley & Sons, pp. 1626–1633.
- [4] Asmussen, S. (2001). *Ruin Probabilities*, World Scientific, Singapore.
- [5] Athreya, K.B. & Ney, P.E. (1972). *Branching Processes*, Springer, Berlin.
- [6] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.
- [7] Mikosch, T. (2004). *Non-Life Insurance Mathematics*, Springer, Berlin.
- [8] Rolski, T., Schmidli, H., Schmidt, V. & Teugels, J. (1999). *Stochastic Processes for Insurance and Finance*, John Wiley & Sons, Chichester.
- [9] Pakes, A.G. (2004). Convolution equivalence and infinite divisibility, *Journal of Applied Probability* **41**, 407–424.
- [10] Pakes, A.G. (2007). Convolution equivalence and infinite divisibility: Corrections and corollaries, *Journal of Applied Probability* **44**, 295–305.
- [11] Watanabe, T. (2008). Convolution equivalence and distributions of random sums, *Probability Theory and Related Fields*, to appear.
- [12] Bingham, N.H., Goldie, C.M. & Teugels, J.L. (1987). *Regular Variation*, Cambridge University Press, Cambridge.
- [13] Resnick, S.I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer, New York.
- [14] Doney, R. & Kyprianou, A. (2006). Overshoots and undershoots of Lévy processes, *Annals of Applied Probability* **16**, 91–106.
- [15] Kyprianou, A. (2006). *Introductory Lectures on Fluctuation of Lévy Processes with Applications*, Springer, Berlin.
- [16] Schoutens, W. (2003). *Lévy Processes in Finance*, John Wiley & Sons, Chichester.
- [17] Bertoin, J. (1996). *Lévy Processes*, Cambridge University Press, Cambridge.
- [18] Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge.
- [19] Huzak, M., Perman, M., Šikć, H. & Vondraček, Z. (2004). Ruin probabilities and decompositions for general perturbed risk processes, *Annals of Applied Probability* **14**, 1378–1397.
- [20] Klüppelberg, C., Kyprianou, A.E. & Maller, R.A. (2004). Ruin probabilities and overshoots for general Lévy insurance risk processes, *Annals of Applied Probability* **14**, 1766–1801.
- [21] Klüppelberg, C. & Kyprianou, A. (2006). On extreme ruinous behavior of Lévy insurance processes, *Journal of Applied Probability* **43**, 594–598.
- [22] Gerber, H.U. (1970). An extension of the renewal equation and its application in the collective theory of risk, *Skandinavisk Aktuari Tidskrift* **53**, 205–210.
- [23] Veraverbeke, N. (1993). Asymptotic estimates for the probability of ruin in a Poisson model with diffusion, *Insurance Mathematics and Economics* **13**, 57–62.
- [24] Dufresne, F. & Gerber, H.U. (1991). Risk theory for a compound poisson process that is perturbed by diffusion, *Insurance Mathematics and Economics* **10**, 51–59.
- [25] Gerber, H.U. & Lundry, F. (1998). On the discounted penalty at ruin in a jump-diffusion and perpetual put option, *Insurance Mathematics and Economics* **22**, 263–276.
- [26] Gerber, H.U. & Shiu, E.S.W. (1997). The joint distribution of the time of ruin, the surplus immediately before ruin, and the deficit at ruin, *Insurance Mathematics and Economics* **21**, 129–137.
- [27] Furrer, H. (1998). Risk processes perturbed by α -stable Lévy motion, *Scandinavian Actuarial Journal* **59**–74.
- [28] Schmidli, H. (2001). Distribution of the first ladder height of a stationary risk process perturbed by α -stable Lévy motion, *Insurance Mathematics and Economics* **28**, 13–20.
- [29] Dufresne, F., Gerber, H.U. & Shiu, E.W. (1991). Risk theory with the gamma process, *Astin Bulletin* **21**, 177–192.

- [30] Yang, H. & Zhang, L. (2001). Spectrally negative Lévy processes with applications in risk theory, *Advances in Applied Probability* **33**, 281–291.
- [31] Embrechts, P., Goldie, C.M. & Veraverbeke, N. (1979). Subexponentiality and infinite divisibility, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **49**, 335–347.
- [32] Embrechts, P. & Goldie, C.M. (1982). On convolution tails, *Stochastic Processes and their Applications* **13**, 263–278.
- [33] Cline, D.B.H. (1987). Convolution of distributions with exponential and subexponential tails, *Journal of the Australian Mathematical Society, Series A* **43**, 347–365.
- [34] Goldie, C.M. & Resnick, S. (1988). Distributions that are both subexponential and in the domain of attraction of an extreme-value distribution, *Advances in Applied Probability* **20**, 706–718.
- [35] Klüppelberg, K. & Kostadinova, T. (2007). Integrated insurance risk models with exponential Lévy investment, *Insurance Mathematics and Economics*, to appear.
- [36] Protter, P.E. (2004). *Stochastic Integration and Differential Equations*, 2nd Edition, Springer, Berlin.
- [37] Frolova, A., Kabanov, Y. & Pergamenschikov, S. (2002). In the insurance business risky investments are dangerous, *Finance and Stochastics* **6**, 227–235.
- [38] Wang, G. & Wu, R. (2001). Distributions for the risk process with a stochastic return on investment, *Stochastic Processes and their Applications* **95**, 329–341.
- [39] Schmidli, H. (2005). On optimal investment and subexponential claims, *Insurance Mathematics and Economics* **36**, 25–35.
- [40] Gaier, J. & Grandits, P. (2004). Ruin probabilities and investment under interest force in the presence of regularly varying tails, *Scandinavian Actuarial Journal* **104**, 256–278.
- [41] Paulsen, J. (1998). Ruin with compounding assets—a survey, *Insurance Mathematics and Economics* **22**, 3–16.
- [42] Lindner, A. & Maller, R. (2005). Lévy integrals and the stationarity of generalised Ornstein-Uhlenbeck processes, *Stochastic Processes and their Applications* **115**, 1701–1722.
- [43] Nyrhinen, H. (2001). Finite and infinite time ruin probabilities in a stochastic economic environment, *Stochastic Processes and their Applications* **92**, 265–285.
- [44] Goldie, C.M. (1991). Implicit renewal theory and tails of solutions of random equations, *Annals of Applied Probability* **1**, 126–166.
- [45] Paulsen, J. (2002). On Cramér asymptotics for risk processes with stochastic return on investments, *Annals of Applied Probability* **12**, 1247–1260.
- [46] Kalashnikov, V. & Norberg, R. (2002). Power tailed ruin probabilities in the presence of risky investments, *Stochastic Processes and their Applications* **98**, 211–228.
- [47] Kostadinova, T. (2007). Optimal investment for insurers, when stock prices follows an exponential Lévy process, *Insurance Mathematics and Economics* **41**, 250–263.
- [48] Brokate, M., Klüppelberg, C., Kostadinova, R., Maller, R. & Seydel, R.S. (2007). On the distribution tail of an integrated risk model: a numerical approach, *Insurance Mathematics and Economics*, to appear.

VICKY FASEN AND CLAUDIA KLÜPPELBERG

Latin Hypercube Sampling

Background

Latin hypercube sampling (LHS) has gained widespread use since its development by W. J. Conover of Texas Tech University in the summer of 1975 while he was serving as a consultant to Los Alamos National Laboratory. LHS has been used worldwide in such computer modeling applications as hurricane loss projection modeling in Florida [1–4]; risk assessment for geologic isolation of radioactive waste at Yucca Mountain in Nevada [5, 6] and the WIPP site in New Mexico [7–9]; and safety assessments for nuclear power plants throughout the world [10, 11]. It is also used in computer modeling for reliability analyses for manufacturing equipment, particularly in optimization schemes for repairable equipment [12]. Other applications include the petroleum industry [13]; transmission of HIV [14]; and subsurface stormflow modeling [15]. Iman and Conover [16] present a lengthy and detailed application of LHS to risk assessment. Surveys of sensitivity analysis procedures that can be used in conjunction with LHS can be found in Kleijnen and Helton [17], Helton and Davis [18], and Helton *et al.* [19]. In addition, a Google search for LHS reveals a multitude of wide ranging applications.

An overview article on LHS was given by Iman [20]. Helton and Davis [21] present 325 references for the use of LHS in the propagation of uncertainty in complex systems. Software to implement the LHS strategy was developed in 1975 by R.L. Iman. This software was refined and was first released formally in 1980 [22]. A later revision [23] has been the most widely distributed mainframe version of the program. Commercial vendors of LHS software include @RISK® and Crystal Ball®. The first journal article on LHS appeared in *Technometrics* [24].

Latin Hypercube Sampling

In the summer of 1975, Conover was asked by Los Alamos National Laboratory to develop a method for improving the efficiency of simple Monte Carlo used to characterize the uncertainty in inputs to

computer models. Conover's work (Conover [25, 26, Appendix A]) resulted in the development of a stratified Monte Carlo sampling scheme to improve on the coverage of the input space. The stratification is accomplished by dividing the vertical axis on the graph of the distribution function of a random variable X_j into n nonoverlapping intervals of equal length, where n is the number of computer runs to be made. Through the inverse function $F^{-1}(x)$, these n intervals divide the horizontal axis into n equiprobable, but not necessarily equal length, intervals. Thus, the x axis has been stratified into n equiprobable and nonoverlapping intervals. The next step in the LHS scheme requires the random selection of a value within each of the n intervals on the vertical axis. When these values are mapped through $F^{-1}(x)$, exactly one value will be selected from each of the intervals previously defined on the horizontal axis. Let $\mathbf{X}$ be an $n \times k$ matrix whose j th column contains the LHS for the random variable X_j . A random process must be used to ensure a random ordering of the values within each column of this matrix. This mixing process serves to emulate the pairing of observations in a simple Monte Carlo process.

To fully appreciate the value of the underlying structure of LHS, it is helpful to be familiar with computer models used in actual applications. Such models are usually characterized by a large number of input variables (perhaps as many as a few hundred) and usually only a handful of these inputs are important for a given response. In addition, the model response is frequently multivariate and time dependent. If the input values were based on a factorial design, each level of each factor would be repeated many times. Moreover, the experimenter usually has a particular response in mind when constructing the factorial design and this design may be totally ineffective with multiple responses. On the other hand, LHS ensures that the entire range of each input variable is completely covered without regard to which single variable or combination of variables might dominate the computer model response(s). This means that a single sample will provide useful information when some input variable(s) dominate certain responses (or certain time intervals), while other input variables dominate other responses (or time intervals). By sampling over the entire range, each variable has the opportunity to show up as important, if it is indeed important. If an input

variable is not important, then the method of sampling is of little or no concern. The efficiency of LHS relative to simple random sampling is considered by Iman [20] for a range of conditions.

Many computer models used for risk assessment and other applications require correlated input values. A technique for inducing rank correlation structures among input variables for both simple Monte Carlo and LHS was given by Iman and Conover [27]. Sallaberry *et al.* [28] consider an extension of LHS with correlated variables.

Application to Hurricane Loss Projection Methodology

Iman *et al.* [29] provide a detailed overview of problems associated with projecting losses associated with hurricanes, which is a complex and difficult undertaking that is fraught with uncertainties. Hurricane Charley, which struck southwest Florida on August 13, 2004, illustrates the uncertainty of forecasting damages from these storms. Owing to shifts in the track and the rapid intensification of the storm, real time estimates grew from \$2 to \$3 billion in losses late on the 12th, to a peak of \$50 billion for a brief time, as the storm appeared to be headed for the Tampa Bay area. The storm struck the resort areas of Charlotte Harbor and moved across the densely populated central part of the state, with early poststorm estimates in the \$28 to \$31 billion dollar range, and final estimates converging at \$15 billion as the actual intensity at landfall became apparent. The Florida Commission on Hurricane Loss Projection Methodology (FCHLPM) has a great appreciation for the role of computer models in projecting losses from

hurricanes. The FCHLPM contracts with a professional team to perform on-site (confidential) audits of computer models developed by several different companies in the United States who seek to have their models approved for use in insurance rate filings in Florida. The team's members represent the fields of actuarial science, computer science, meteorology, statistics, and wind and structural engineering. An important part of the auditing process requires uncertainty analyses (UA) and sensitivity analyses (SA) to be performed with the applicant's proprietary model. LHS is an integral part of these analyses.

The approach to both UA and SA is based on sampling the range of uncertainty associated with some key computer input variables and then using the LHS values (correlated and uncorrelated) with the computer model to produce wind speed and subsequent loss. Figure 1 shows a schematic representation of the input-output process for generating wind velocity and loss for k input variables based on an LHS of size n with outputs computed for t hours.

The goal of SA is to determine which of the k inputs in Figure 1 influence both wind velocity at time t at a given location (latitude, longitude) and cause maximum loss cost. The general goal of UA is to determine which of the k inputs contribute to the uncertainty in the same output variables at time t at a given location.

Iman *et al.* [3, 4] give a detailed SA and UA for a Holland [30] wind model, and a cubic damage function [31] is used to convert wind speed to percent damage. Those analyses are now briefly summarized.

The Holland model was used to calculate wind speed (WS) for Category 1 ($74 \leq WS < 96$ mph); Category 3 ($110 \leq WS < 130$ mph); and Category

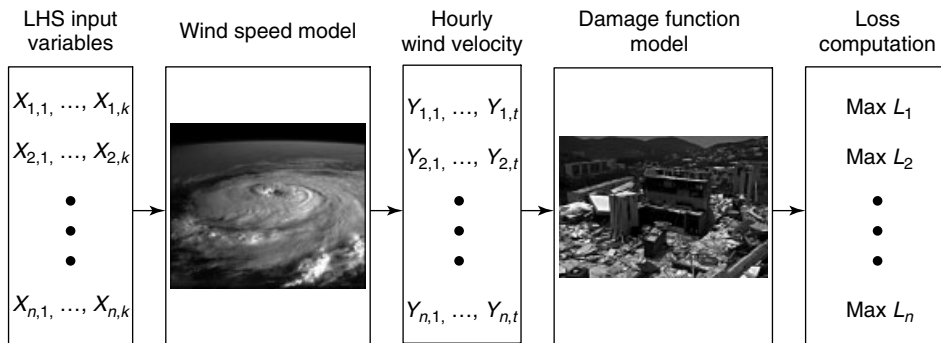


Figure 1 Schematic representation of the input-output process for UA/SA

5 ($WS \geq 155$ mph) hurricanes. Wind speeds were calculated hourly for 12 h at each vertex in a 21×46 grid (966 vertices) that covered a portion of the path of each simulated hurricane along a $60 \text{ mi} \times 135 \text{ mi}$ area crossing southern Florida.

The following four input variables were selected for this analysis ($k = 4$ in the first box of Figure 1):

- central pressure (CP) – the barometric pressure (mb) at the center of the storm (in general, the lower the CP compared to the far field pressure, the stronger the storm in terms of wind speed);
- radius of maximum winds ($R_{\max}$) – the radius (mi) from the center of the storm to the point of maximum winds (the eye wall) surrounding the storm;
- forward speed (V_T) – speed (mph) at which the storm is moving along the earth’s surface (not the speed at which winds are circulating around the storm center);
- Holland B parameter (B) – a parameter in the wind field model that defines the shape of the wind field, and is dependent on the maximum wind speed and the difference between environmental pressure and CP as explained in the next section.

Each of these input characteristics (variables) has an associated degree of uncertainty for a specific storm, which were described by triangular distributions. CP and $R_{\max}$ are likely to have some degree of correlation or linear relationship, particularly for a Category 5 hurricane.

Standardized regression coefficients (SRC) provide a very convenient metric for presenting SA results. SRCs are easily calculated from the inverse of the simple correlation matrix for the four input variables and the wind speeds produced by the Holland wind field at a given set of grid coordinates for a fixed time and a given storm category. These can also be calculated for the maximum loss at each grid point. Iman, Johnson and Schroeder [21] provide details of the SRC calculations.

Graphical displays of SRCs are very useful in interpreting the results of the SA. Figure 2 shows how the influence of $R_{\max}$ changes over time at a fixed point in the grid as the eye of Category 1, 3, and 5 hurricanes pass this point at $t = 3$ h. Figure 3 shows how $R_{\max}$ influences the maximum wind speed over the entire $60 \text{ mi} \times 135 \text{ mi}$ grid, ranging

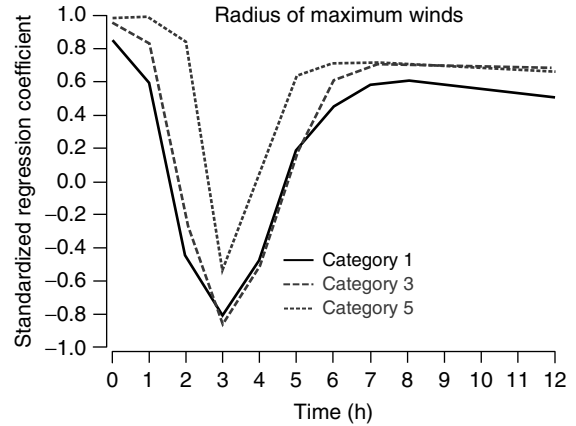


Figure 2 SRCs for $R_{\max}$ at a single grid point over time

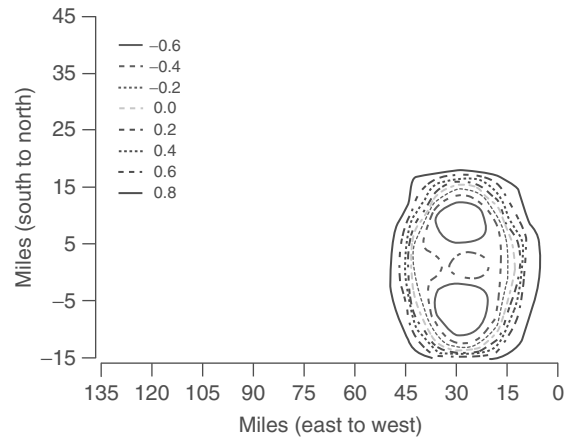


Figure 3 SRCs for $R_{\max}$ for a category 5 at $t = 2$ h over the entire grid

from a negative influence inside the eye wall to a positive influence outside the eye wall.

As previously stated, the goal of uncertainty analysis is to quantify the contributions of the input variables to the uncertainty in total wind speed and maximum loss. These contributions are referred to as expected percentage reduction (EPR) in the variance of the output. There are several steps in the calculation of these contributions, which involve the processing of additional LHSs. Complete details of the calculation and an illustrated example with a simple analytic model can be found in Iman *et al.* [2]. The same reference also provides a detailed step-by-step illustration of how to estimate the EPR in

the variance of the output variable(s) based on actual computer input and output. The calculation of EPR is based on a method given by Iman [32], referred to as *uncertainty importance*.

As with SA, graphical displays are useful for conveying the results of UA. Figures 4 and 5 are the respective analogs of Figures 2 and 3. Figure 4 shows how the EPR owing to $R_{\max}$ changes over time at a fixed point in the grid as the eye of Category 1, 3, and 5 hurricanes pass this point at $t = 3$ h. Figure 3 shows how $R_{\max}$ influences the variation in the maximum wind speed over the entire $60 \text{ mi} \times 135 \text{ mi}$ grid, ranging from a small contribution inside the eye wall to a very strong contribution at and near the eye wall as one would expect.

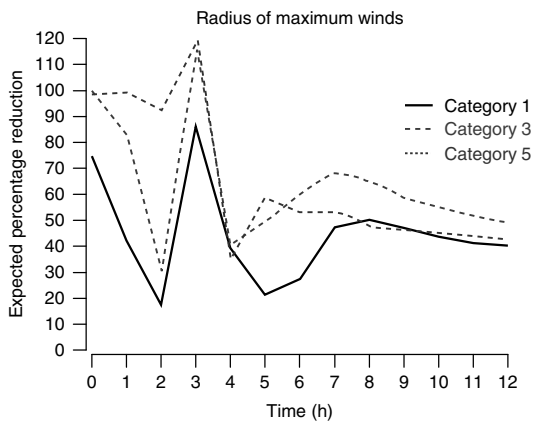


Figure 4 EPRs for $R_{\max}$ at a single grid point over time

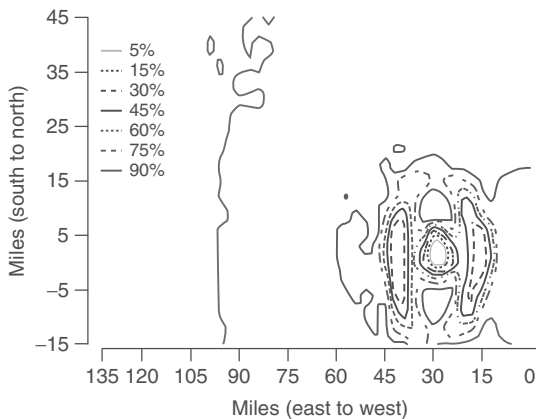


Figure 5 EPRs for $R_{\max}$ for a category 5 at $t = 2$ h over the entire grid

The interested reader can check out the references given in this article to gain a more complete understanding of how LHS provides for efficient and effective uses of what are frequently very time consuming computer model calculations and also sets up the subsequent uncertainty and sensitivity analyses.

References

- [1] Iman, R.L., Johnson, M.E. & Schroeder, T.A. (2002). Assessing hurricane effects: part 1-sensitivity analysis, *Reliability Engineering and Systems Safety* **78**(2), 36–50.
- [2] Iman, R.L., Johnson, M.E. & Schroeder, T.A. (2002). Assessing hurricane effects: part 2 - uncertainty analysis, *Reliability Engineering and Systems Safety* **78**(2), 51–59.
- [3] Iman, R.L., Johnson, M.E. & Watson, C.C. (2005). Sensitivity analysis for computer model projections of hurricane losses, *Risk Analysis* **25**(5), 1277–1297.
- [4] Iman, R.L., Johnson, M.E. & Watson, C.C. (2005). Uncertainty analysis for computer model projections of hurricane losses, *Risk Analysis* **25**(5), 1299–1312.
- [5] CRWMS M&O (Civilian Radioactive Waste Management System Management and Operating Contractor) (2000). *Total System Performance Assessment for the Site Recommendation*, NV TDR-WIS-PA-000001 REV 00, Las Vegas.
- [6] U.S. DOE (U.S. Department of Energy) (2002). *Yucca Mountain Science and Engineering Report*, NV DOE/RW-0539-1, Rev. 1, Las Vegas.
- [7] Campbell, J.E. & Longsine, D.E. (1990). Application of generic risk assessment software to radioactive waste disposal, *Reliability and Systems Safety* **30**, 183–193.
- [8] Helton, J.C. (1999). Uncertainty and sensitivity analysis in performance assessment for the waste isolation pilot plant, *Computer Physics Communications* **117**, 156–180.
- [9] Helton, J.C., Anderson, D.R., Jow, H.-N., Marietta, M.G. & Basabilvazo, G. (1999). Performance assessment in support of the 1996 compliance certification application for the waste isolation pilot plant, *Risk Analysis* **19**, 959–986.
- [10] Breeding, R.J., Helton, J.C., Gorham, E.D. & Harper, F.T. (1992). Summary description of the methods used in the probabilistic risk assessments for NUREG-1150, *Nuclear Engineering and Design* **135**, 1–27.
- [11] Helton, J.C. & Breeding, R.J. (1993). Calculation of reactor accident safety goals, *Reliability Engineering and System Safety* **39**, 129–158.
- [12] Painton, L.A. & Campbell, J.E. (1995). Genetic algorithms in optimization of system reliability, *IEEE Transactions on Reliability* **44**(2), 172–178.
- [13] MacDonald, R.C. & Campbell, J.E. (1986). Hydrocarbon economic risk analysis, *Journal of Petroleum Technology* **38**, 57–69.

- [14] Blower, S.M. & Dowlatbadi, H. (1994). Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example, *International Statistical Review* **62**, 229–243.
- [15] Gwo, J.P., Toran, L.E., Morris, M.D. & Wilson, G.V. (1996). Subsurface stormflow modeling with sensitivity analysis using a Latin hypercube sampling technique, *Ground Water* **34**, 811–818.
- [16] Iman, R.L. & Conover, W.J. (1980). Small sample sensitivity analysis techniques for computer models, with an application to risk assessment, *Communications in Statistics* **A9**(17), 1749–1842, Rejoinder to Comments, 1863–1874.
- [17] Kleijnen, J.P.C. & Helton, J.C. (1999). Statistical analyses of scatterplots to identify important factors in large-scale simulations, 1: review and comparison of techniques, *Reliability Engineering and System Safety* **65**, 147–185.
- [18] Helton, J.C. & Davis, F.J. (2002). Illustration of sampling-based methods for uncertainty and sensitivity analysis, *Risk Analysis* **22**, 591–622.
- [19] Helton, J.C., Johnson, J.D., Sallaberry, C.J. & Storlie, C.B. (2006). Survey of sampling-based methods for uncertainty and sensitivity analysis, *Reliability Engineering and System Safety* **91**, 1175–1209.
- [20] Iman, R.L. (1999). Latin hypercube sampling, *Encyclopedia of Statistical Sciences*, Update Volume 3, John Wiley & Sons, New York.
- [21] Helton, J.C. & Davis, F.J. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, *Reliability Engineering and System Safety* **81**, 23–69.
- [22] Iman, R.L., Davenport, J.M. & Zeigler, D.K. (1980). Latin hypercube sampling (program user's guide), Technical Report SAND79-1473, Sandia National Laboratories, Albuquerque.
- [23] Iman, R.L. & Shortencarier, M.J. (1984). A FORTRAN 77 program and user's guide for the generation of Latin hypercube and random samples for use with computer models, Technical Report NUREG/CR-3624, SAND83-2365, Sandia National Laboratories, Albuquerque.
- [24] McKay, M.D., Conover, W.J. & Beckman, R.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics* **21**(2), 239–245.
- [25] Conover, W.J. (1975). *On a Better Method of Selecting Values of Input Variables for Computer Codes*, Unpublished manuscript (see Appendix A of [11]).
- [26] Helton, J.C. & Davis, F.J. (2002). *Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems*, SAND2001-0417, App A, B, Sandia National Laboratories, Albuquerque.
- [27] Iman, R.L. & Conover, W.J. (1982). A distribution-free approach to inducing rank correlation among input variables, *Communications in Statistics* **B11**(3), 311–334.
- [28] Sallaberry, C.J., Helton, J.C. & Hora, S.C. (2006). *Extension of latin hypercube samples with correlated variables*, Technical Report SAND2006-6135, Sandia National Laboratories, Albuquerque.
- [29] Iman, R.L., Johnson, M.E. & Watson, C.C. (2006). Statistical aspects of forecasting and planning for hurricanes, Invited paper, *The American Statistician* **60**(2), 105–121.
- [30] Holland, G.J. (1980). An analytic model of the wind and pressure profiles in hurricanes, *Monthly Weather Review* **108**(8), 1212–1218.
- [31] Howard, R.A., Matheson, J. & North, W. (1972). The decision to seed hurricanes, *Science* **176**, 1191–1201.
- [32] Iman, R.L. (1987). A matrix-based approach to uncertainty and sensitivity analysis for fault trees, *Risk Analysis* **7**(1), 21–33.

RONALD L. IMAN

Lead

Background

As with all metals, there are three forms of lead: elemental, inorganic, and organic. Elemental lead is a dense, blue-gray, soft metal that occurs naturally. Its softness, ability to resist corrosion, high density, and low melting point make it useful for plumbing, solder, batteries, weights, and as a shield for radiation. The inorganic form of lead is the most common; it is mined from the earth as an ore. Among many other uses, lead was used as a pigment in paint and ceramics; it is also used in ammunition. The organic forms of lead—mainly tetraethyl and tetramethyl—were used as antiknock agents^a in gasoline, beginning in the 1920s. Most countries have since phased it out, but it is still commonly used in aviation fuel. In the past three centuries, the amount of lead in the environment has increased over 1000-fold because of human activity. However, most uses of lead are being reduced owing to the adverse health effects associated with exposure.

Purpose

This article summarizes the unique aspects of lead that must be considered during the risk assessment process and discusses the wealth of data that exists for each of the four risk assessment steps. Since the majority of exposure today is to inorganic lead, that is the emphasis of this article. Also, since there is extensive data on the effects of lead exposure in humans, we do not address animal studies. Finally, the article discusses the United States Occupational Safety and Health Administration (OSHA)'s current risk management policies.

Special Considerations

Lead has several characteristics that require careful attention in considering its human health effects. First, lead exposure patterns, toxicokinetics, and health effects differ between children and adults. Second, lead has well-validated biomarkers of exposure, which must be interpreted carefully with regard

to what they represent. This is because interpreting associations of health effects with measures of recent dose differs substantially from that of cumulative dose. More specifically, health effects associated with recent dose may be more likely to be acute and reversible, while those associated with cumulative dose may be more likely to be chronic and irreversible. Third, occupational exposures differ from environmental exposures as they are usually of higher intensity, occur over an 8-h period (*versus* a 24-h period), mainly occur in adults within a narrow age range (e.g., 18–60 years), and generally only occur in healthy persons, as unhealthy individuals are often selected out of the workplace (i.e., the healthy worker effect). Further discussion of these points follows.

Children have a greater risk of environmental exposure to lead owing to their proximity to the ground, hand-to-mouth activity, increased time outdoors, pica behavior, the sweet taste associated with lead paint, and their inability to judge risk. Additionally, per kilogram of body weight, children take in more fluids, food, and air compared to adults. Absorption of lead through the gastrointestinal tract is higher in children and it more readily crosses their blood-brain barrier [1]. Furthermore, children exhibit more severe health effects at lower doses than adults. Lead also crosses the placenta; although some of the data are mixed, children born to lead-exposed mothers may be born early, have decreased birth weight, cognitive sequelae, or birth defects.

There are two well-validated biomarkers of lead exposure: lead in whole blood and lead in bone. Other markers, used less frequently and with less supportive validation evidence, include tooth lead, urinary lead, chelatable lead (after dimercaptosuccinic acid or ethylene diamine tetraacetic acid, and hair or nail lead. Blood lead levels are the most commonly obtained estimate of dose, probably due to cost, convenience, and the fact that most occupational standards utilize it extensively. Lead has a clearance half-time of approximately 30 days from blood, and thus reflects integrated exposure from external and internal sources, over approximately 120 days; a single measurement, for example, cannot distinguish between a high-level recent exposure and lower-level long-term exposures. Serial measurements of blood lead give a better picture of lead exposure over time and their integral (i.e., area-under-the-curve of blood

lead *versus* time) correlates well with tibia lead, a measure of cumulative dose.

The characteristics of the biomarkers discussed above are important since the health effects associated with recent dose differ from those associated with cumulative dose. High-level, short-term exposures can result in lead's direct action on cellular receptors, other proteins, and calcium agonism or antagonism, leading to death, encephalopathy, cognitive dysfunction, and transient hypertension. If patients are removed from the source of exposure, some of these effects are reversible. In contrast, the chronic effects of cumulative dose are associated with more permanent changes in cellular architecture and synaptic density, leading to persistent or progressive health effects, such as chronic hypertension, cognitive decline that continues even after cessation of exposure, and increased risk of death because of cardiovascular outcomes.

Existing Risk Assessments

Several existing assessments estimate the risk of adverse health effects due to lead exposure (*see Environmental Hazard; Environmental Health Risk*). For example, in the United States, the Agency for Toxic Substances and Disease Registry (ATSDR) and the Environmental Protection Agency (EPA), as well as state governments, perform risk assessments at specific sites suspected to contain high levels of lead, such as superfund sites or homes built before the ban on lead-based paint [2–4]. Additionally, every 3 years the EPA releases an assessment of health risks from toxic air pollutants, including lead. This assessment is based on emissions, and uses models to describe ambient levels, human exposure, and risk [5]. With regard to occupational exposure, the National Institute for Occupational Safety and Health (NIOSH) performs health hazard evaluations of lead in specific workplaces, as requested [6].

Hazard Identification

Lead exposure at high doses can cause death. Although this is rarely encountered now, reports at the turn of the last century estimated hundreds of deaths per year because of childhood or occupational lead poisoning. Lead is a nonspecific toxicant and acts throughout the body, not only to antagonize

calcium, thus interfering with basic biochemical processes such as cell messaging, but also to bind with proteins, thus altering protein structure and inhibiting enzyme function. Because of this nonspecificity, lead affects many organ systems, including the neurological, cardiovascular, circulatory, gastrointestinal, hematopoietic, musculoskeletal, and renal systems. Some of the main health effects are discussed below. For all health outcomes, an increasing number of studies over the past decade have documented health effects associated with tibia lead, suggesting that lead separately causes acute health effects as a function of recent dose, and chronic health effects as a function of cumulative dose.

Neurological effects include both central nervous system and peripheral nervous system impairments. High doses of lead are associated with encephalopathy in both adults and children. In adults, central nervous system symptoms include anxiety, depression, dizziness, fatigue, forgetfulness, headache, impotence, irritability, lethargy, malaise, nervousness, and weakness. Adult neurobehavioral test results show dose–response relationships across many cognitive domains, with some of the most consistent and strongest relationships occurring on tests of manual dexterity, executive abilities, and verbal memory and learning. Peripheral nervous system symptoms include paresthesias and weakness; some authors consider lead to have a larger influence on peripheral motor function than sensory function. Testing shows impaired nerve conduction velocities and postural balance. Additionally, epidemiologic evidence exists for a link between lead exposure and development of both amyotrophic lateral sclerosis and Parkinson's disease; biochemical evidence suggests lead may interact with both amyloid and tau protein, providing a possible link to Alzheimer's disease.

In children, lead dose is consistently associated with impaired intellectual function as measured by intelligence (IQ) tests in epidemiologic studies in several countries. Central nervous system impairments include decreased attention, memory, nonverbal reasoning ability, and visual-motor performance. Peripheral neuropathy is demonstrated *via* impaired fine motor skills and decreased nerve conduction velocities. Behaviorally, lead-exposed children are more likely to display antisocial behaviors such as aggression, delinquency, and marijuana use.

Occupational epidemiological studies in adults comprise the majority of studies on renal effects

and most use blood lead levels as their biomarker, although an increasing number measure tibia lead. Lead causes well-documented and consistent adverse effects in the renal system including proximal tubular nephropathy, glomerular sclerosis, and interstitial fibrosis. Impairments in function include depressed glomerular filtration rate, enzymuria, impaired transport of organic anions and glucose, and proteinuria.

Cardiovascular effects include increased risk of mortality from cardiovascular disease, coronary heart disease, and stroke. Well-documented data from numerous epidemiologic studies, both occupational and environmental, demonstrate a consistently increased prevalence of both hypertension and peripheral arterial disease. Intermediate effects include changes in cardiac conduction or rhythm as well as changes in left ventricular function or mass. These effects have been more studied in adults, but some studies indicate that exposure in childhood might be associated with the development of hypertension in adulthood. Although most studies evaluated associations with blood lead levels, bone lead levels are increasingly recognized as a better predictor of hypertension, implying that cumulative dose is the more relevant metric.

Some epidemiological studies associate lead exposure with reproductive effects in adults. Moderately high blood lead levels appear to be linked to spontaneous abortion and premature delivery in pregnant women, while men may exhibit decreased fertility and changes in sperm. Most of these studies measured blood lead levels in occupationally exposed workers, or people living near a lead-related industry. Studies of *in utero* effects measuring lead in maternal or child blood found neurodevelopmental effects, as well as decreased height and weight gain rates.

Most studies of the relationship between lead and cancer are occupational epidemiology studies of blood lead levels (*see Occupational Cohort Studies*). Many of the studies do not control for concurrent risk factors such as smoking, arsenic exposure, or dietary habits. However, there are fairly consistent associations between high lead levels and cancer, specifically cancers of the digestive and respiratory system. EPA classifies inorganic lead as a probable human carcinogen, based on inadequate evidence in humans and sufficient evidence in animals. The US Department of Health and Human Services states that lead and lead compounds are reasonably anticipated to be human carcinogens. The International Agency

for Research on Cancer has determined that inorganic lead is probably carcinogenic to humans.

Dose–Response Relationship

For adults, cardiovascular and neurological outcomes comprise the most sensitive health endpoints associated with lead exposure. Researchers have documented acute neurological effects at blood lead levels over $20\text{ }\mu\text{g dl}^{-1}$. Recent evidence indicates that cardiovascular changes occur at even lower levels of exposure. Lead levels above $5\text{ }\mu\text{g dl}^{-1}$ are associated with increased risk of coronary heart disease, peripheral arterial disease, and elevations in blood pressure. Analyses indicate that for each doubling of blood lead levels there is a 1 mmHg increase in systolic blood pressure and a 0.6 mmHg increase in diastolic blood pressure, although there is some inconsistency in effect estimates across studies (*see Dose–Response Analysis; Threshold Models*).

In children, neurological effects are considered the most sensitive endpoint. Neurodevelopmental effects are consistently seen below $10\text{ }\mu\text{g dl}^{-1}$; most researchers agree that there is no evidence for a threshold below which adverse effects are not expected to occur. Analyses indicate that for every $1\text{ }\mu\text{g dl}^{-1}$ increase in blood lead there is an approximate decrease in IQ of 0.9 points. The relation between blood lead and IQ may not be linear; an increasing body of evidence suggests the adverse slope is steeper below $10\text{ }\mu\text{g dl}^{-1}$.

Certain factors, such as age and genetics, may make individuals more susceptible to the effects of lead exposure, giving the dose–response curve a steeper slope or shifting it to the left.

On the basis of epidemiologic evidence in 1991, and with the goal of being protective of the most sensitive subpopulation (children), the Centers for Disease Control and Prevention (CDC) determined that blood lead concentrations above $10\text{ }\mu\text{g dl}^{-1}$ present a risk to human health. In the past, this value was as high as $60\text{ }\mu\text{g dl}^{-1}$, but CDC decreased it repeatedly [7]. Some scientists continue to push for CDC to lower it even further. Neither ATSDR nor EPA have determined levels of external exposure below which adverse health effects are not expected to occur in children.

Exposure Assessment

Numerous monitoring programs in the United States track blood lead levels in the population. CDC's National Health and Nutrition Examination Survey (NHANES) analyzes blood lead data in citizens across the United States. In 2001–2002, the geometric mean blood levels in the population was between 1.5 and 1.6 $\mu\text{g dl}^{-1}$. NIOSH's Adult Blood Lead Epidemiology and Surveillance Program collects data from participating states, which require that all laboratories report adult blood lead levels that exceed 25 $\mu\text{g dl}^{-1}$ to the state department of health. Finally, CDC's Childhood Blood Lead Surveillance System is similar to NIOSH's program, but reports only on children less than 72-months old.

To estimate current and future exposures, it is also necessary to measure the levels of lead in the environment. EPA uses two models that relate environmental lead levels to predicted blood lead levels: (a) the integrated exposure uptake biokinetic model for children, and (b) the adult lead methodology for adults. These models take into account lead's multiple exposure pathways (ingestion and inhalation) as well its multiple exposure media (air, dust, food, soil, and water). The models allow scientists to use measures of external exposure and knowledge about the toxicokinetics of lead to estimate total current external exposure as well as to predict lifetime dose [8].

Risk Characterization

The goal of this step is to determine the risk of having an exposure above the "safe" level in the population. Although many scientists believe this level should be lower, the US government generally considers 10 $\mu\text{g dl}^{-1}$ to be safe for all populations exposed nonoccupationally, including children. OSHA's standards, on the other hand, use a blood lead level of 40 $\mu\text{g dl}^{-1}$ across the working lifetime [9]. This is higher than CDC's level owing to the fact that OSHA must balance health with feasibility concerns.

According to CDC, in 2006, there were approximately 310 000 children with blood lead levels above 10 $\mu\text{g dl}^{-1}$ living in the US. With regard to adults, 35 states in 2002 reported over 10 500 adults with blood lead levels of 25 $\mu\text{g dl}^{-1}$ or greater. Almost 1900 of these possessed levels of 40 $\mu\text{g dl}^{-1}$ or greater.

It should be noted that many studies have now documented health effects at much lower levels than

these. Additionally, these values are not indicative of lifetime dose, nor do they necessarily reflect peak exposure levels during some of the critical windows of vulnerability, such as during development. Finally, these statistics may not reflect high exposures occurring in the small population of people who work with lead or who live near lead-contaminated sites.

Occupational Risk Management in the United States

OSHA's standard for lead in general industry was finalized in 1978 and allows workers to attain blood lead levels up to 40 $\mu\text{g dl}^{-1}$ over their working lifetimes. Many scientists now believe, based on an extensive body of new evidence, that OSHA's lead standards should protect workers both from the acute effects of recent dose and also from the chronic effects of cumulative dose, and that allowing blood lead levels up to 40 $\mu\text{g dl}^{-1}$ does not accomplish either goal. Given evidence for chronic health effects at doses lower than the current standard, the ability to measure cumulative dose, and the availability of technology to limit exposure, some scientists believe a reassessment of the lead standards is overdue [10, 11].

End Notes

^a. Antiknock agents are gasoline additives used to reduce engine knocking and increase the fuel's octane rating.

References

- [1] ATSDR (2005). *Toxicological Profile for Lead: Draft for Public Comment*, US Department of Health and Human Services, Public Health Service, Atlanta.
- [2] ATSDR (2006). *Public Health Assessments & Health Consultations*, at <http://www.atsdr.cdc.gov/hac/phal/index.html> (accessed Nov 2006).
- [3] EPA (2006). *Superfund Risk Assessment*, at http://www.epa.gov/oswer/riskassessment/risk_superfund.htm (accessed Nov 2006).
- [4] Abadin, H.G. & Wheeler, J.S. (1994). Guidance for risk assessment of exposure to lead: a site-specific, multi-media approach, in *Hazardous Waste and Public Health: International Congress on the Health Effects of Hazardous Waste*, J.S. Andrews Jr, H. Frumkin, B.L. Johnson, M.A. Mehlman, C. Xintaras & J.A.

- Bucseila, eds, Princeton Scientific Publishing Company, Princeton, pp. 477–485.
- [5] EPA (2006). *The National-Scale Air Toxics Assessment*, at <http://www.epa.gov/ttn/atw/natamain/> (accessed Nov 2006).
- [6] NIOSH (2006). *Health Hazard Evaluations*, <http://www.cdc.gov/niosh/hhe/> (accessed Nov 2006).
- [7] CDC (1991). *Preventing Lead Poisoning in Young Children: A Statement by the Centers for Disease Control and Prevention*, at <http://www.cdc.gov/nceh/lead/publications/books/plpyc/contents.htm> (accessed Nov 2006).
- [8] National Academies of Science (2005). *Superfund and Mining Megaprojects: Lessons from the Coeur d'Alene River Basin*, National Research Council, Washington, DC, at <http://www.nap.edu/openbook/0309097142/html/index.html> (accessed Nov 2006).
- [9] OSHA (1978). *Table of Contents for Lead*, at [http://www.osha.gov/pls/oshaweb/owasrch.search_form?p_doc_type=PREAMBLES&p_toc_level=1&p_keyvalue=Lead~\(November~1978\)](http://www.osha.gov/pls/oshaweb/owasrch.search_form?p_doc_type=PREAMBLES&p_toc_level=1&p_keyvalue=Lead~(November~1978)) (accessed Nov 2006).
- [10] Schwartz, B.S. & Hu, H. (2007). Adult lead exposure: time for change – introduction to the minimonograph. *Environmental Health Perspectives* in press **115**, 451–454 .
- [11] Silbergeld, E., Landrigan, P.J., Froines, J.R. & Pfeffer, R.M. (1997). The occupational lead standard: a goal unachieved, a process in need of repair, in *Work, Health, and Environment: Old Problems, New Solutions*, C. Levenstein & J. Wooding, eds, The Guilford Press, New York, pp. 131–148.

Related Articles

Occupational Risks

What are Hazardous Materials?

MEGAN WEIL LATSHAW AND BRIAN
S. SCHWARTZ

Lévy Processes in Asset Pricing

The main empirical motivation for using Lévy processes in finance comes from fitting asset return distributions. Consider the daily (either continuous or simple) returns of Standard & Poor's 500 index (SPX) from January 2, 1980 to December 31, 2005. We plot the histogram of normalized (mean zero and variance one) daily simple returns in Figure 1, along with the standard normal density function. The maximum and minimum (which all occurred in 1987) of the normalized daily returns are about 7.9967 and -21.1550 . Note that, for a standard normal random variable Z , $P(Z < -21.1550) \approx 1.4 \times 10^{-107}$; as a comparison the whole universe is believed to have existed for 15 billion years or 5×10^{17} seconds.

Clearly the histogram of SPX displays a high peak and two asymmetric heavy tails. This is not only true for SPX but also for almost all financial prices, e.g., worldwide stock indices, individual stocks, foreign exchange rates, interest rates. In fact it is so evident that a name “leptokurtic distribution” is given, which means the kurtosis of the distribution is large. More precisely, the kurtosis and skewness are defined as $K = E\left(\frac{(X-\mu)^4}{\sigma^4}\right)$, $S = E\left(\frac{(X-\mu)^3}{\sigma^3}\right)$. For the standard normal density $K = 3$, and if $K > 3$, the distribution is called *leptokurtic* and the distribution will have a higher peak and two heavier tails than that of the normal distribution. For the SPX data, the sample kurtosis is about 42.23. The skewness is about -1.73 ; the negative skewness indicates that the left tail is heavier than the right tail. The classical geometric Brownian motion model (see **Simulation in Risk Management; Volatility Smile; Weather Derivatives**), which models the stock price as $S(t) = S(0)e^{\mu t + \sigma W_t}$ with W_t being the standard Brownian motion (see **Large Insurance Losses Distributions; Risk-Neutral Pricing; Importance and Relevance; Default Correlation**), is inconsistent with this feature, because in this model the return, $\ln(S(t)/S(0))$, has a normal distribution. Lévy processes, among other processes, have been proposed to incorporate the leptokurtic feature.

Overview

A stochastic process X_t is a Lévy process if it has independent and stationary increments and a stochastically continuous sample path, i.e., for any $\varepsilon > 0$, $\lim_{h \downarrow 0} P(|X_{t+h} - X_t| > \varepsilon) = 0$. Therefore, Lévy processes provide a natural generalization of the sum of independent and identically distributed (i.i.d.) random variables. The simplest possible Lévy processes are the standard Brownian motion $W(t)$, Poisson processes $N(t)$, and compound Poisson processes $\sum_{i=1}^{N(t)} Y_i$, where Y_i are i.i.d. random variables.

Of course, one can combine the above processes to form other Lévy processes. For example, an important class of Lévy processes is the jump-diffusion process (see **Default Risk**) given by $\mu t + \sigma W(t) + \sum_{i=1}^{N(t)} Y_i$, where μ and σ are constants. Interestingly, as per the famous Lévy–Itô decomposition the converse is also true. More precisely, any Lévy process can be written as a drift term μt , a Brownian motion with variance and covariance matrix $\mathbf{A}$, and a possibly infinite sum of independent compound Poisson processes that are related to an intensity measure $\nu(\mathbf{dx})$. This implies that a Lévy process can be approximated by jump-diffusion processes. This has important numerical applications in finance, as jump-diffusion models are widely used in finance.

The triplet $(\mu, \mathbf{A}, \nu)$ is also linked to the Lévy–Khinchin representation, which states that the characteristic function of a Lévy process X_t can be written in terms of $(\mu, \mathbf{A}, \nu)$ as

$$\frac{1}{t} \log E[e^{iz'X_t}] = -\frac{1}{2} z' \mathbf{A} z + i \mu z + \int_{\mathbb{R}^d} (e^{iz'x} - 1 - iz'x I_{\{|x| \leq 1\}}) \nu(\mathbf{dx}) \quad (1)$$

The representation suggests that it is easier to study Lévy processes *via* Laplace transforms and then numerically invert Laplace transforms. For numerical algorithms of Laplace inversion, see Abate and Whitt [1], Craddock *et al.* [2], and Petrella [3]. Some of the infinite activity Lévy processes may not have an analytical form for the intensity measure ν , although the probability density of X_t can be obtained explicitly; see, e.g., the generalized hyperbolic model in Eberlein and Prause [4].

There are two types of Lévy processes, jump-diffusion and infinite activity Lévy processes. In jump-diffusion processes, jumps are considered rare events, and in any given finite interval there are only finite number of jumps. Examples of jump-diffusion models in finance include Merton's [5]

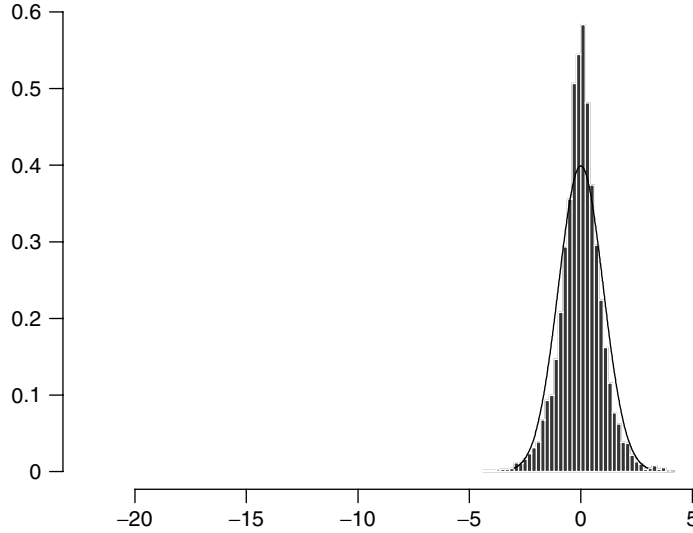


Figure 1 Comparison of the histogram of the normalized daily returns of S&P 500 index (from January 2, 1980 to December 31, 2005) and the density of $N(0,1)$. Note the feature of a high peak and two heavy tails (i.e., the leptokurtic feature)

model, in which the jump size Y has a normal distribution, and the double exponential jump-diffusion model in Kou [6], in which Y has a density $f_Y(y) = p \cdot \eta_1 e^{-\eta_1 y} 1_{\{y \geq 0\}} + q \cdot \eta_2 e^{\eta_2 y} 1_{\{y < 0\}}$. The double exponential jump-diffusion model leads to analytical solutions for path-dependent options, such as barrier and lookback options, and analytical approximations of finite-horizon American options, thanks to the memoryless property of exponential density (see [7–10]; **Equity-Linked Life Insurance; Credit Migration Matrices**). The economic justification of the double exponential jump-diffusion model, including the rational expectations equilibrium and the consideration from behavioral finance, is discussed in [6].

For infinite activity Lévy processes, in any finite time interval there are infinitely many jumps. Many of these models can be constructed *via* Brownian subordination, i.e., $W(\tau(t))$, where $\tau(t)$ is a subordinator, which is an increasing Lévy process. The subordinator must have no diffusion component (i.e., with only positive jumps and a positive drift); see Cont and Tankov [11, p. 88]. For example, if the subordinator $\tau(t)$ is a gamma process, then $W(\tau(t))$ leads to the variance gamma model (see Madan and Seneta [12] and Madan *et al.* [13]; **Extreme Values**

in Reliability). If the subordinator $\tau(t)$ is an inverse Gaussian process, then $W(\tau(t))$ leads to the normal inverse Gaussian model (see Barndorff-Nielsen [14] and Rydberg [15]).

There is a large literature on Lévy processes in finance, including several excellent books, e.g., the books by Cont and Tankov [11] and Kijima [16]. The book [11] also discusses the issue of hedging in incomplete markets (see **Equity-Linked Life Insurance; Individual Risk Models; Nonlife Insurance Markets; Risk Measures and Economic Capital for (Re)insurers**), as Lévy processes lead to incomplete markets, and the complete replication of an option payoff is impossible. One can also use rational expectations in Lucas [17] and Stokey and Lucas [18] to choose a risk-neutral measure to price derivative as in Kou [6]. Some additional topics related to Lévy processes are provided below:

1. In terms of computational issues, see Cont and Voltchkova [19] and d’Haullin *et al.* [20] on numerical methods *via* solving partial integrodifferential equations, and Feng and Linetsky [21] and Feng *et al.* [22] on how to price path-dependent options numerically *via* variational methods and extrapolation.

2. In terms of applications, see the references in [23] for applications of Lévy processes in fixed income derivatives (*see Credit Risk Models; From Basel II to Solvency II – Risk Management in the Insurance Sector; Credit Migration Matrices*) and term structure models, and the references in [24] for applications in credit risk and credit derivatives (*see Dynamic Financial Analysis*).
3. Statistical inference and econometric analysis for Lévy processes are discussed in the book by Singleton [25].

Some Difficulties

The Volatility Clustering Effect

In addition to the leptokurtic feature, returns distributions also have an interesting dependent structure, called the *volatility clustering effect* (*see* Engle [26]; **Default Correlation**). More precisely, the volatility of returns (which are related to the squared returns) are correlated, but asset returns themselves have almost no autocorrelation. In other words, a large movement in asset prices, either upside or downside, tends to generate large movements in the future asset prices, although the directions of the movements are unpredictable.

In particular, any model for stock returns with independent increments (such as Lévy processes) cannot incorporate the volatility clustering effect. However, one can combine Lévy processes with other processes (e.g. [27, 28]) or consider time-changed Brownian motion and Lévy processes to incorporate the volatility clustering effect. More precisely, if $\tau(t)$ contains a diffusion component (i.e., not a subordinator), then $W(\tau(t))$ and $X(\tau(t))$ may have dependent increments and no longer be Lévy processes (*see* Carr *et al.* [29, 30] and Carr and Wu [31]).

Difficulties in Distinguishing Tail Behavior

Although the main empirical motivation for using Lévy processes in finance comes from the fact that asset return distributions tend to have tails heavier than those of normal distribution, it is not clear how heavy the tail distributions are, as some people favor power-type distributions while others prefer exponential-type distributions, although, as pointed out in Kou [6, p. 1090], the power-type right tails cannot be used in models with continuous

compounding as they lead to infinite expectation for the asset price. We will stress that, quite surprisingly, it is very difficult to distinguish power-type tails from exponential-type tails from empirical data, unless one has extremely large sample size perhaps in the order of tens of thousands or even hundreds of thousands (*see* Heyde and Kou [32]). Therefore, it is very difficult to choose a good model based on the limited empirical data alone.

A good intuition may be obtained by simply looking at the quantiles for both standardized Laplace (with a symmetric density $f(x) = \frac{1}{2}e^{-x}I_{[x>0]} + \frac{1}{2}e^x I_{[x<0]}$) and standardized t distributions with mean 0 and variance 1. The right quantiles for the Laplace and normalized t densities with degrees of freedom from 3 to 7 are given in Table 1.

This table shows that Laplace distribution may have higher tail probabilities than those of t distributions, even if asymptotically the Laplace distribution should have lighter tails than those of t distributions. For example, regardless of the sample size, the Laplace distribution may appear to be heavier tailed than a t distribution with degrees of freedom (d.f.) 6 or 7, up to the 99.99 percentile. To distinguish the distributions, it is necessary to use quantiles with very low p-values and correspondingly large samples. If the true quantiles have to be estimated from data, then the problem is even worse, as the sample standard deviations need to be considered, resulting in sample sizes typically in the tens of thousands or even hundreds of thousands necessary to distinguish power-type tails from exponential-type tails.

Of course, one can use intraday data to have more sample sizes. However, intraday data may involve market microstructures, resulting in totally different empirical methods. For example, there is no single price due to the bid–ask spread, and the order size information is also relevant (*see* O’Hara [33]).

Table 1 Right quantiles for the Laplace and normalized t distributions with d.f. from 3 to 7

Probability (%)	Laplace	t_7	t_6	t_5	t_4	t_3
1	2.77	2.53	2.57	2.61	2.65	2.62
0.1	4.39	4.04	4.25	4.57	5.07	5.90
0.01	6.02	5.97	6.55	7.50	9.22	12.82
0.001	7.65	8.54	9.82	12.04	16.50	27.67

Risk due to Model Selection

The difficulties in distinguishing tail distributions indicate that there is a risk associated with choosing a proper model (see **Causal Diagrams**), as whether one prefers to use power-type distributions or exponential-type distributions is a subjective issue, which cannot be easily justified empirically. This also has implications in risk management, which is sensitive to the choice of models.

For example, a controversy in axiomatic approaches to risk measures is whether one should use value-at-risk or VaR (see **Simulation in Risk Management; Credit Scoring via Altman Z-Score; Extreme Value Theory in Finance**), which is a measure based on quantiles, or the tail conditional expectation.

The tail conditional expectation satisfies a set of axioms based on subadditivity (Artzner *et al.* [34]), while the VaR also satisfies a different set of axioms based on comonotonic (see **Comonotonicity**) subadditivity (Heyde *et al.* [35]), which is consistent with the prospect theory in behavioral finance. The intuition behind subadditivity is that merger should reduce risk, although the intuition may not be valid in the presence of the limited liability law. Furthermore, both VaR and the tail conditional expectation take into consideration the loss beyond the threshold, because VaR at a higher quantile (e.g., at 97.5%) can also be represented as tail conditional median (e.g., at 95%).

Whether one should measure risk by using VaR or the tail conditional expectation depends on whether the measure is used for internal or external risk management. The main advantage of VaR is that it is more robust against model assumptions and misspecifications, thus making it more suitable to generate more consistent results for external risk regulations and law enforcement. Indeed, VaR is widely used in governmental regulation, e.g., the recent Basel (II) (see **From Basel II to Solvency II – Risk Management in the Insurance Sector; Credit Risk Models; Informational Value of Corporate Issuer Credit Ratings**) banking regulations. On the other hand both the tail conditional expectation and VaR may be suitable for internal risk management.

Acknowledgment

I thank an anonymous referee and Prof. Ngai-Hang Chan for helpful comments.

References

- [1] Abate, J. & Whitt, W. (1992). The Fourier series method for inverting transforms of probability distributions, *Queueing Systems* **10**, 5–88.
- [2] Craddock, M., Heath, D. & Platen, E. (2000). Numerical inversion of Laplace transforms: a survey of techniques with applications to derivative pricing, *Journal of Computational Finance* **4**, 57–81.
- [3] Petrella, G. (2004). An extension of the Euler Laplace transform inversion algorithm with applications in option pricing, *Operations Research Letters* **32**, 380–389.
- [4] Eberlein, E. & Prause K. (2002). The generalized hyperbolic model: financial derivatives and risk measures, in *Mathematical Finance-Bachelier Congress 2000*, H. Geman, D. Madan, S. Pliska, T. Vorst, eds, Springer-Verlag, 245–267.
- [5] Merton, R.C. (1976). Option pricing when underlying stock returns are discontinuous, *Journal of Financial Economics* **3**, 125–144.
- [6] Kou, S.G. (2002). A jump-diffusion model for option pricing, *Management Science* **48**, 1086–1101.
- [7] Kou, S.G. & Wang, H. (2003). First passage time of a jump diffusion process, *Advances in Applied Probability* **35**, 504–531.
- [8] Kou, S.G. & Wang, H. (2004). Option pricing under a double exponential jump-diffusion model, *Management Science* **50**, 1178–1192.
- [9] Kou, S.G., Petrella, G. & Wang, H. (2005). Pricing path-dependent options with jump risk via Laplace transforms, *Kyoto Economic Review* **74**, 1–23.
- [10] Huang, Z. & Kou, S.G. (2006). *First Passage Times and Analytical Solutions for Options on Two Assets with Jump Risk*, Columbia University, Preprint.
- [11] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, 2nd Printing, Chapman & Hall/CRC Press, London.
- [12] Madan, D.B. & Seneta, E. (1990). The variance gamma (V.G.) model for share market returns, *Journal of Business* **63**, 511–524.
- [13] Madan, D.B., Carr, P. & Chang, E.C. (1998). The variance gamma process and option pricing, *European Finance Review* **2**, 79–105.
- [14] Barndorff-Nielsen, O. (1998). Processes of normal inverse Gaussian type, *Finance and Stochastics* **2**, 41–68.
- [15] Rydberg, T.H. (1997). The normal inverse Gaussian Levy process: simulation and approximation, *Communications in Statistics Stochastic Models* **13**, 887–910.
- [16] Kijima, M. (2002). *Stochastic Processes with Applications to Finance*, Chapman & Hall, London.
- [17] Lucas, R.E. (1978). Asset prices in an exchange economy, *Econometrica* **46**, 1429–1445.
- [18] Stokey, N.L. & Lucas, R.E. (1989). *Recursive Methods in Economic Dynamics*, Harvard University Press.

-
- [19] Cont, R. & Voltchkova, E. (2005). Finite difference methods for option pricing in jump diffusion and exponential Lévy models, *SIAM Journal of Numerical Analysis* **43**, 1596–1626.
 - [20] D'Halluin, Y., Forsyth, P.A. & Vetzal, K.R. (2003). *Rubust Numerical Methods for Contingent Claims Under Jump-Diffusion Processes*. Working Paper, University of Waterloo.
 - [21] Feng, L. & Linetsky, V. (2005). *Pricing Options in Jump-Diffusion Models: An Extrapolation Approach*, Northwestern University, Preprint.
 - [22] Feng, L., Linetsky, V. & Marozzi, M. (2004). *On the Valuation of Options in Jump-Diffusion Models by Variational Methods*, Northwestern University, Preprint.
 - [23] Glasserman, P. & Kou, S.G. (2003). The term structure of simple forward rates with jump risk, *Mathematical Finance* **13**, 383–410.
 - [24] Chen, N. & Kou, S.G. (2005). *Credit Spreads, Optimal Capital Structure, and Implied Volatility with Endogenous Default and Jump Risk*, Columbia University, Preprint.
 - [25] Singleton, K. (2006). *Empirical Dynamic Asset Pricing*, Princeton University Press.
 - [26] Engle, R. (1995). *ARCH, Selected Readings*, Oxford University Press.
 - [27] Duffie, D., Pan, J. & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions, *Econometrica* **68**, 1343–1376.
 - [28] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck based models and some of their uses in financial economics (with discussion), *Journal of Royal Statistical Society, Series B* **63**, 167–241.
 - [29] Carr, P., Geman, H., Madan, D. & Yor, M. (2002). The fine structure of asset returns: an empirical investigation, *Journal of Business* **75**, 305–332.
 - [30] Carr, P., Geman, H., Madan, D. & Yor, M. (2003). Stochastic volatility for Lévy processes, *Mathematical Finance* **13**, 345–382.
 - [31] Carr, P. & Wu, L. (2004). Time-changed Lévy processes and option pricing, *Journal of Financial Economics* **71**, 113–141.
 - [32] Heyde, C.C. & Kou, S.G. (2004). On the controversy over tailweight of distributions, *Operations Research Letters* **32**, 399–408.
 - [33] O'Hara, M. (1995). *Market Microstructure Theory*, Blackwell, London.
 - [34] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**, 203–228.
 - [35] Heyde, C.C., Kou, S.G. & Peng, X.H. (2006). *What is a Good Risk Measure: Bridging the Gaps Between robustness Subadditivity, Prospect Theory, and Insurance Risk Measures*, Columbia University, Preprint.

STEVEN KOU

Life Insurance Markets

What is Life Insurance?

There are two main types of life insurance policy: protection and investment policies. Protection policies are designed to provide a benefit (typically a lump-sum payment) following a specific event (such as death) if it occurs during the period (or term) for which the policy is valid, but nothing if the event does not occur during that time. Investment policies last until the policy is cancelled, and accumulate a cash value in the form of savings over the life of the policy. Pensions (*see* **Longevity Risk and Life Annuities; Risk Classification/Life; Securitization/Life**) are a form of life assurance that allow policyholders to build up a fund over the course of their working life, which is then converted into a payment on retirement. This payment may take the form of an annuity (*see* **Equity-Linked Life Insurance; Insurance Applications of Life Tables; Multistate Models for Life Insurance Mathematics**), which in its simplest form provides a stream of income until the insured person dies; thus insuring them against the risk that they will run out of resources because they live too long.

Life insurance policies can be bought either by the person whose life is insured, or by a third party. In cases where the third party is also the beneficiary, insurers try and limit the sale of life insurance policies to those who have an insurable interest (sometimes referred to as *cestui qui vit*), or in other words, to people who will materially suffer if the insured person dies. This is to reduce the possibility of moral hazard (*see* **Options and Guarantees in Life Insurance; Moral Hazard**), or that the beneficiary will actively seek the death of the insured.

Life insurance policies are also a popular form of employee benefit, in which case the third-party purchasing insurance and the beneficiary will differ. Within this group market, individuals will, typically, not be rated on their risk profile separately. Therefore, the value of the benefit to the individual will, in part, depend on how risky they are compared to the rest of the group.

Why Purchase Life Insurance?

Households typically have three motives for buying life insurance products: the desire to smooth

consumption over time, bequests, and tax advantages. The strength of the first two of these motives will depend on the individual or household, while the tax advantages associated with life insurance policies will depend on the fiscal regimes operating in individual countries.

The classic explanation of the consumption-smoothing motive for entering life insurance markets is provided by Yaari [1]. This uses a life-cycle model to show that for a consumer facing an uncertain lifetime, with no bequest motive, the optimal investment would be 100% annuitization. If the consumer does not wish to leave any bequests, then the best way to insure themselves against the risk of living too long is annuitization. More recently, [2] calculates that a 65-year-old male would be willing to give up one-third of his wealth to gain access to an actuarially fair annuities market.

At the level of the household, rather than the individual, annuities are not the only life insurance policy that can help with consumption smoothing. Life insurance policies such as term assurance (which pays out if the individual dies within a given period) can help working-age individuals protect their family against the loss of future labor earnings [3]. If life insurance products are not available, then one potential outcome could be inefficient overcaution, particularly on the part of those with a strong bequest motive, such as parents with young children.

There is some debate over the extent to which the bequest motive drives the purchase of life insurance. In the United States, 78% of couples aged over 70 own a protection-style life insurance policy on at least one member [4]. These holdings do not appear to be driven by the desire to protect the survivor against a drop in income. This leads to the hypothesis put forward by Bernheim [5] that life insurance is being held by elderly households to offset an excessive level of mandatory annuitization in the form of social security. This annuity-offset model would suggest the existence of a very strong bequest motive. However, tests of the annuity-offset model by Brown [4] could not find evidence that elderly households were buying term life insurance policies to leave a bequest. This finding has important implications for government policy – if mandatory annuitization leads to overannuitization because of a strong bequest motive, this would suggest that welfare could be improved by allowing households greater choice over the extent of annuitization. The strength of the

bequest motive is also questioned by Holtz-Eakin *et al.* [6], who find that purchase of life insurance by small business owners is insufficient to cover estate taxes. As this implies that the business may have to be sold, rather than stay within the family, following the owner's death, it indicates that the bequest motive is weaker than what the debate around estate taxes on small businesses would suggest.

Do Households Buy Enough Life Insurance?

There are different definitions of adequate life insurance ranging from a definition based on the utility of the survivor to simpler definitions that look at the ability of survivors to maintain their lifestyle. Whichever definition is used, however, the consensus is that there is significant underinsurance.

For example, [7] examines the adequacy of life insurance among married American households approaching retirement and finds that more than half the wives and one-quarter of the husbands in the sample are inadequately protected against the death of their spouse. In the case of wives, 15% would have suffered a reduction in living standards of greater than 40%. In other words, they suffered from severe underinsurance. A further 15% of wives would have suffered a reduction in income of between 20 and 40% if their husband had died in the year of the sample, and so also faced significant underinsurance. For husbands, 5 and 6% faced severe and significant underinsurance, respectively. One explanation put forward for this underinsurance is that people do not like thinking about death.

Similarly, the results of [8] suggest that there is significant underannuitization outside the social security system, even allowing for the existence of incomplete markets (*see Equity-Linked Life Insurance; Individual Risk Models; Weather Derivatives*).

The Role of Government

The extent to which people buy life insurance policies will depend partly on the role of government. Governments, for example, set the tax regime; and tax advantages play an important role in the demand for life insurance. In addition, some government policies act as a substitute for life insurance, which potentially leads to crowding out. The extent to which social

security policies provide a backup income for survivors will, in part, determine the importance of life insurance for the household as a whole.

Most public pension policies combine insurance with redistribution. The redistribution may be deliberate, for example, lower-paid workers may be entitled to higher replacement ratios (the proportion of income that the pension replaces). However, differences in life expectation mean that systems with mandatory annuitization at a uniform price will also result in redistribution in favor of those with higher life expectancy, typically those on higher incomes. Brown [9] investigates the impact of this type of program and shows that, provided administration costs are low, complete annuitization is welfare enhancing even for those with lower than average life expectancy. This is because, in the absence of annuities, even individuals with a high-mortality risk would need to set aside resources to cover the risk that they will live longer than expected. In addition, the associated redistribution is much lower if it is measured on a utility-adjusted basis, rather than simply in financial terms.

The other key role of government is in setting the regulatory environment in which insurance companies operate. This will have far-reaching implications for their business models, including the extent to which they can use risk pricing; the way in which they interact with customers, particularly on consumer protection issues such as suitability and treating customers fairly; and the amount of regulatory capital (*see Credit Migration Matrices; Operational Risk Modeling*) that they need to hold.

Commitment Issues

One of the features of life insurance products, particularly investment policies, is that they often involve long-term contracts. Unfortunately, events may make it difficult for consumers to commit over the long run. For example, [10] shows that in the United Kingdom only 60% of regular premium policies are still in force after 4 years. This lack of persistency is attributable to many factors. The Financial Services Authority [11] shows that 58% of lapses are because of income and affordability issues, with a further 10% attributed to marital or domestic reasons. The extent to which these types of events are predictable will have an impact on the desirability of an individual taking out a long-run investment policy.

However, there may be other reasons for a lack of commitment. When thinking about the purchase of life insurance products over time, there are two types of risk that consumers consider – the short-term event risk that an accident occurs within a period, and the long-term classification risk, or that changes in the consumer's type will lead to variability in future premiums. Consumers who receive positive (better than average) news about their risk characteristics will have an incentive to switch providers to take advantage of their improved health. This will make it difficult for them to commit to a long-term contract, and for insurance companies, in turn, to provide full insurance against classification risk. Hendel and Lizzeri [12] show that markets have got round the problem of lack of commitment from consumers by front-loading premium payments within long-term policies. This acts as a partial lock-in for consumers. The results show that contracts that are more front loaded have higher persistency (lower lapses), and hence keep a higher proportion of customers in the long run. As better risk types are more likely to switch provider, contracts with higher lapse rates retain worse risk pools. While front loading comes at a cost because it makes it harder for households to smooth consumption, [12] estimates that this cost is low. The benefits of front loading are that consumers suffer very little reclassification risk, so that insurance companies in the United States are able to commit to guaranteed renewable contracts for life insurance without the need for regulation.

Risk Pricing

Risk pricing (*see Actuary; Risk Measures and Economic Capital for (Re)insurers*) is the method that insurers use to try to price the risks associated with a specific group of individuals to reflect the likelihood that they will claim. If insurers are unable to use risk-pricing, for example because of government legislation designed to reduce age or sex discrimination, the result is likely to be that: either the higher risk groups will find it difficult to obtain life insurance; or that those with a better risk profile will tend not to buy insurance, because the price that they face is too high. One of the benefits of risk pricing has been the development of “impaired life” insurance for those with significant health problems such as diabetes – a group who would previously have struggled to get cover.

At the decumulation (in payment) stage, the risk for the insurer from pensions is longevity risk – in other words, the possibility that the insurer may have underestimated the length of time the pensioner will survive. For other forms of life insurance, the typical risk is mortality risk, the risk that the policyholder will die. One of the problems with correctly pricing annuities is predicting changes to longevity risk, as life expectancy has been rising sharply over recent decades. The extent to which the rate of change of improvements in life expectancy can be expected to continue will have a big impact on the profitability of annuity contracts, as well as on bulk-buyout transactions (whereby life insurers take over either another insurer's book of business or the pension risks associated with individual companies' defined benefit pension schemes). Blake *et al.* [13] provides a discussion on how these risks have changed over time by examining longevity fan charts. Given significant differences in life expectancy for men and women, it will be particularly important to account for gender differences in risk-pricing calculations for life insurance products such as annuities.

One of the problems for an insurer of accurately pricing the risks associated with a particular product offering is the existence of asymmetric information. Customers will typically know more about their own personal risk characteristics than their insurers, and this can lead to adverse selection – the people taking out insurance are, in fact, more risky for the insurance company than their observable characteristics would suggest. In general, with insurance, it can be difficult to disentangle the impact of adverse selection and moral hazard (whereby the existence of insurance changes people's behavior). In other words, if claims are higher for those with insurance than would be the case for a seemingly identical population without insurance, it is impossible to tell if this is the result of moral hazard or adverse selection. For most life insurance products, however, the problems of moral hazard are typically thought to be less than would be the case for some other types of insurance. They can also be more easily managed, for example, through the use of clauses that exclude payouts in the event of a suicide within a certain time period. However, adverse selection definitely exists, although it may not show up in the simple headline number of amount insured.

For example, Finkelstein and Poterba [14] demonstrates that in the UK annuities market, mortality patterns are consistent with the existence of asymmetric

information and that high-risk individuals self-select features of insurance contracts that are more valuable to them, for a given price, than would be true for low-risk individuals. The contract features where this self-selection occurs are the time profile of annuity payments, and whether the annuity will make payments to the annuitant's estate in the event of death. Annuitants who live longer tend to have selected back-loaded payment streams (where income rises over time), while those who die early will often have selected an annuity type that makes payments to their estate in the event of death. These selection effects are large – the mortality differences between those who opt for back-loaded *versus* non-back-loaded annuities are larger than the mortality differences between male and female annuitants. Self-selection does not appear to be prevalent in the amount of the initial annuity payment, which is equivalent to a measure of the amount insured. This may explain the mixed nature of the results in previous studies on adverse selection in insurance markets, which tended to concentrate on the amount of payment in the event that the insured risk occurs.

Buying Life Insurance

One of the key issues in the sale of life insurance products is the role of advice. Life insurance products can be complex and difficult to compare, meaning that consumers face high search costs. This suggests a natural role for tailored, face-to-face advice. The important role of full advice, in the regulatory sense, is reinforced by regulation such as the “suitability” requirement in the United Kingdom, whereby sellers need to demonstrate that financial products are suitable for purchasers.

However, other routes to market can have a big impact on the search costs that consumers face, particularly on simpler products such as term assurance. The most obvious source of reduced search costs is the rise of online insurance comparison sites. Brown and Goolsbee [15] looks at the impact of online insurance sites on the prices paid for term assurance by different groups. They show that the faster a group adopted the Internet, the faster the prices of term assurance dropped for that group and that these changes cannot be explained by changes in mortality. They show that the rise of the Internet in the United States reduced the price of term life assurance by between 8 and 15%, equivalent to an annual increase

in the consumer surplus of \$115 and \$215 million on these policies.

References

- [1] Yaari, M. (1965). Uncertain lifetime, life insurance and the theory of the consumer, *Review of Economic Studies* **32**, 137–150.
- [2] Mitchell, O., Poterba, J.M., Warshawsky, M.J. & Brown, J.R. (1999). New evidence on the money's worth of individual annuities, *American Economic Review* **89**, 1299–1318.
- [3] Lewis, F.D. (1989). Dependents and the demand for life insurance, *American Economic Review* **79**, 452–467.
- [4] Brown, J.R. (2001). Are the elderly really over-annuitized? New evidence on line insurance and bequests, in *Themes in the Economics of Aging*, D. Wise, ed, University of Chicago Press.
- [5] Bernheim, B.D. (1991). How strong are bequest motives? Evidence based on estimates of demand for life insurance and annuities, *Journal of Political Economy* **99**, 899–927.
- [6] Holtz-Eakin, D., Phillips, J.W. & Rosen, H.S. (2001). Estate taxes, life insurance and small businesses, *Review of Economics and Statistics* **83**, 52–63.
- [7] Bernheim, B.D., Forni, L., Gokhale, J. & Kotlikoff, L.J. (2003). The mismatch between life insurance holdings and financial vulnerability: evidence from the health and retirement survey, *American Economic Review* **93**, 354–365.
- [8] Davidoff, T., Brown, J.R. & Diamond, P.A. (2005). Annuities and individual welfare, *American Economic Review* **95**, 1573–1590.
- [9] Brown, J.R. (2003). Redistribution and insurance: mandatory annuitization with mortality heterogeneity, *Journal of Risk and Insurance* **70**, 17–41.
- [10] Financial Services Authority (2006). *2006 Survey of the Persistency of Life and Pension Policies*, http://www.fsa.gov.uk/pubs/other/Persistency_2006.pdf.
- [11] Financial Services Authority (2000). *Persisting: why consumers stop paying into policies*, Consumer Research 6.
- [12] Hendel, I. & Lizzeri, A. (2003). The role of commitment in dynamic contracts: evidence from life insurance, *Quarterly Journal of Economics* **118**, 299–327.
- [13] Blake, D., Cairns, A.J. & Dowd, K. (2007). *Facing Up to the Uncertainty of Life: the Longevity of Fan Charts*, Pensions Institute, Discussion Paper No 0703.
- [14] Finkelstein, A. & Poterba, J. (2004). Adverse selection in insurance markets: policy holder evidence from the UK annuity market, *Journal of Political Economy* **112**, 193–208.
- [15] Brown, J.R. & Goolsbee, A. (2002). Does the Internet make markets more competitive? Evidence from the life insurance industry, *Journal of Political Economy* **110**, 481–507.

Lifetime Models and Risk Assessment

Researchers in different fields have investigated risk assessment using methods for lifetime data. Lifetime data have been referred to as *survival time data* in biomedical research, as *reliability data* in engineering applications, and as *time-to-event data* in social sciences. In general, lifetime data can be defined as the times to the occurrence of a certain event. Statistical methods for analyzing lifetime data include the definition of survival time, censored data, prediction of probability of survival, prediction of instantaneous hazards, evaluation of cure rates, etc. We review different types of lifetime data, methods for estimating hazard and survival functions, and methods for risk assessment. Detailed information can be found in the following books: Kalbfleisch and Prentice [1], Nelson [2], Cox and Oakes [3], Fleming and Harrington [4], Andersen *et al.* [5], Marubini and Valsecchi [6], Klein and Moeschberger [7], Lawless [8], Lee and Wang [9], and others. Hougaard [10] gives good descriptions of analytical methods for multivariate survival data. Bayesian survival analysis can be found in Ibrahim *et al.* [11]. A Bayesian approach for reliability models and risk assessment is given by Singpurwalla [12]. Lifetime models, based on stochastic processes hitting a boundary, have been investigated by Whitmore [13, 14], Lu [15], Whitmore *et al.* [16], Onar and Padgett [17], Aalen and Gjessing [18], Padgett and Tomlinson [19, 20], and others. A review of threshold regression for risk assessment, based on the concept of a first hitting time (FHT), can be found in Lee and Whitmore [21].

Important Parametric Distributions for Lifetime Models

Several parametric distributions have been used in analyzing lifetime data, including the exponential, Weibull, inverse Gaussian, gamma and generalized gamma, lognormal, log-logistic, and various accelerated failure-time distributions.

Censored and Truncated Data

In practical situations, it often happens that we cannot observe all the subjects in the study for a complete lifetime data. For example, patients may be either lost to follow-up or still alive at the end of the study. These kinds of data are called *censored data*. There are several types of censoring mechanisms.

1. Type I censoring

Each individual has a fixed potential censoring time $C_i > 0$ such that lifetime T_i is observed if $T_i \leq C_i$; otherwise, we only know that $T_i > C_i$.

2. Type II censoring

In a random sample of size n , we only observe the r smallest lifetimes $t_{(1)} \leq t_{(2)} \leq \dots \leq t_{(r)}$, where $r < n$. Both type I and type II censored data, as defined above, are also referred to as *being right censored*.

3. Independent random censoring

It is common that the censoring mechanism is random and independent of the failure process. In this situation, the lifetimes T_i and censoring times C_i are independent, for $i = 1, \dots, n$. Letting I denote an indicator function and defining $\delta_i = I(T_i \leq C_i)$, the data observed from n individuals in a random censoring scheme are therefore pairs (t_i, δ_i) , for $i = 1, \dots, n$, where $t_i = \min(T_i, C_i)$.

4. Interval censoring

In longitudinal studies or clinical trials, interval censoring may occur when a patient has periodic follow-ups and the patient's event time is only known to fall in an interval $(L_i, R_i]$, for $i = 1, \dots, n$.

5. Current status data

Interval censored data where the intervals are either (C_i, ∞) or $(0, C_i]$ are called *current status data*. These data arise when subject i is examined once at time C_i and the event occurrence is determined at the same time C_i . For example, the time to tumor occurrence of a mouse can only be checked at the time of autopsy.

6. Truncation

Truncation is a condition where only those individuals who experience certain events can be observed by the investigator. For example, left truncation occurs when individuals of different ages enter a retirement study (i.e., delayed entry) and are followed until either an event or right censoring occurs. Individuals

who die before retirement cannot be included in the study. In AIDS studies, right truncation often occurs in that individuals who are yet to develop AIDS are neither observed nor are they included in the study sample.

Calendar or Running Time

In many applications, the natural time scale of the failure process is not calendar or clock time but rather some measure of use or cumulative exposure. Cox and Oakes [3, Section 1.2, pp. 3–4] point out that “often the *scale* for measuring time is clock time, although other possibilities certainly arise, such as the use of operating time of a system, mileage of a car, or some measure of cumulative load encountered” Lawless [8, p. 6] also states that “Miles driven might be used as a time scale for motor vehicles, and number of pages of output for a computer printer or photocopier.” This kind of alternative time scale is sometimes referred to as an *operational time* or *running time*. Survival analysis can be done in terms of either calendar time or operational time.

Estimation of Survival and Hazard Functions

The following are the basic theoretical quantities in lifetime models and related risk assessment:

1. Failure-time probability density function $f(t)$

This function describes the rate at which items will fail at any given time $T = t$.

$$f(t) dt = Pr(T \in [t, t + dt)) \quad (1)$$

2. Survival function $S(t)$

This function is the probability that an item will fail after time t .

$$S(t) = Pr(T > t) \quad (2)$$

3. Hazard function $h(t)$

This function describes the rate of failure at time t among items that have survived to time t .

$$h(t) = -\frac{d \ln S(t)}{dt} = \frac{f(t)}{S(t)} \quad (3)$$

4. Cumulative hazard function $H(t)$:

This function is the aggregate hazard to which an item is exposed up to time t .

$$H(t) = \int_0^t h(u) du = -\ln[S(t)] \quad (4)$$

Estimating the hazard function and survival function is the major focus of lifetime data analysis. From a process point of view, Aalen and Gjessing [22] present a good understanding of the hazard function.

If the data have already been grouped into intervals, a life table can be constructed to assess survival probabilities and hazard rates at different intervals. For parametric models, likelihood-ratio testing procedures and parametric regression models for handling censored and/or truncated data have been well developed.

Kaplan – Meyer Nonparametric Estimator for Survival Function

When individual lifetimes can be observed, the product-limit nonparametric estimator of the survival function developed by Kaplan and Meyer [23] is the most widely used method for estimating survival functions. In the situation where n distinct lifetimes are observed without censoring such that their ordered values are $t_{(1)} < t_{(2)} < \dots < t_{(n)}$, the survival function can be estimated by

$$\hat{S}(t) = \prod_{t_{(i)} \leq t} \frac{(n - i)}{(n - i + 1)} \quad (5)$$

In general, suppose d_i subjects have lifetime $t_{(i)}$, $1 \leq i \leq n$. Let Y_i denote the stochastic process indicating the number of individuals at risk at time $t_{(i)}$. Then

$$\hat{S}(t) = \prod_{t_{(i)} \leq t} \left(1 - \frac{d_i}{Y_i}\right) \quad (6)$$

Nelson – Aalen Estimator for Cumulative Hazard Function

In a reliability context, Nelson [24] introduced the following estimator for the cumulative hazard function

$$\hat{H}(t) = \sum_{t_{(i)} \leq t} \left(1 - \frac{d_i}{Y_i}\right) \quad (7)$$

Using methods of counting processes, Aalen [25] also derived this estimator, which is now referred to as the *Nelson – Aalen estimator*.

Semiparametric Proportional Hazards Models and Cox Regression

Cox [26] introduced the proportional hazards (PH) model. Given a covariate vector $\mathbf{X} = \mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, this model assumes that the hazard function for lifetime T is of the form

$$h(t|\mathbf{X} = \mathbf{x}) = h_0(t) \exp(\boldsymbol{\beta}'\mathbf{x}) \quad (8)$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients and $h_0(t)$ denotes the baseline hazard function.

For any two individuals with covariate vectors $\mathbf{X}_1 = \mathbf{x}_1$ and $\mathbf{X}_2 = \mathbf{x}_2$, the ratio of their hazard rates is constant through time.

$$\frac{h(t|\mathbf{X}_2 = \mathbf{x}_2)}{h(t|\mathbf{X}_1 = \mathbf{x}_1)} = \frac{h_0(t) \exp(\boldsymbol{\beta}'\mathbf{x}_2)}{h_0(t) \exp(\boldsymbol{\beta}'\mathbf{x}_1)} = \exp[\boldsymbol{\beta}'(\mathbf{x}_2 - \mathbf{x}_1)] \quad (9)$$

This quantity is called the *hazard ratio* (or *relative risk*) and is very useful in applications. For PH models, it can be shown that

$$S(t|\mathbf{X} = \mathbf{x}) = [S_0(t)]^{\exp(\boldsymbol{\beta}'\mathbf{x})} \quad (10)$$

where $S_0(t)$ is the baseline survival function.

Additive Hazards Models

In a PH model, as mentioned above, the covariates act multiplicatively on some unknown baseline hazard function. In an additive hazards regression model, the hazard rate at time t is assumed to be a linear combination of the covariate vector $\mathbf{X}$.

$$h(t|\mathbf{X} = \mathbf{x}) = \beta_0(t) + \boldsymbol{\beta}'\mathbf{x} \quad (11)$$

Therefore, in this model, the covariates act in an additive manner on an unknown baseline hazard function $\beta_0(t)$.

Accelerated Failure-Time Model

Given any lifetime t , the accelerated failure-time model is defined by the equation

$$S(t|\mathbf{X} = \mathbf{x}) = S_0[t \exp(\boldsymbol{\theta}'\mathbf{x})] \quad (12)$$

The factor $\exp(\boldsymbol{\theta}'\mathbf{x})$ is called an *acceleration factor* and indicates how a change in a covariate value changes the baseline time scale. Specifically, the hazard rate for an individual with covariate vector $\mathbf{X} = \mathbf{x}$ can be written as

$$h(t|\mathbf{x}) = \exp(\boldsymbol{\theta}'\mathbf{x}) h_0[t \exp(\boldsymbol{\theta}'\mathbf{x})] \quad (13)$$

Frailty Models

In multivariate survival analysis, the frailty model has become popular. In practice, a frailty is an unobservable random effect shared by subjects within a subgroup. In one formulation of frailty, a common random effect acts multiplicatively on the hazard rates of all members in the subgroup. The shared frailty model extends the PH regression model as follows. The hazard rate for the j th individual in the i th group is of the form

$$h_{ij}(t|\mathbf{x}_{ij}) = h_0(t) \exp[\sigma w_i + \boldsymbol{\beta}'\mathbf{x}_{ij}] \quad (14)$$

where w_i is the frailty shared by the i th group.

First Hitting Time Models and Threshold Regression

FHTs arise naturally in many types of stochastic processes. In a lifetime context, the state of the underlying process represents the strength of an item or the health of an individual. The item fails or the individual experiences a clinical endpoint when the process reaches an adverse threshold state for the first time. In many applications, the process is latent (i.e., unobservable). *Threshold regression* refers to FHT models with regression structures that accommodate covariate data. The parameters of the process, threshold state, and time scale may depend on the covariates.

An FHT model has two basic components: (a) a *parent stochastic process* $\{X(t)\}$ and (b) a *threshold*. The FHT is the time when the stochastic process first crosses the *threshold*. Whether the sample path of the parent process is observable or latent (unobservable) is an important distinguishing characteristic of the FHT model. Latent processes are the most common by far. As an example, we consider a Wiener process $\{X(t), t \geq 0\}$ with mean parameter μ , and variance parameter σ^2 as the parent stochastic process, and initial value $X(0) = x_0 > 0$.

In threshold regression, parameters μ , σ^2 , and initial value x_0 will be connected to linear combinations of covariates using suitable regression link functions, as illustrated below for some parameter, say θ .

$$g_\theta(\theta_i) = \mathbf{z}_i \boldsymbol{\beta} \quad (15)$$

Here g_θ is the link function, parameter θ_i is the value of parameter θ for individual i , $\mathbf{z}_i = (1, z_{i1}, \dots, z_{ik})$ is the covariate vector of individual i (with a leading unit to include an intercept term), and $\boldsymbol{\beta}$ is the associated vector of regression coefficients.

An identity function of form

$$\mu = \mathbf{z} \boldsymbol{\beta} = \beta_0 + \beta_1 z_1 + \dots + \beta_k z_k \quad (16)$$

might be used to link parameter μ to the covariates and a logarithmic function

$$\ln(x_0) = \mathbf{z} \boldsymbol{\gamma} = \gamma_0 + \gamma_1 z_1 + \dots + \gamma_k z_k \quad (17)$$

for the initial value parameter x_0 . Parameter σ^2 can be given an arbitrary value (say 1) because the process $\{X(t)\}$ is latent.

A running-time scale transformation can be accommodated by the threshold regression model. If $r(t)$ denotes the transformation of calendar time t to running time r , with $r(0) = 0$, and $\{X(r)\}$ is the parent process defined in terms of running time r , then the resulting process expressed in terms of calendar time is the process $X^*(t) = X[r(t)]$. A *composite running time* might be defined by

$$r(t) = \sum_{j=1}^J \alpha_j r_j(t) \quad (18)$$

where the $r_j(t)$ are different accumulation measures that can advance degradation or disease progression and the α_j are positive parameters that weight the contributions of the different measures.

As an example, consider a Wiener process with $x_0 > 0$ and the boundary as the zero level. It is well known that the FHT has an inverse Gaussian distribution if $\mu < 0$. The corresponding survival function $S(r|\mu, \sigma^2, x_0)$ can be written as

$$\Phi \left[\frac{\mu r + x_0}{\sqrt{\sigma^2 r}} \right] - \exp(-2x_0 \mu / \sigma^2) \Phi \left[\frac{\mu r - x_0}{\sqrt{\sigma^2 r}} \right] \quad (19)$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. Lee

et al. [27] demonstrated how to use this model in analyzing AIDS data. Lee *et al.* [28] used the threshold regression model in assessing lung cancer risk to a large cohort of US railroad workers from diesel exhaust exposure. Bayesian methods for FHT models have been discussed by Pettit and Young [29] and Shubina [30, 31].

Acknowledgment

This project was supported in part by NIH Grant OH008649.

References

- [1] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The statistical Analysis of Failure Time Data*, John Wiley & Sons, New York.
- [2] Nelson, W. (1982). *Applied Life Data Analysis*, John Wiley & Sons, New York.
- [3] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*, Chapman & Hall, New York.
- [4] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*, John Wiley & Sons, New York.
- [5] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.
- [6] Marubini, E. & Valsecchi, M.G. (1995). *Analyzing Survival Data from Clinical Trials and Observational Studies*, John Wiley & Sons, New York.
- [7] Klein, J.P. & Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer.
- [8] Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd Edition, John Wiley & Sons.
- [9] Lee, E.T. & Wang, J.W. (2003). *Statistical Methods for Survival Data Analysis*, John Wiley & Sons.
- [10] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer, New York.
- [11] Ibrahim, J., Chen, M.-H. & Sinha, D. (2001). *Bayesian Survival Analysis*, Springer-Verlag.
- [12] Singpurwalla, N.D. (2006). *Reliability and Risk: A Bayesian Perspective*, John Wiley & Sons.
- [13] Whitmore, G.A. (1986). First passage time models for duration data regression structures and competing risks, *The Statistician* **35**, 207–219.
- [14] Whitmore, G.A. (1995). Estimating degradation by a Wiener diffusion process subject to measurement error, *Lifetime Data Analysis* **1**, 307–319.
- [15] Lu, J. (1995). *A reliability model based on degradation and lifetime data*, Ph.D. thesis, McGill University, Montreal.
- [16] Whitmore, G.A., Crowder, M.J. & Lawless, J.F. (1998). Failure inference from a marker process based on

- a bivariate Wiener model, *Lifetime Data Analysis* **4**, 229–251.
- [17] Onar, A. & Padgett, W.J. (2000). Inverse Gaussian accelerated test models based on cumulative damage, *Journal of Statistical Computation and Simulation* **66**, 233–247.
 - [18] Aalen, O.O. & Gjessing, H.K. (2004). Survival models based on the Ornstein-Uhlenbeck process, *Lifetime Data Analysis* **10**, 407–423.
 - [19] Padgett, W.J. & Tomlinson, M.A. (2004). Inference from accelerated degradation and failure data based on Gaussian process models, *Lifetime Data Analysis* **10**, 191–206.
 - [20] Padgett, W.J. & Tomlinson, M.A. (2005). Accelerated degradation models for failure based on geometric Brownian motion and Gamma processes, *Lifetime Data Analysis* **11**, 511–527.
 - [21] Lee, M.-L.T. & Whitmore, G.A. (2006). Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary, *Statistical Sciences* **21**, 501–513.
 - [22] Aalen, O.O. & Gjessing, H.K. (2001). Understanding the shape of the hazard rate: a process point of view, *Statistical Sciences* **16**, 1–22.
 - [23] Kaplan, E.L. & Meyer, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.
 - [24] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data, *Technometrics* **14**, 945–965.
 - [25] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.
 - [26] Cox, D.R. (1972). Regression models and lifetables (with discussion), *Journal of the Royal Statistical Society (JRSSB), Series B* **34**, 187–220.
 - [27] Lee, M.-L.T., DeGruttola, V. & Schoenfeld, D. (2000). A model for markers and latent health status, *Journal of the Royal Statistical Society, Series B* **62**, 747–762.
 - [28] Lee, M.-L.T., Garshick, E., Whitmore, G.A., Laden, F. & Hart, J. (2004). Assessing lung cancer risk to rail workers using a first hitting time regression model, *Environmetrics* **15**, 1–12.
 - [29] Pettit, L.I. & Young, K.D.S. (1999). Bayesian analysis for inverse Gaussian lifetime data with measures of degradation, *Journal of Statistical Computation and Simulation* **63**, 217–234.
 - [30] Shubina, M. (2005). *Bayesian analysis for markers and degradation*, Ph.D. Dissertation, Biostatistics Department, Harvard School of Public Health, Boston.
 - [31] Shubina, M. (2005). *Threshold models with markers measured before observed event times*, Ph.D. Dissertation, Biostatistics Department, Harvard School of Public Health, Boston.

MEI-LING TING LEE

Linkage Analysis

In biology, a *trait* or character is a feature of an organism. Most biological traits including human health and physical characteristics are determined primarily by the chemical composition of the proteins created in cells. The composition and formation of these proteins are in turn determined by the genetic material. Genetic material is aligned in pairs of long strands called *chromosomes*. Roughly, for any given protein, there is a particular responsible location (called *locus*) on one of the chromosomes. The chemical sequence of the genetic locus on the chromosome determines the composition of a particular protein which, in turn, determines traits and their *phenotypes* [1]. The term *phenotype* refers to the state of the trait (e.g., the trait eye color has the phenotypes blue, brown, and hazel).

Discovering the genetic loci that correspond to particular traits of interest is an important scientific endeavor that can have profound impact on improving public health and/or creating major economic benefits. Many human diseases including various types of cancer, heart disease, and diabetes are known to have some underlying genetic predisposition [1, 2]. In medical studies, the determination of disease susceptibility loci is an important initial step in identifying genetic etiology and for effective intervention of these diseases. The identification of genetic loci that affect certain traits also has applications in plant and animal studies, such as for improving grain yield in rice and increasing milk production in cows.

Identification of the genetic loci that correspond to particular traits cannot be done by simply measuring the traits of interest in a sample of subjects and then comparing the traits with the subjects' chromosomal chemical sequences. Even with modern biotechnology and computing power, it is still impractical to determine the whole sequence, which is too long to be done under usual time and cost constraints. An often useful idea is to first localize the genetic loci that correspond to particular traits to a relatively small chromosomal region, which can be done using available fragments of the sequence from a subset of loci in the chromosomes. Such genetic loci are termed *marker loci*. There are numerous marker loci with known chromosomal positions spreading densely throughout the entire human genome. Thus,

the genetic loci that correspond to particular traits are likely to be close to some of these marker loci. Moreover, genetic materials at close-by loci tend to be inherited together, and thus a genetic marker close to a trait locus tends to be inherited together, which means the marker tends to be coinherited with the trait. *linkage analysis* aims to test for coinheritance of marker loci with particular traits.

Many useful methods for linkage analysis have been developed during the last few decades. The classical linkage analysis method uses the logarithm of odds (LOD) score, which is equivalent to the likelihood-ratio test statistic [3]. This method requires the specification of a detailed inheritance model for the trait. When such models can be correctly specified as in the case of a simple Mendelian disease, the LOD score method is the method of choice owing to the power optimality of the likelihood-ratio test (LRT). For complex traits, however, it is often impossible to correctly specify a detailed inheritance model for the trait. The so-called model-free method for linkage analysis may be used, which does not require to specify, *a priori*, an inheritance model for the trait of interest. Frequently, model-based methods are called *parametric methods*, whereas model-free methods are called *nonparametric methods*. However, there is no sharp boundary that can clearly distinguish model-based analysis from model-free analysis, as both the theory and applications have considerable overlap [4, 5]. Traditionally, linkage methods for mapping qualitative traits are developed separately from those methods for mapping quantitative traits; many methods that are useful for both qualitative and quantitative traits are currently available. Most methods have been developed first for establishing linkage between two genetic loci (two point) and then extended to multipoint linkage analysis. The latter can be better suited for mapping complex traits; however, this often involves heavy computational burden. The rest of the article is devoted to a brief overview of the commonly used methods for linkage analysis.

Biological Basis of Linkage

A large number of loci spread densely throughout the whole genome (called *marker loci*) have been identified and their exact positions are known in the

chromosomes. The genetic material at these marker loci are segments of the DNA (deoxyribonucleic acid) whose inheritance from parents to offspring can be followed. The DNA segments at these marker loci are considerably variable between individuals. However, with modern biotechnology, it is straightforward to determine the polymorphic DNA sequences at these marker loci. Typically, at any marker locus, there exist only a few possible sequence values that the marker DNA sequences can take on. These values are termed *alleles*, and the pair of alleles obtained from an individual's marker loci in the pair of chromosomes comprise the *genotype*. An individual's genotype at a locus is called *homozygous* if the pair of alleles are identical and *heterozygous* if they differ. The arrangement of the alleles on the pairs of chromosomes is called the *phase*. The genotype data for multiple loci with the relevant phase information are the *haplotype*. Consider the simple case of a locus defined with two alleles, A_1 and A_2 . Denote the allele frequency of A_i in the population as $p_i, i = 1, 2$. If the genotype frequencies of A_1A_1 , A_1A_2 , and A_2A_2 in the population are p_1^2 , $2p_1p_2$, and p_2^2 , respectively, the so-called Hardy-Weinberg equilibrium (HWE) holds. If these frequencies are violated, it is said that there is a disequilibrium. Similarly, if frequencies of two alleles at two or more genetic loci (haplotype) do not satisfy the HWE, it is called *linkage disequilibrium* (LD) or *allelic association* [1].

The transmission pattern of genetic material from parents to offspring is of fundamental importance to linkage analysis. For each individual, one member of each pair of chromosomes is inherited maternally, and the other is inherited paternally. An offspring does not inherit a complete copy of the sequence that forms either of the parental chromosomes. Instead, each inherited chromosome represents a mixture of subsequences of the corresponding parental pair. The number of subsequences and the locations of their boundaries are random and independent for different offspring. The events that generate the boundaries are termed *crossovers*. An odd number of crossovers between two loci is termed a *recombination* [1].

Importantly, DNA segments at close-by loci on a single chromosome are often transmitted intact from parents to offspring without recombination between them. The recombination fraction, θ , between two loci is the relative frequency of recombination. If the two loci on a single chromosome are very close to

each other, the recombination fraction θ is close to 0, and if the two loci are far away then θ is close to 1/2. If two loci are on different chromosomes, $\theta = 1/2$. Two loci are often said to have linkage or said to be linked if they are close enough on the same chromosome that their segments cosegregate, i.e., $\theta < 1/2$ [1, 4]. The statistical test of no linkage between two loci is to test

$$H_0 : \theta = \frac{1}{2} \quad \text{against} \quad H_1 : \theta < \frac{1}{2} \quad (1)$$

The classical linkage test is based on the likelihood ratio, the logarithm of which is called a *LOD score* [1, 3], which is discussed next.

LOD Score Analysis

Frequently, a parametric model is employed to evaluate the probability of the observed pedigree data under certain assumptions on the recombination fraction θ and other parameters such as gene frequencies. The likelihood function for the i th pedigree will be denoted as $L_i(\theta)$. It is natural to use the LRT for linkage analysis. The likelihood-ratio function for the i th pedigree can be written as $\Lambda_i(\theta) = L_i(\theta)/L_i(0.5)$ and the LRT statistic is $LR = 2 \max_{0 \leq \theta \leq 0.5} \sum_i \log \Lambda_i(\theta)$. The LOD score function for the i th pedigree is $Z_i(\theta) = \log_{10} \Lambda_i(\theta)$, and for all independent pedigrees it is

$$Z(\theta) = \sum_i \log_{10} \Lambda_i(\theta) \quad (2)$$

and the maximum LOD score is simply $\max_{0 \leq \theta \leq 0.5} Z(\theta)$. Traditionally, the LOD score test is called *significant* if $\max_{0 \leq \theta < 0.5} Z(\theta) > 3$ [1, 3]. Next we discuss a few special cases of linkage analysis using the LOD score.

Phase-Known Pedigrees

In some animal cross studies, scientists can breed animals to the point where it is possible to determine which of their offspring are recombinants and which are nonrecombinants from the genotype data [1]. For the i th pedigree, suppose that the number of recombinants is r_i out of N_i meioses. We can estimate the recombination fraction θ using the likelihood $L(\theta) = \prod_i L_i(\theta)$, where

$$L_i(\theta) = \theta^{r_i} (1 - \theta)^{N_i - r_i}, \quad \theta \in [0, 0.5] \quad (3)$$

The maximum-likelihood estimate (MLE) of the recombination fraction θ , denoted as $\hat{\theta}$, is $\hat{\theta} = \min(\sum r_i / \sum N_i, 0.5)$, where the summation is over independent pedigrees.

Notice that $\theta = 1/2$ is on the boundary of the parameter space $[0, 1/2]$, and the classical LRT statistic $LR = \max_{0 \leq \theta \leq 0.5} 2 \log[L(\theta)/L(0.5)]$ for equation (3) has an asymptotic distribution with 50% point mass at 0 and 50% as $\chi^2(1)$, which is often denoted as

$$0.5\chi^2(0) + 0.5\chi^2(1) \quad (4)$$

If we assume the above distribution for the ordinary LRT statistic, the maximum LOD score, $\max_{\theta} Z(\theta) = LR/(2 \log 10)$, exceeds 3, and corresponds to a P value of 0.0001. In many genetic linkage analysis, a large number of loci are often being tested, and thus the P values are often required to be small to declare statistical significance due to concerns of multiple testing or multiple comparisons [6, 7].

Phase-Unknown Pedigrees

Consider a nuclear family with a pair of biallelic loci. Denote the alleles at the first locus by A_1/A_2 and the alleles at the second locus by B_1/B_2 . Consider the case that the father of the nuclear family is heterozygous at both loci, whereas the mother is homozygous at both loci. In this situation, there are two possible phases for the father: either he could have phase A_1B_1/A_2B_2 or he could have phase A_1B_2/A_2B_1 . Suppose the father's phase is A_1B_1/A_2B_2 , it is possible to determine whether each offspring is recombinant or nonrecombinant based on the genotype data, say we observe r_i recombinants out of N_i offspring ($N_i - r_i$ nonrecombinants). However, if the father's phase is actually A_1B_2/A_2B_1 , then with the same offspring data, these N_i offspring would have $N_i - r_i$ recombinants and r_i nonrecombinants. Because the two phases are equally likely in the absence of LD, the likelihood of the i th pedigree is given by

$$L_i(\theta) = 0.5\theta^{r_i}(1-\theta)^{N_i-r_i} + 0.5(1-\theta)^{r_i}\theta^{N_i-r_i}, \quad \theta \in \left[0, \frac{1}{2}\right] \quad (5)$$

The LOD score in equation (2) and the maximum LOD score can then be computed. The LRT statistic can be compared to the mixture distribution in equation (4) to obtain P values.

Locus Heterogeneity

It is possible that, in a proportion (say λ) of the pedigrees the two genetic loci are linked with a $\theta < 0.5$, and in others the two loci are not linked, which corresponds to $\theta = 0.5$. However, it is not known which pedigrees have linkage and which do not. Then the likelihood for a given pedigree may be written as

$$L_i(\lambda, \theta) = \lambda L_i(\theta) + (1 - \lambda) L_i\left(\frac{1}{2}\right), \quad \theta \in [0, 0.5] \quad (6)$$

where $L_i(\theta)$ can be either the phase-known or the phase-unknown likelihood in equation (3) or (5), or can be other more complicated pedigree likelihood functions. The corresponding LOD score is sometimes called *admixture LOD score* or *heterogeneity LOD score* [1, 8, 9]. Complex diseases often exhibit locus heterogeneity so that alleles at more than one locus confer susceptibility to the disease. Different pedigrees can have different alleles responsible for the same disease [1, 2]. The admixture likelihood in equation (6) is obviously the simplest case. One can detect linkage in the presence of locus heterogeneity using the LRT based on the admixture likelihood:

$$LR = 2\sum_i [\log L_i(\hat{\lambda}, \hat{\theta}) - \log L_i(1, 0.5)] \quad (7)$$

where $\hat{\lambda}$ and $\hat{\theta}$ are the MLEs of the mixture proportion λ and the recombination fraction θ . The asymptotic null distribution of the LRT in this case is nonstandard because under the null hypothesis λ is unidentifiable [9, 10].

Complicated Cases

In the above, we have been dealing with simple situations where the likelihood can be written down by inspecting the haplotypes passed from parents to offspring. However, various complications generally make such a simple approach impossible. For example, some key individuals may be unavailable for study and the trait locus is unknown and has to be estimated. Parametric LOD score can still be written down by assuming some *penetrance* models and other parameters. Penetrance models are used to quantify the probability of the observed phenotype conditional on each of the possible genotypes at the trait locus. For example, consider a disease locus with two alleles, Q and q . For a fully penetrant recessive disease,

we have

$$\Pr(\text{disease}|QQ) = 1 \quad (8)$$

and

$$\Pr(\text{disease}|qq) = \Pr(\text{disease}|Qq) = 0 \quad (9)$$

The LOD score method has been widely used for linkage analysis where the mode of inheritance of the loci must be specified. Under some conditions, however, several modes of inheritance appear plausible, and/or the penetrance may not be 0 or 1. Hence, non-parametric methods that require no assumptions about the mode of inheritance might be more appropriate for such situations which is common for mapping complex traits [2].

Identity-by-Descent Analysis

If a trait locus and a marker locus are linked, then pairs of relatives who have similar trait values should share more alleles *identity by descent* (IBD) (i.e., direct copies of the same parental allele), and relative pairs with different trait values should share fewer allele IBD, than expected by chance alone. Early studies often evaluated marker IBD relationships of sibling pairs, and this was later extended to pairs of other types of relatives. Thus these methods are often referred to as *sib-pair*, or *relative-pair* methods. Next we review methods for the evaluation of IBD-sharing probabilities.

Data of full siblings (same father and same mother) are often studied. If the siblings' parents are also typed for a marker, or if a large number of siblings are typed to permit the parental marker genotypes to be deduced, it is often possible to count the number of marker alleles that a sib pair shares IBD. There are also situations where for some sib pairs it is impossible to actually count the number of marker alleles shared IBD. Haseman and Elston [11] proposed to estimate the IBD allele sharing probabilities at a marker locus by utilizing the marker information available on the siblings and their parents. Let f_i be the prior probability that a pair of siblings shares, by virtue of their relationship alone, i alleles IBD. Thus for full siblings, $f_0 = 1/4$, $f_1 = 1/2$, $f_2 = 1/4$. Then by Bayes' theorem, the posterior probability that the siblings share i alleles IBD, given

the available marker data information I_m , is simply

$$\hat{f}_i = \frac{[f_i P(I_m | \text{the siblings share } i \text{ alleles IBD})]}{P(I_m)} \quad (10)$$

where $P(I_m) = \sum_{k=0}^2 f_k P(I_m | \text{the siblings share } k \text{ alleles IBD})$.

Haseman and Elston [11] considered only the availability of marker information on a pair of full sibs and up to two parents, but the same general expression can be used when other relatives are present, and pairs of relatives other than siblings can be easily accommodated [12]. In general, these estimates $\hat{f}_i$ depend on accurate estimates of the marker genotype relative frequencies, but, provided both parents and the sibs are genotyped, they are independent of the genotype frequencies. The proportion of marker alleles that a sib pair shares IBD, π (which in fact can take on only the values 0, 1/2, or 1), can then be estimated by $\hat{\pi} = \hat{f}_2 + (1/2)\hat{f}_1$. The values of $\hat{\pi}$ for different pairs of sibs in a sibship of more than two members are pairwise independent [13, 14].

Kruglyak and Lander [15] showed how the IBD allele-sharing probabilities can be estimated in a multipoint fashion with greater accuracy at each marker location and how they can be estimated at any intervening point along the genome. The algorithm has been implemented in the computer program package MAPMAKER/SIBS. The computational burden for exact multipoint IBD calculation is considerable even in nuclear families. The computation of the Elston–Stewart [16] algorithm increases exponentially with the markers. For the Lander–Green hidden Markov model used in Kruglyak *et al.* [17] to estimate the multipoint IBD, the computation increases exponentially with the number of nonfounders in a pedigree.

Fulker *et al.* [18] derived an alternative multipoint approximation algorithm for estimating π_Q , the proportion of the alleles IBD at any location Q for a pair of siblings, from the single point estimates of IBD at a set of m marker loci on the same chromosome. Their method is based on a multiple linear regression of π_Q on the m estimates $\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_m$ for m marker loci, using the fact that for any two loci i and j with recombination fraction θ_{ij} between them, $E[\text{Cov}(\hat{\pi}_i, \hat{\pi}_j)] = 8V(\hat{\pi}_i)V(\hat{\pi}_j)(1 - 2\theta_{ij})^2$. Their regression equation provides multipoint estimates of π_Q at any location Q other than the locations for which marker information is available; the estimates at those locations are the m

estimates themselves. The method has been shown to be as effective as MLE of the exact multipoint IBD distribution [19].

Almasy and Blangero [20] extended the sib-pair multipoint mapping approach of Fulker *et al.* [18] to general relative pairs, which allows multipoint analysis in pedigrees of unlimited size and complexity. This multipoint IBD method uses the proportion of alleles shared IBD at genotyped loci to estimate IBD sharing at arbitrary point along a chromosome for each relative pair. They derived correlations in IBD sharing as a function of chromosomal distance for relative pairs in general pedigree and provided a simple framework whereby these correlations can be easily obtained for any relative pair by a single line of descent or by multiple independent lines of descent. The multipoint algorithm has been implemented in the software SOLAR.

Model-Free Analysis of Qualitative Traits

The principal advantage of IBD-sharing model-free methods is that one does not need to specify, *a priori*, a detailed inheritance model for the trait of interest, which is often unknown for complex phenotypes of interest. We now present some examples of IBD-based model-free analysis of linkage for qualitative traits.

Mean Test for Affected Sib Pairs (ASP)

The data for affected sib-pair (ASP) methods are the frequencies with which two affected offspring share copies of the same parental marker allele IBD. Under the hypothesis of no linkage between a trait locus and a marker locus, it is expected that the probabilities of sharing 0, 1, or 2 alleles IBD are 0.25, 0.5, and 0.25, respectively. If there are two affected siblings in each family, let N be the total number of affected sib pairs and T be the total number of alleles shared by the N sib pairs IBD. Then the *mean test* is based on the mean proportion of marker alleles shared IBD by the ASP:

$$Z = \frac{(T - N)}{\sqrt{\frac{N}{2}}} \quad (11)$$

which approximately has a standard normal distribution $N(0, 1)$ for large N . The mean test is more

powerful than many alternative tests based on IBD sharing as found in [14]. Interestingly, for ASP data, for recessive disease mode, Knapp *et al.* [21] proved that the one-sided mean test is the uniformly most powerful test assuming no LD.

A Maximum LOD Score Method

A maximum LOD score (MLS) method was introduced by Risch [22] for assessing the significance of linkage based on IBD sharing. Let α_i be the prior probability that two relatives share i alleles IBD, Z_i be the posterior probability that two relatives share i marker alleles IBD given that they are both affected ($i = 0, 1, 2$) (i.e., $Z_i = P(\text{IBD}_m = i | \text{both affected})$), and let w_{ij} be the probability of the observed marker data of the j th pair (and possibly other relevant relatives) given that they share i marker alleles IBD, $i = 0, 1, 2$. That is $w_{ij} = P(\text{marker data of the } j\text{th pair} | \text{IBD}_m = i)$. The likelihood for the observed data on the j th pair is $L_j = \sum_{i=0}^2 Z_i w_{ij}$. Then the likelihood ratio for the N independent pairs is given by $\Lambda = \prod_{j=1}^N (\sum_{i=0}^2 Z_i w_{ij}) / (\sum_{i=0}^2 \alpha_i w_{ij})$. Then $T = \log_{10} \Lambda$ can be interpreted as an LOD score. The MLS can be obtained by maximizing with respect to Z_i , $i = 0, 1, 2$. Analogous to the criterion of an MLS > 3 in model-based linkage analysis [3], the same criterion may be applied to $T = \log_{10} \Lambda$.

The above MLS method is based on IBD sharing and is thus model-free, but it can be parameterized in terms of the proportion of marker alleles shared IBD, making them amenable to likelihood-type analysis. Next we present another model-free method, which is also in the form of an LRT.

Linkage Analysis Using Both IBD and LD

When a disease mutation arises in a population, the mutant allele is in LD with other alleles. The LD is attenuated in successive generations owing to recombinations, but this happens very slowly for loci that are very close to the disease locus. With the numerous genetic markers (e.g., the single nucleotide polymorphisms (SNPs)) that are densely spread along the genome currently available, some markers are likely to be very close to the disease locus and thus exhibit LD that may be used for fine mapping. However, allelic association might be due to population admixture, which can be a confounding factor for

gene mapping [23]. One way to overcome population admixture is to type markers on individuals with disease and their two parents to determine whether particular parental alleles are more often transmitted to the affected offspring rather than not transmitted, which is called the *transmission/disequilibrium test* (TDT) [24]. In essence, the linkage tests based on IBD sharing (e.g., the mean test) use recombination information in the current pedigree, whereas the linkage tests based on LD (e.g., the TDT) utilize historical recombination information. It is natural to consider combining these two approaches and utilizing both sources of recombination information. To this end, a marginal likelihood approach based on transmission of marker alleles has been developed recently in [25–27]. The method extends the maximum-binomial likelihood method of [28], which was designed to use all affected siblings for linkage analysis in a natural way. For the special case of affected sib-pair data, the score test of the marginal likelihood is the mean test in the absence of LD, and is the sum of the mean test and the TDT in the presence of LD. This method has better power than the mean test in the presence of strong LD, and has better power than the TDT when LD is weak. Thus it has good power regardless of the level of LD, which is good because LD is typically unknown and quite variable along the genome. This method is a valid model-free linkage test while some people have argued that the TDT is not a valid linkage test for various reasons, including that it has no power in the absence of LD.

Analysis of Quantitative Traits

Many traits such as height, blood pressure, glucose levels, and cholesterol levels can be quantified, with the resulting measurements being distributed in a continuous fashion across a population. Practical applications of quantitative-trait linkage analysis are increasing rapidly because of the superior information content of quantitative traits. Traditionally, the linkage analysis for general quantitative traits have been done using the regression approach of Haseman and Elston [11] based on IBD sharing of sib pairs. When trait values have an approximate normal distribution, the variance-component method is very useful and can use IBD sharing of general pedigrees [20]. Next we give an overview of both approaches.

Haseman–Elston Regression Methods and Extensions

The idea of the method proposed by Haseman and Elston [11] was to take pairs of siblings and regress the squared differences in their trait values on their IBD sharing at a marker. More precisely, let the i th sibling pair have trait values $(Y_{i,1}, Y_{i,2})$, and define the squared trait difference as $\hat{Y}_i^D = (Y_{i,1} - Y_{i,2})^2$. Summarize the estimated mean IBD sharing at the locus for the pair as $\hat{\pi}_i$. Perform a simple linear regression of $\hat{Y}_i^D$ on $\hat{\pi}_i$ using least squares. Under the null hypothesis of no linkage, the regression slope is zero. Under the alternative hypothesis that the locus is linked to the trait, high proportions of IBD sharing (large $\hat{\pi}_i$) should be associated with a small difference in trait values (small $\hat{Y}_i^D$). Because of this expected negative correlation between $\hat{Y}_i^D$ and $\hat{\pi}_i$, the regression slope is negative. Linkage can be tested with a one-sided t test of the regression slope estimate using the least-squares method. For example, assume a single quantitative trait locus (QTL) with two alleles Q and q . For an observed trait value Y , the genetic model may be written as follows:

$$Y = \mu + \gamma + e \quad (12)$$

where μ is an overall mean, γ is the effect of the QTL, and e is a residual effect. Suppose the two alleles Q and q have frequencies p_Q and $(1 - p_Q)$, respectively. The genotype-specific means of the trait are

$$\mu_{QQ} = \mu + a, \quad \mu_{Qq} = \mu + d, \quad \mu_{qq} = \mu - a \quad (13)$$

The total genetic variance is σ_γ^2 , which can be decomposed into two parts corresponding to additive effect and dominant effect, $\sigma_\gamma^2 = \sigma_a^2 + \sigma_d^2$, where the variances of the additive effect is $\sigma_a^2 = 2p_Q(1 - p_Q)[a + (1 - 2p_Q)d]^2$, and the variance of the dominant effect is $\sigma_d^2 = [2p_Q(1 - p_Q)d]^2$. Let θ denote the recombination fraction of the QTL and marker loci. Haseman and Elston [11] showed that for the special case of nondominance (i.e., only additive genetic variance) at the trait loci,

$$E(\hat{Y}^D | I_m) = \alpha_s + \beta_s \hat{\pi}_i \quad (14)$$

where $\beta_s = -2(1 - \theta)\sigma_\gamma^2$. It is clear that when $\theta = 0.5$, or $\sigma_\gamma = 0$, then $\beta_s = 0$; otherwise, $\beta_s < 0$. Thus, a one-sided test of linkage can be used. Several

research groups [29–35] have proposed approaches for improving the power of this regression method by use of information in both the squared trait difference $\hat{Y}_i^D$ and the squared trait sum $\hat{Y}_i^S = (Y_{i,1} + Y_{i,2})^2$. In particular, Wright [29] pointed out that Haseman and Elston's choice of $\hat{Y}_i^D$ discards some useful trait information contained in $\hat{Y}_i^S$. Drigalenko [30] showed that regression of $\hat{Y}_i^D$ on $\hat{\pi}_i$ and $\hat{Y}_i^S$ on $\hat{\pi}_i$ produces separate estimates of the same slope. Using average of the two resulting slope estimates is shown [30] to be equivalent to performing a single regression using the mean-corrected trait product, $\hat{Y}_i^P = [(Y_{i,1} - \mu)(Y_{i,2} - \mu)]$, as the dependent variable. Elston *et al.* [31] further developed the idea of regression based on $\hat{Y}_i^P$ and expanded it to consider larger sibships, covariates, and other complexities. Xu *et al.* [32], Forrest [33], and Sham and Purcell [34] suggested to improve the power of this regression method using different weighting of the two slope estimates. Visscher and Hopper [35] suggested using a combination of $\hat{Y}_i^D$, $\hat{Y}_i^S$ and the trait correlation between the pair of siblings as response variable and regressing on $\hat{\pi}_i$. Others have proposed score statistics to improve the power of this regression method [36–38].

All these methods are largely limited to sibships. Sham *et al.* [39] offered a regression method for general pedigrees. The idea is to reverse the Haseman–Elston paradigm by regressing the IBD sharing on an appropriate function of the quantitative-trait values. The actual regression that is suggested is a multivariate (not multiple) regression, within each family, of the pairwise IBD sharing scores on the pairwise squared sums and squared differences. This method is computationally manageable and has been implemented in the software package called *Merlin* [40]. Feingold [41] provides an excellent overview of regression-based QTL mapping methods.

Variance-Component Methods and Extensions

An alternative approach that simultaneously examines all pedigree relationships has also been developed from the classical variance-component analysis. The classical technique simply separates the total variance in the trait values into components due to genetic and environment effects [42]. Hopper and Matthews [43] first suggested adapting the method to linkage analysis by modeling an additional variance component

for a hypothesized QTL near a marker locus. Linkage to the locus is indicated by a statistically significant nonzero value for the QTL component. The earliest version of this method was based on analysis of only one or two markers at a time [44–46]. Goldgar [44] proposed a variance-component approach for analyzing nonexperimental sibships. Amos [46] developed a mixed-effects variance-component model for evaluating covariate effects, as well as evidence for genetic linkage to a single trait-affecting locus for pedigree data. Almasy and Blangero [20] improved on this approach by extending the approach to arbitrary pedigree and using an approximation to a multipoint algorithm.

Let the quantitative phenotypic value, Y , of an individual be written as a linear function of the n QTLs that influence it:

$$Y = \mu + \sum_{i=1}^n \gamma_i + e \quad (15)$$

where μ is the grand mean, γ_i is the random effect of the i th QTL, and e represents a random environmental effect. Assume $\{\gamma_i\}$ and e are uncorrelated with mean zero and so that the variance of Y is

$$\sigma_y^2 = \sum_{i=1}^n \sigma_{\gamma_i}^2 + \sigma_e^2 \quad (16)$$

When one allows for both additive and dominate effects, $\sigma_{\gamma_i}^2 = \sigma_{ai}^2 + \sigma_{di}^2$, where σ_{ai}^2 is the additive genetic variance due to the i th locus, and σ_{di}^2 is the dominance variance.

The covariance between the phenotype of different individuals $\{Y_j\}$ depends on the relatedness and the pedigree structure. It is determined by the IBD-sharing probabilities and the component variances σ_{ai}^2 and σ_{di}^2 . For the above random-effects model, the phenotypic covariance between the trait values of any pair of relative can be obtained as

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &= E(Y_1 - \mu)(Y_2 - \mu) \\ &= \sum_{i=1}^n \left[\left(\frac{k_{1i}}{2 + k_{2i}} \right) \sigma_{ai}^2 + k_{2i} \sigma_{di}^2 \right] \end{aligned} \quad (17)$$

where k_{ji} is the i th QTL-specific probability of the pair of relatives sharing j alleles IBD. Let $\hat{\pi}_i = (k_{1i}/2 + k_{2i})$ be the coefficient of relationship or the probability of a random allele being shared IBD

at the i th QTL. When $n = 1$, the above model is reduced to the simple model in equation (12). The $\hat{\pi}_i$ and k_{2i} coefficients and their expectations effectively structure the expected phenotypic variance, and are the basis for much of the quantitative linkage analysis just as in the sib-pair difference method of Haseman and Elston [11] discussed above. For any given chromosomal location, $\hat{\pi}_i$ and k_{2i} coefficients can be estimated from genetic marker data I_m , e.g., using the software SOLAR developed in [20] or other software. By assuming multivariate normality as a working model for the trait distribution within pedigrees, the pedigree likelihood can be easily written down, which is totally determined by the mean and the covariance parameters. Numerical algorithms can be used to estimate the variance-component parameters. The likelihood-ratio test or the LOD scores can be used to test for linkage [20].

Frequently, only a few QTLs are examined at a time, and one can reduce the number of parameters in equation (15) that are to be considered. Let $\phi = (1/2)E(k_{1i}/2 + k_{2i})$ denote the expected kinship coefficient over the genome with $2\phi = R$ giving the expected coefficient of relation, and $\delta_\tau = E[k_{2i}]$ denote the expected probability of sharing two alleles IBD. Using 2ϕ to approximate $(k_{1i}/2 + k_{2i})$ and δ_τ for k_{2i} yields an approximation of the $\text{Cov}(Y_1, Y_2)$ as follows:

$$\text{Cov}(Y_1, Y_2) \approx 2\phi \sum_{i=1}^n \sigma_{ai}^2 + \delta_\tau \sum_{i=1}^n \sigma_{di}^2 \quad (18)$$

For example, if one is focusing on the analysis of the i th QTL in equation (15), one can absorb the effects of all other remaining QTLs in one random effect Γ_i term at unlinked loci. Then the expected phenotypic covariance between relatives is well approximated by

$$\text{Cov}(Y_1, Y_2) = \hat{\pi}_i \sigma_{ai}^2 + k_{2i} \sigma_{di}^2 + 2\phi \sigma_\Gamma^2 + \delta_\tau \sigma_d^2 \quad (19)$$

where σ_Γ^2 represents the additive genetic variance besides that of the i th QTL of interest, and σ_d^2 represents the residual dominant genetic variance. One can further replace μ by $\mu + \beta^T \mathbf{X}$, i.e., adding a fixed effect term to incorporate covariates, where $\mathbf{X}$ is a vector of directly observable covariates and β is a vector of regression coefficients. Then the mixed-effects model for the phenotypic value can be written as

$$Y = \beta^T \mathbf{X} + \gamma + \Gamma + e \quad (20)$$

This model includes covariates fixed effect, a major gene effect, random polygenic effects, and random environmental effects. This variance-component-based model can be used for mapping QTL in general human pedigrees and has good power when the trait values have an approximate normal distribution. This variance-component approach has also been extended by de Andrade *et al.* [47] to the context of longitudinal studies where quantitative traits are measured repeatedly over time, such as in the Framingham Heart Study [48]. For such data, in addition to the correlation among pedigree members, there is also within-subject correlation among the repeated measures of the same individual over time.

A practical concern of the variance-component approach is the robustness with respect to the normality assumption for the trait values. Violation of the normality assumption might have serious adverse effects on both type I errors and power [49–51]. In this case, the variance-component approach is less robust than Haseman–Elston regression method [52] and the distribution-free method suggested in [15]. One might consider using generalized estimating equations for estimating variance components [46, 49]. Another natural strategy is to perform a parametric transformation $H(\cdot)$, such as the log-transformation or square-root transformation on the trait values to approximate normality [48, 52, 53]. However, it can be hard to find a generally useful parametric transformation, and different transformations might lead to conflicting results.

Diao and Lin [54] proposed a novel extension of the above variance-component model that allows the true transformation function $H(\cdot)$ to be completely unspecified:

$$H(Y) = \beta^T \mathbf{X} + \gamma + \Gamma + e \quad (21)$$

Diao and Lin [54] also present efficient likelihood-based procedures to estimate variance components and to test for genetic linkage. The method allows choice of software to estimate the multipoint IBD allele-sharing probabilities. Thus one may choose appropriate software according to the size and complexity of the pedigrees as well as the number of markers. GENEHUNTER [17] and ACT [46] perform multipoint calculation that are based on hidden Markov models [55] and can handle an arbitrary number of markers for small pedigrees, whereas the approximation method implemented in SOLAR [20]

can handle large pedigrees. The above semiparametric transformation model for variance component in equation (21) provides great flexibility on the trait distribution. However, it still requires that the phenotype is fully observed and has a known distribution after some unknown transformation. These assumptions may not be satisfied when the phenotype pertains to the survival time or age of onset, which has a skewed distribution and is usually subject to censoring due to random loss of follow-up or limited duration of the experiment. In addition, the QTL may only affect the survival of a latent subpopulation. An extension to censored trait data is given in [56] for human pedigrees. For experimental crosses, semiparametric survival methods using Cox's regression model for interval mapping of QTL with time-to-event phenotypes under random censoring have also been developed in [57, 58], and others.

Discussion

With the wide range of methodologies currently available, it is natural to ask which method should be used for the given data. There is no simple answer to this question. Choice of methods may depend on many considerations including the statistical design, size and complexity of the pedigrees, the phenotypes (qualitative or quantitative), current knowledge of the pattern of trait inheritance, etc. In general, the desirable characteristics of a good linkage analysis method may include high statistical power, correct significance level (validity), and certain robustness [4, 41].

In terms of statistical power, likelihood-based analysis is the benchmark when the genetic model is correctly specified. The nonparametric methods (e.g., using score tests) are often designed to emulate this power optimality. On the other hand, the model-free methods are often relatively easy to compute besides being more robust than parametric methods.

Determining appropriate genome-wide significant levels is critical for whole genome scans, which are becoming the commonly used designs [59, 60]. Using liberal genome-wide significance levels can lead to an abundance of false discoveries whereas using overly conservative levels would diminish the statistical power and lead to many lost opportunities. Much progress has been made in developing useful statistical methods to evaluate genome-wide significance levels. These include using maximums of the

Gaussian process as in Lander and Botstein [61] for mapping QTL in a backcross design and as in Feingold *et al.* [62] for human genetics with IBD-sharing statistics. Lin [63] discussed a resampling (or random weighting) scheme that is useful for finding genome-wide significance levels for a general class of score-type test statistics, which can be used in a variety of genome-wide studies.

In terms of effective sampling designs for genome scan, Elston [64] and Elston *et al.* [65] proposed a two-stage procedure in which at the first stage affected pairs of relatives are typed for widely spaced markers and a relatively large significance level is used to determine where, at a second stage, more narrowly spaced markers should be placed around the markers that were significant at the first stage. Statistical efficiency is gained because a tight grid of markers is typed only at locations presenting evidence that a trait locus might occur. Further developments can be found in [60, 66].

Motivated by mapping complex traits, and spurred by the recent successes of various genome projects and rapid advancements of biotechnology, enormous progress in developing new linkage analysis methods has been made, which remains an active area of current research. Owing to page limitations, this article provides a very brief overview of the likelihood-based (LOD score) parametric methods and the nonparametric methods based on IBD sharing. For in-depth overviews of many other aspects of linkage analysis, interested readers may consult the classical textbook by Ott [1], and a large collection of excellent overview articles contained in Volume 42 of *Advance in Genetics* (2001).

References

- [1] Ott, J. (1999). *Analysis of Human Genetic Linkage*, 3rd Edition, Johns Hopkins University Press.
- [2] Lander, E.S. & Schork, N.J. (1994). Genetic dissection of complex traits, *Science* **265**, 2037–2048.
- [3] Morton, N.E. (1955). Sequential tests for the detection of linkage, *American Journal of Human Genetics* **7**, 277–318.
- [4] Elston, R.C. (1998). Methods of linkage analysis - and the assumptions underlying them, *American Journal of Human Genetics* **63**, 931–934.
- [5] Hodge, S.E. (2001). Model-free vs. model-based linkage analysis: a false dichotomy? *American Journal of Medical Genetics* **105**, 62–64.

- [6] Lander, E.S. & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results, *Nature Genetics* **11**, 241–247.
- [7] Terwilliger, J.D. & Ott, J. (1994). *Handbook of Human Genetic Linkage*, Johns Hopkins University Press.
- [8] Smith, C. (1963). Testing for heterogeneity of recombination fraction value in human genetics, *Annals of Human Genetics* **27**, 175–181.
- [9] Abreu, P., Hodge, S.E. & Greenberg, D.A. (2002). Quantification of type I error probabilities for heterogeneity lod scores, *Genetic Epidemiology* **22**, 156–169.
- [10] Liu, X. & Shao, Y. (2003). Asymptotics for likelihood ratio test under loss of identifiability, *Annals of Statistics* **31**, 807–832.
- [11] Haseman, J.K. & Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics* **2**, 3–19.
- [12] Amos, C., Dawson, D.V. & Elston, R.C. (1990). The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees, *American Journal of Human Genetics* **47**, 842–853.
- [13] Hodge, S.E. (1984). The information contained in multiple sibling pairs, *Genetic Epidemiology* **1**, 109–122.
- [14] Blackwelder, W.C. & Elston, R.C. (1985). A comparison of sib-pair linkage tests for disease susceptibility loci, *Genetic Epidemiology* **2**, 85–97.
- [15] Kruglyak, L. & Lander, E.S. (1995). Complete multipoint sib pair analysis of qualitative and quantitative traits, *American Journal of Human Genetics* **57**, 439–454.
- [16] Elston, R.C. & Stewart, J. (1971). A general model for the analysis of pedigree data, *Human Heredity* **21**, 523–542.
- [17] Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach, *American Journal of Human Genetics* **58**, 1347–1363.
- [18] Fulker, D.W., Cherny, S.S. & Cardon, L.R. (1995). Multipoint interval mapping of quantitative trait loci, using sib pairs, *American Journal of Human Genetics* **56**, 1224–1233.
- [19] Fulker, D.W. & Cherny, S.S. (1996). An improved multipoint sib-pair analysis of quantitative traits, *Behavior Genetics* **26**, 527–532.
- [20] Almasy, L. & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees, *American Journal of Human Genetics* **62**, 1198–1211.
- [21] Knapp, M., Seuchter, S.A. & Baur, M.P. (1994). Linkage analysis in nuclear families 2: relationship between affected-sib-pair tests and lod score analysis, *Human Heredity* **44**, 44–51.
- [22] Risch, N. (1990). Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs, *American Journal of Human Genetics* **46**, 242–253.
- [23] Rabinowitz, D. (2002). Adjusting for population heterogeneity and misspecified haplotype frequencies when testing nonparametric null hypotheses in statistical genetics, *Journal of the American Statistical Association* **97**, 742–758.
- [24] Spielman, R.S., McGinnis, R.E. & Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *American Journal of Human Genetics* **52**, 506–516.
- [25] Huang, J. & Jiang, Y. (1999). Linkage detection adaptive to linkage disequilibrium: the disequilibrium-likelihood-binomial test for affected-sibship data, *American Journal of Human Genetics* **65**, 1741–1759.
- [26] Lo, S.H., Liu, X. & Shao, Y. (2003). A marginal likelihood model for family-based data, *Annals of Human Genetics* **67**, 357–366.
- [27] Shao, Y. (2005). Adjustment for transmission heterogeneity in mapping complex genetic diseases using mixture models and score tests, *Proceedings (American Statistical Association)* **2005**, 383–393.
- [28] Abel, L., Alcais, A. & Maller, A. (1998). Comparison of four sib-pair linkage methods for analyzing sibships with more than two affected: interest of the binomial-maximum-likelihood approach, *Genetic Epidemiology* **15**, 371–390.
- [29] Wright, F.A. (1997). The phenotypic difference discards sib-pair QTL linkage information, *American Journal of Human Genetics* **60**, 740–742.
- [30] Drigalenko, E. (1998). How sib pairs reveal linkage, *American Journal of Human Genetics* **63**, 1242–1245.
- [31] Elston, R.C., Buxbaum, S., Jacobs, K.B. & Olson, J.M. (2000). Haseman and Elston revisited, *Genetic Epidemiology* **19**, 1–17.
- [32] Xu, X., Weiss, S., Xu, X. & Wei, L.J. (2000). A unified Haseman-Elston method for testing linkage with quantitative traits, *American Journal of Human Genetics* **67**, 1025–1028.
- [33] Forrest, W. (2001). Weighting improves the “New Haseman-Elston” method, *Human Heredity* **52**, 47–54.
- [34] Sham, P.C. & Purcell, S. (2001). Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs, *American Journal of Human Genetics* **68**, 1527–1532.
- [35] Visscher, P.M. & Hopper, J.L. (2001). Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data, *Annals of Human Genetics* **65**, 583–601.
- [36] Tang, H.-K. & Siegmund, D. (2001). Mapping quantitative trait loci in oligogenic models, *Biostatistics* **2**, 147–162.
- [37] Putter, H., Sandkuijl, L.A. & van Houwelingen, J.C. (2002). Score test for detecting linkage to quantitative traits, *Genetic Epidemiology* **22**, 345–355.
- [38] Wang, K. & Huang, J. (2002). A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size, *American Journal of Human Genetics* **70**, 412–424.
- [39] Sham, P.C., Purcell, S., Cherny, S.S. & Abecasis, G.R. (2002). Powerful regression-based quantitative-trait

- linkage analysis of general pedigrees, *American Journal of Human Genetics* **71**, 238–253.
- [40] Abecasis, G., Cherny, S., Cookson, W. & Cardon, L. (2002). Merlin: rapid analysis of dense genetic maps using sparse gene flow trees, *Nature Genetics* **30**, 97–101.
- [41] Feingold, E. (2002). Regression-based quantitative-traitlocus mapping in the 21st century (Invited Editorial), *American Journal of Human Genetics* **71**, 217–222.
- [42] Lange, K., Westlake, J. & Spence, M.A. (1976). Extensions to pedigree analysis. III. Variance components by the scoring method, *Annals of Human Genetics* **39**, 485–491.
- [43] Hopper, J.L. & Mathews, J.D. (1982). Extensions to multivariate normal models for pedigree analysis, *Annals of Human Genetics* **46**, 373–383.
- [44] Goldgar, D.E. (1990). Multipoint analysis of human quantitative genetic variation, *American Journal of Human Genetics* **47**, 957–967.
- [45] Schork, N.J. (1993). Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations, *American Journal of Human Genetics* **53**, 1306–1319.
- [46] Amos, C.I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees, *American Journal of Human Genetics* **54**, 535–543.
- [47] de Andrade, M., Gueguen, R., Visvikis, S., Sass, C., Siest, G. & Amos, C.I. (2002). Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis, *Genetic Epidemiology* **22**, 221–232.
- [48] Geller, F., Dempfle, A. & Gorg, T. (2003). Genome scan for body mass index and height in the Framingham Heart Study, *BMC Genetics (Suppl)* **4**, S91.
- [49] Amos, C.I., Zhu, D.K. & Boerwinkle, E. (1996). Assessing genetic linkage and association with robust components of variance approaches, *Annals of Human Genetics* **60**, 143–160.
- [50] Allison, D.B., Neale, M.C., Zannolli, R., Schork, N.J., Amos, C.I. & Blangero, J. (1999). Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci mapping procedure, *American Journal of Human Genetics* **65**, 531–544.
- [51] Feingold, E. (2001). Methods for linkage analysis of quantitative trait loci in humans, *Theoretical Population Biology* **60**, 167–180.
- [52] Allison, D.B., Fernández, J.R., Heo, M. & Beasley, T.M. (2000). Testing the robustness of the new Haseman-Elston quantitative-trait loci mapping procedure, *American Journal of Human Genetics* **67**, 249–252.
- [53] Strug, L., Sun, L. & Corey, M. (2003). The genetics of cross-sectional and longitudinal body mass index, *BMC Genetics (Suppl)* **4**, S14.
- [54] Diao, G. & Lin, D.Y. (2005). A powerful and robust method for mapping quantitative trait loci in genetic pedigrees, *American Journal of Human Genetics* **77**, 97–111.
- [55] Lander, E.S. & Green, P. (1987). Construction of multilocus genetic linkage maps in humans, *Proceedings of the National Academy of Sciences USA* **84**, 2363–2367.
- [56] Diao, G. & Lin, D.Y. (2006). Semiparametric variance-component models for linkage and association analysis of censored trait data, *Genetic Epidemiology* **30**, 570–581.
- [57] Diao, G. & Lin, D.Y. (2005). Semiparametric methods for mapping quantitative trait loci with censored data, *Biometrics* **61**, 789–798.
- [58] Liu, M., Lu, W. & Shao, Y. (2006). Interval mapping of quantitative trait loci for time-to-event data with the proportional hazards mixture cure model, *Biometrics* **62**, 1053–1063.
- [59] Whittemore, A.S. (1996). Genome scanning for linkage: an overview, *American Journal of Human Genetics* **59**, 704–716.
- [60] Guo, X. & Elston, R.C. (2001). One-stage versus two-stage strategies for genome scans, *Advances in Genetics* **42**, 459–471.
- [61] Lander, E.S. & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics* **121**, 185–199.
- [62] Feingold, E., Brown, P.O. & Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high resolution maps of IBD, *American Journal of Human Genetics* **53**, 234–251.
- [63] Lin, D.Y. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies, *Bioinformatics* **21**, 781–787.
- [64] Elston, R.C. (1992). Designs for the global search of the human genome by linkage analysis, in *Proceedings of the XVth International Biometric Conference*, Hamilton, pp. 39–51, December 1992.
- [65] Elston, R.C., Guo, X. & Williams, L.V. (1996). Two-stage global search designs for linkage analysis using pairs of affected relatives, *Genetic Epidemiology* **13**, 535–558.
- [66] Ghosh, S. & Majumder, P.P. (2000). A two-stage variable stringency semi-parametric method for mapping quantitative trait loci with the use of genome-wide scan data on sib pairs, *American Journal of Human Genetics* **66**, 1046–1061.

Related Articles

Microarray Analysis

YONGZHAO SHAO

Logistic Regression

A common question faced by researchers in epidemiologic studies is whether a particular exposure (or exposures) is causally related to (*see* **Causality/Causation**) an outcome or disease process. In many studies, the outcome is a dichotomous variable – a variable that can be assigned one of two possible values, for example, presence/absence of disease or occurrence/nonoccurrence of an event. Logistic regression is the most commonly used statistical tool to answer such questions, especially, when the outcome is dichotomous. In this chapter, we describe the logistic model, explain estimation of the *odds ratio* (OR) (*see* **Odds and Odds Ratio**) using logistic regression, describe hypothesis testing and confidence interval (CI) estimation in the logistic regression framework, and finally, illustrate the use of logistic regression using data from an epidemiologic study.

The Logistic Model

The logistic regression model belongs to a family of statistical models known as *generalized linear models*. Logistic regression can be used to describe the relationship of several predictor/independent variables (usually denoted by $X_1, X_2, \dots, X_k$) to a dichotomous outcome/dependent variable (denoted by Y), which is typically coded as 1 or 0 for the two possible values that it can take. The logistic model, as shown in equation (1) below, describes the expected value of Y (denoted by $E(Y)$):

$$E(Y) = \frac{1}{1 + \exp \left[- \left(\beta_0 + \sum_{j=1}^k \beta_j X_j \right) \right]} \quad (1)$$

where $E(Y)$ is equivalent to the probability $\Pr(Y = 1)$ for the (0, 1) random variable Y , β_0 is the intercept in the model, and the β_j 's are the slope parameters for the independent X_j 's.

There are three broad steps involved in logistic regression analysis:

1. We first postulate a mathematical model on the basis of prior knowledge about and experience with the outcome under study that describes the

mean of outcome Y as a function of the independent variables X_j and their slope parameters β_j .

2. We then fit the model to the data obtained from the epidemiologic study using maximum likelihood (ML) procedures that enable us to estimate the model parameters (the β_0 and the β_j).
3. Finally, after assessing the fit of the model and relevant regression diagnostic indices, we make appropriate statistical inferences about measures of effect (typically *ORs*) that are derived from estimated model parameters.

There are two reasons for the popularity of the logistic model. Both reasons relate to the mathematical expression on the right side of the logistic regression formula in equation (1). It is evident that the right side expression in equation (1) is of the general mathematical form given in equation (2):

$$f(z) = \frac{1}{1 + e^{-z}} \quad (2)$$

where

$$z = \left(\beta_0 + \sum_{j=1}^k \beta_j X_j \right) \quad (3)$$

The function $f(z)$ is called the *logistic function*. The values of $f(z)$ range from 0 to 1 as z varies from $-\infty$ to $+\infty$. This property makes the logistic function very useful to model probability, and in epidemiologic studies, such a probability can be used to describe an individual's risk of getting the outcome.

Another reason for the popularity of the logistic function relates to its general shape as shown in Figure 1. A plot of the logistic function is in the form of a sigmoid curve which epidemiologists believe applies to a variety of disease conditions. In this context, the variable z can be thought of as a combination of various risk factors (the X_j), so that $f(z)$ represents the risk of the outcome for a given value of z . As seen from the figure, the risk is near zero for low z values, rises over a range of intermediate z values, and remains close to 1 once z gets large enough.

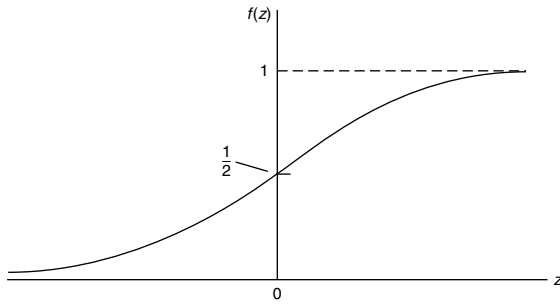


Figure 1 The shape of the logistic function $f(z) = \frac{1}{1+e^{-z}}$. Kleinbaum *et al.* [1] [From Applied Regression Analysis and Multivariable Methods 3rd edition by KLEINBAUM/KUUPER/MULLER/NIZAM, 1998. Reprinted with permission of Brooks/Cole, a division of Thomson Learning: www.thomsonrights.com.]

Estimating the Odds Ratio Using Logistic Regression

The purpose of regression models is to provide us with estimates of model parameters that can then be used to describe the relationship between the outcome variable (Y) and one or more predictors (X 's). Some of these predictors, often called *exposure variables* are of primary interest, whereas other predictors, often called *control variables* are not of primary interest, but are, nevertheless, considered in one's model to control for their possible effects on the outcome (see **Confounding**).

For the logistic regression model, quantification of these relationships typically involves a measure of effect known as the *odds ratio* (see **Odds and Odds Ratio**) that is derived from the parameter estimates. Briefly, an odds can be defined as the ratio of the probability of some event occurring to the probability of that event not occurring. An *OR*, by definition, is a ratio of two odds. For example, let us consider an epidemiologic study evaluating the occurrence of disease D in two groups of individuals A and B , one with and the other without an exposure of interest. Let $A = E$ denote the group with exposure E and $B = NE$ denote the group without the exposure E ; then D_E describes persons with exposure E that develop D and D_{NE} describes persons without exposure E that develop D . If $\Pr(D_E) = 0.25$ and $\Pr(D_{NE}) = 0.10$, the *OR* that compares the odds of developing disease D for those with exposure E with the odds of developing D for those without exposure E is given

by equation (4):

$$\begin{aligned} OR_{E \text{ vs } NE} &= \frac{\text{odds}(D_E)}{\text{odds}(D_{NE})} = \frac{\frac{\Pr(D_E)}{1 - \Pr(D_E)}}{\frac{\Pr(D_{NE})}{1 - \Pr(D_{NE})}} \\ &= \frac{0.25}{\frac{1 - 0.25}{0.10}} = \frac{1/3}{1/9} = 3 \end{aligned} \quad (4)$$

In other words, the odds of developing disease D in those with exposure E is three times the corresponding odds for those without exposure E , suggesting roughly that those with E have a threefold higher risk of developing D than those without E . An *OR* of one would mean that the odds for the two groups are the same; that is, exposure E does not affect the development of disease D .

Why is the *OR* the measure of effect when we do logistic regression analysis? To answer this question, let us consider another way of describing the logistic regression model, called the *logit form* of the model. The *logit* term is defined as the natural log odds of the event ($Y = 1$), and can be thought of as a transformation of the probability $\Pr(Y = 1)$. We can thus present the logistic model as shown in equation (5): E vs NE

$$\begin{aligned} \text{logit}[\Pr(Y = 1)] &= \log_e[\text{odds}(Y = 1)] \\ &= \log_e \left[\frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)} \right] \end{aligned} \quad (5)$$

If we then substitute the logistic model formula (equation (1)) for $\Pr(Y = 1)$ into equation (5), it follows that

$$\text{logit}[\Pr(Y = 1)] = \beta_0 + \sum_{j=1}^k \beta_j X_j \quad (6)$$

Equation (6) is called the *logit form* of the model. An attractive quality of the logit form is that it expresses the model as a linear combination of the independent variables and their parameters.

For example, if Y denotes disease D status (1 = yes, 0 = no) and there is only one (i.e., $k = 1$) predictor X_1 – say, exposure to E (1 = exposed to E , 0 = not exposed to E) – then the logistic model in logit form can be written as

$$\text{logit}[\Pr(Y = 1)] = \beta_0 + \beta_1 X_1 = \beta_0 + \beta_1 E \quad (7)$$

To obtain an expression for the *OR* from a logistic model, we must compare the odds for two groups of individuals. For the example given above involving only one independent variable, the two groups are those exposed to E (i.e., $E = 1$) and those not exposed to E (i.e., $E = 0$). The log odds for these two groups can be written as

$$\log_e \text{odds}(\text{exposed to } E) = \beta_0 + (\beta_1 \times 1) = \beta_0 + \beta_1 \quad (8)$$

and

$$\log_e \text{odds}(\text{not exposed to } E) = \beta_0 + (\beta_1 \times 0) = \beta_0 \quad (9)$$

It follows that the *OR* comparing exposed to unexposed is given by the following formula:

$$\begin{aligned} OR_{E \text{ vs } NE} &= \frac{\text{odds}(\text{exposed to } E)}{\text{odds}(\text{not exposed to } E)} \\ &= \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1) \end{aligned} \quad (10)$$

In other words, for the simple example involving one (0, 1) predictor, the *OR* comparing the two categories of the predictor is obtained by exponentiating the coefficient of the predictor in the logistic model, as shown in equation (10).

Generally, when computing an *OR*, we can define two groups (or individuals) that are to be compared in terms of two different specifications of the set of predictors ($X_1, X_2, \dots, X_k$). For example, let $\mathbf{X}_A = (X_{A1}, X_{A2}, \dots, X_{Ak})$ and $\mathbf{X}_B = (X_{B1}, X_{B2}, \dots, X_{Bk})$ denote the collection of X 's for groups (or individuals) A and B , respectively. On the basis of the logit form of the model shown in equation (6), the log odds for the two groups are

$$\log_e \text{odds for } \mathbf{X}_A = \beta_0 + \sum_{j=1}^k \beta_j X_{Aj} \quad (11)$$

and

$$\log_e \text{odds for } \mathbf{X}_B = \beta_0 + \sum_{j=1}^k \beta_j X_{Bj} \quad (12)$$

To obtain a general formula for the *OR*, we must divide the odds for group (or individual) A by the odds for group (or individual) B , which gives

us an expression involving the logistic regression parameters. Equation (13) shows the general *OR* formula:

$$\begin{aligned} OR_{X_A \text{ vs } X_B} &= \frac{\text{odds for } X_A}{\text{odds for } X_B} \\ &= \frac{\exp\left(\beta_0 + \sum_{j=1}^k \beta_j X_{Aj}\right)}{\exp\left(\beta_0 + \sum_{j=1}^k \beta_j X_{Bj}\right)} \\ &= \exp \sum_{j=1}^k \beta_j (X_{Aj} - X_{Bj}) \end{aligned} \quad (13)$$

The constant term β_0 in the logistic model (equation (1)) drops out of the *OR* expression (equation (13)), so the *OR* depends only on exponentiating the sum of the β_j coefficients multiplied by the difference in corresponding values of the X_j 's for groups (or individuals) A and B .

Equation (13) describes a “population” *OR* parameter because the β_j terms in this expression are themselves unknown population parameters. An estimate of this population *OR* can be obtained by fitting the logistic model using *ML* estimation and substituting the *ML* parameter estimates $\hat{\beta}_j$ (note the $\hat{\cdot}$ on β_j that is used to denote an estimated value of the unknown population parameter), together with values of X_{Aj} and X_{Bj} , into the formula shown in equation (13) to obtain a numerical value that estimates the *OR*.

In general, when the logistic model involves more than one independent (X_j) variable and only one variable (i.e., the exposure variable) changes, while the other (i.e., control) variables are held constant, we say that the *OR* comparing two categories of the changing variable is an *adjusted OR* that controls for the other variables in the model. When the exposure variable is a (0, 1) variable and there are no product terms involving the exposure variable and other (control) variable(s) in the model, equation (13) simplifies to the following:

$$OR_{X_A \text{ vs } X_B} = \exp(\beta_j(1 - 0)) = \exp(\beta_j) \quad (14)$$

where β_j is the exposure variable of interest. However, if the exposure variable is not a (0, 1) variable and/or there are product terms in the model involving the exposure variable (referred to as *interaction*

terms or effect modifiers – see **Effect Modification and Interaction**), then we cannot use equation (14) to get an estimate for the *OR*. In such situations, we use the general formula shown in equation (13) to obtain the *OR* estimate(s).

Hypothesis Testing and Confidence Interval Estimation

The estimated *OR* calculated above is known as the *point estimate* of the unknown population parameter – the true *OR*. Such point estimates have a certain amount of variability associated with them, as represented by the standard error estimates of the coefficients. This variability has to be considered when we make statistical inference on the parameters of interest. We can use two kinds of inference-making procedures. One is testing hypothesis about certain parameters, the other is deriving interval estimates of certain parameters.

Statistical inference-making requires three quantities, which can be readily obtained by the default output provided by most *ML* estimation programs in most statistical software bundles. These are the *maximized likelihood value*, the *estimated variance-covariance matrix* and a *listing of each variable followed by its ML estimate and standard error*. To illustrate how statistical inferences are made using these quantities, we consider two logit models, as shown in equations (15) and (16):

Model 1:

$$\text{logit}[\text{Pr}(Y = 1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (15)$$

Model 2:

$$\text{logit}[\text{Pr}(Y = 1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad (16)$$

There are two commonly used tests for testing hypothesis in logistic regression – the *likelihood ratio (LR) test* and the *Wald test*. The *LR* test is a comparison of the log likelihood statistic (defined as $-2 \log_e$ of the maximized likelihood value) of the two models that are being compared. The test statistic, called the *LR statistic*, is the difference between the log likelihood statistics of the two models being compared, one of which is a subset of the other (for example, model 1 above is a subset of model 2). The larger model is referred to as the *full model* and the smaller as the *reduced model*; that is, the reduced

model is obtained by setting certain parameters in the full model to zero. For example, setting β_3 to zero in model 2 would reduce it to model 1; thus, model 1 is the reduced model and model 2 is the full model. The *LR* statistic in large samples has an approximate χ^2 distribution and the degrees of freedom (df) for this χ^2 test are equal to the difference between the number of parameters in the two models.

Let us now compare model 1 with model 2 as an example of the *LR* test. The additional parameter in the full model (model 2) that is not part of the reduced model (model 1) is β_3 , the coefficient of the variable X_3 . Thus, the null hypothesis that compares Models 1 and 2 is stated as β_3 equal to 0 (usually denoted as $H_0 : \beta_3 = 0$). If we suppose that X_3 is a (0, 1) exposure variable and X_1 and X_2 are confounders, then the *OR* for the exposure–disease relationship controlling for confounders is given by $\exp(\beta_3)$. Thus, in this case, testing the null hypothesis that $\beta_3 = 0$ is equivalent to testing the null hypothesis that the adjusted *OR* for effect of exposure is equal to $\exp(0) = 1$ (i.e., the null value for *OR* is therefore 1 and not 0).

To test this null hypothesis, the corresponding *LR* statistic is given by the following quantity:

$$\begin{aligned} \text{LR statistic} &= (-2 \log_e L_1) - (-2 \log_e L_2) \\ &= -2 \log_e \left(\frac{L_1}{L_2} \right) \end{aligned} \quad (17)$$

where $L_1 = ML$ value for model 1 and, $L_2 = ML$ value for model 2.

The *LR* statistic is approximately χ^2 if the study size is large with $df = 1$, for this case. The *P* value associated with the *LR* statistic is then calculated using the χ^2 distribution with 1 *df* and we reject the null hypothesis if the *P* value is greater than the prefixed value of α (α is typically fixed at 0.05).

The *Wald test* is another way of carrying out hypothesis testing in logistic regression. This test is usually done when there is only one parameter being tested (for example, comparing models 1 and 2 above). The *Wald test* statistic is computed by dividing the estimated coefficient of interest by its standard error. This test statistic has approximately a normal or *Z* distribution in large samples. The square of this *Z* statistic is approximately a χ^2 statistic with 1 *df*. We can carry out the *Wald test* by looking up the *ML* coefficient of the parameter of interest and its standard error from the logistic regression output.

Several packages, including SAS, also compute the χ^2 statistic and the P value associated with it.

As an example of a Wald test, consider models 1 and 2. The Wald test for testing the null hypothesis that $\beta_3 = 0$ is given by the Z statistic which is equal to $\hat{\beta}_3$ divided by the standard error of $\hat{\beta}_3$. The computed Z can be compared to percentage points from a standard normal table or Z can be squared and then compared to percentage points from a χ^2 distribution with 1 *df*.

The LR statistic and its corresponding Wald statistic give approximately the same value in very large samples; so if the study is large enough, it will not matter which statistic is used. Nevertheless, the LR statistic is generally preferred over (i.e., has better statistical properties than) the Wald statistic whenever numerical values computed for corresponding LR and Wald statistics differ substantially.

The most commonly reported interval estimate is the large sample CI for the parameter obtained by computing the estimate of the parameter plus or minus a percentage point of the normal distribution times the estimated standard error. For example, the CI for the β_3 parameter in model 2 is given by

$$100(1 - \alpha)\% \text{ CI for } \beta_3 = \hat{\beta}_3 \pm Z_{1-\alpha/2} \times s_{\hat{\beta}_3} \quad (18)$$

where $s_{\hat{\beta}_3}$ = standard error of $\hat{\beta}_3$.

In the formula shown in equation (18), the values for $\hat{\beta}_3$ and its standard error estimate are obtained from the relevant computer output. The Z percentage point is obtained from tables of the standard normal distribution. For example, if we want a 95% CI , then $\alpha = 0.05$, $1 - \alpha/2 = 0.975$, and $Z_{0.975} = 1.96$. The formula for the 95% CI can then be represented as

$$95\% \text{ CI for } \beta_3 = \hat{\beta}_3 \pm 1.96 \times s_{\hat{\beta}_3} \quad (19)$$

However, most epidemiologists are more interested in the CI for an OR rather than the CI for the parameter coefficient. When we are considering only one (0, 1) exposure variable (for example, X_3 in model 2), then the CI for the OR is obtained by exponentiating the confidence limits obtained for the parameter:

$$\begin{aligned} 95\% \text{ CI for } OR &= 95\% \text{ CI for } \exp(\beta_3) \\ &= \exp(\hat{\beta}_3 \pm 1.96 \times s_{\hat{\beta}_3}) \quad (20) \end{aligned}$$

The formula in equation (20) is correct provided that the variable X_3 is a (0, 1) variable. If this variable

is coded differently (for example, as $(-1, 1)$) or if it is an ordinal or interval variable, then the CI formula must be modified to reflect the coding.

Equation (20) will help us obtain the lower and upper bounds on the OR . The width of the CI (upper bound minus lower bound) informs us about the precision of our point estimate, a wider interval indicating more uncertainty in the estimate as compared to a narrower interval. Additionally, the CI allows us to comment on the statistical significance of the obtained point estimate. If the null value of one (for the OR estimates) is included in the CI then we can conclude that the point estimate is not statistically significant.

In the following section we use a numerical example to illustrate the concepts outlined above.

Numerical Example of Logistic Regression

(From Applied Regression Analysis and Multivariable Methods 3rd edition by KLEINBAUM/KUPPER/MULLER/NIZAM. 1998. Reprinted with permission of Brooks/Cole, a division of Thomson learning: www.thomsonrights.com. Fax (800)730-2215).

Dengue fever is an acute viral fever transmitted by the *Aedes* mosquito. A retrospective survey of an epidemic of dengue fever was carried out in three Mexican cities in 1984 [2]. We will use a subset of the data obtained from this study that includes a random sample of 196 persons from the city of Puerto Vallarta, 57 of whom were determined to be suffering from dengue fever. The primary aim of the analysis was to identify risk factors associated with having the disease, especially the effect of use of mosquito netting as a determinant of the dengue fever. The analysis was carried out using the SAS statistical procedure. A description of the variables present in the dataset DENGUE.DAT ("The dataset DENGUE.DAT can be freely downloaded from the author's website using the following link <http://www.sph.emory.edu/~dkleinb/ar4.htm>") and used in the analysis is provided below:

Subject ID

Dengue fever status (DENGUE): 1 = yes, 2 = no
Age in years (AGE)

Use of mosquito netting (MOSNET): 0 = yes, 1 = no

Geographical sector in which the subject lived (SECTOR): 1-5

The variable SECTOR was treated as a categorical variable in the regression analysis, so we created four dummy variables to distinguish the five geographical sectors. Sector 5 was chosen as the referent group and the variables were defined as follows:

SECTOR1 = 1 if sector 1, 0 otherwise
 SECTOR2 = 1 if sector 2, 0 otherwise
 SECTOR3 = 1 if sector 3, 0 otherwise
 SECTOR4 = 1 if sector 4, 0 otherwise.

We fit the following logistic model, stated in logit form, to these data:

$$\begin{aligned} \text{logit}[\Pr(Y = 1)] = & \beta_0 + \beta_1(\text{AGE}) + \beta_2(\text{MOSNET}) \\ & + \beta_3(\text{SECTOR1}) + \beta_4(\text{SECTOR2}) \\ & + \beta_5(\text{SECTOR3}) + \beta_6(\text{SECTOR4}) \end{aligned} \quad (21)$$

where Y denotes the outcome variable DENGUE.

The relevant portion of the output obtained when we fit the logistic model using SAS software is shown in Table 1.

From Table 1, we can see that the *ML* coefficients (parameter estimates) obtained for the fitted model are

$$\begin{aligned} \hat{\beta}_0 &= -1.9001, \hat{\beta}_1 = 0.0243, \hat{\beta}_2 = 0.3335, \\ \hat{\beta}_3 &= -2.2200, \hat{\beta}_4 = -0.6589, \hat{\beta}_5 = 0.8121, \\ \hat{\beta}_6 &= 0.5310 \end{aligned} \quad (22)$$

so the fitted model is given in logit form by

$$\begin{aligned} \text{logit}[\hat{\Pr}(Y = 1)] \\ = & -1.9001 + 0.0243(\text{AGE}) + 0.3335(\text{MOSNET}) \\ & - 2.2200(\text{SECTOR1}) - 0.6589(\text{SECTOR2}) \\ & + 0.8121(\text{SECTOR3}) + 0.5310(\text{SECTOR4}) \end{aligned} \quad (23)$$

On the basis of this fitted model (with estimates obtained from the computer output), we can compute the estimated *OR* for contracting dengue fever for persons who did not use mosquito netting relative to persons who did use mosquito netting, controlling for age and sector. We can use the formula shown in equation (14) in this case because MOSNET is a (0, 1) variable and there are no product terms in the model. The estimated *OR* is thus obtained by exponentiating the estimated coefficient ($\hat{\beta}_2$) of the MOSNET variable in the fitted model, as shown below:

$$\begin{aligned} \hat{OR}_{(\text{MOSNET}=1 \text{ vs } \text{MOSNET}=0 | \text{age, sector})} \\ = \exp(\hat{\beta}_2) = \exp(0.3335) = 1.396 \end{aligned} \quad (24)$$

The estimated *OR* shown in equation (24) is the odds of getting dengue in those who do not use mosquito netting as compared to the odds in those who use it *given* we keep age and sector constant. So for this case, we are considering MOSNET as the exposure variable and AGE and SECTOR as confounders. From the estimated *OR*, we can conclude that the odds of contracting dengue is about 1.4 times higher for a person who did not use mosquito netting than for a person who did use mosquito netting while controlling for age and sector.

A 95% *CI* for the estimated *OR* in equation (24) above can be computed by using the formula in equation (20).

$$\begin{aligned} 95\% \text{ CI for } \hat{OR}_{(\text{MOSNET}=1 \text{ vs } \text{MOSNET}=0 | \text{age, sector})} \\ = \exp\left(\hat{\beta}_2 \pm 1.96 \times s_{\hat{\beta}_2}\right) \\ = \exp[0.3335 \pm 1.96(1.2718)] \end{aligned} \quad (25)$$

where 1.2718 is the estimated standard error of $\hat{\beta}_2$ (see Table 1). The resulting lower and upper confidence limits are 0.115 and 16.89, respectively.

Table 1 Maximum likelihood estimates obtained by fitting the logistic model shown in equation (15) using the LOGISTIC procedure in SAS

Parameter	Estimate	Standard error	Wald χ^2	Pr > χ^2	Odds ratio
Intercept	-1.9001	1.3254	2.0551	0.1517	0.150
AGE	0.0243	0.00906	7.1778	0.0074	1.025
MOSNET	0.3335	1.2718	0.0688	0.7931	1.396
SECTOR1	-2.2200	1.0723	4.2861	0.0384	0.109
SECTOR2	-0.6589	0.5536	1.4164	0.2340	0.517
SECTOR3	0.8121	0.4750	2.9235	0.0873	2.253
SECTOR4	0.5310	0.4502	1.3911	0.2382	1.701

We can see that the *CI* is very wide and contains the null value of one.

We can also conduct hypothesis testing by using Wald and *LR* tests to test the following null hypothesis:

$$H_0 : OR_{(MOSNET=1 \text{ vs } MOSNET=0 | \text{age, sector})} = 1 \quad (26)$$

or equivalently,

$$H_0 : \beta_2 = 0 \quad (27)$$

The Wald (χ^2) statistic for the null hypothesis and the corresponding *P* value are 0.0688 and 0.7931, respectively (see the row for parameter MOSNET in Table 1). The *P* value of 0.7931 being greater than $\alpha = 0.05$, we conclude that *OR* is not statistically significantly different from 1.

As stated earlier, the *LR* test can be conducted by fitting two models – the full and the reduced models. In this case, the full model is the same as shown in equation (21). The reduced model is obtained under the null hypothesis by dropping the exposure variable (MOSNET) from the full model. The reduced form can be written in the logit form as

$$\begin{aligned} \text{logit}[\Pr(Y = 1)] \\ = \beta_0 + \beta_1(\text{AGE}) + \beta_3(\text{SECTOR1}) \\ + \beta_4(\text{SECTOR2}) + \beta_5(\text{SECTOR3}) \\ + \beta_6(\text{SECTOR4}) \end{aligned} \quad (28)$$

We can obtain the respective log likelihoods for the two models from the computer output (not shown) generated by fitting these models. The $-2 \log_e L_F$ was 203.706 and that of the reduced model (denoted by $-2 \log_e L_R$) was 203.778. The *LR* statistic for these two models can be computed by using the formula shown in equation (17).

$$\begin{aligned} LR \text{ statistic} &= (-2 \log_e L_R) - (-2 \log_e L_F) \\ &= 203.778 - 203.706 = 0.072 \end{aligned} \quad (29)$$

This statistic is χ^2 with 1 *df* because the null hypothesis $H_0 : \beta_2 = 0$ involves only one parameter; the test is nonsignificant at the $\alpha = 0.05$ level. This agrees with our prior observations obtained from looking at the *CI* and the Wald test. These observations support the conclusion that there is no statistical evidence in these data that the nonuse of mosquito netting, significantly increases the probability (or risk) of contracting dengue fever.

We can also evaluate the effect of predictors other than MOSNET on the probability of getting dengue fever from the logistic model in equation (21), involving the predictors AGE, MOSNET, and SECTOR1 through SECTOR4. For example, let us examine the effect of AGE, controlling for MOSNET and the four SECTOR dummy variables. From the row corresponding to the parameter AGE in Table 1, we find that the Wald χ^2 statistic is 7.1778. The null hypothesis being tested here is

$$H_0 : \beta_1 = 0 \quad (30)$$

where β_1 is the coefficient of the AGE variable in the model shown in equation (21). The *P* value for this test is 0.0074 (see Table 1), so we can reject the null hypothesis (at $\alpha < 0.01$ significance level) and conclude that the AGE variable is a significant predictor of dengue fever status. In particular, the positive estimated coefficient for AGE suggests that the risk of dengue fever increases with increasing age.

The *LR* test of the same null hypothesis (i.e., $H_0 : \beta_1 = 0$) would require subtracting the log likelihood statistic for the full model (equation (21)) from the log likelihood statistic for the following reduced model, which is obtained by dropping the AGE variable from equation (21).

$$\begin{aligned} \text{logit}[\Pr(Y = 1)] \\ = \beta_0 + \beta_2(\text{MOSNET}) + \beta_3(\text{SECTOR1}) \\ + \beta_4(\text{SECTOR2}) + \beta_5(\text{SECTOR3}) \\ + \beta_6(\text{SECTOR4}) \end{aligned} \quad (31)$$

The resulting *LR* test also gives a significant result (data not shown), which is in concordance with the significant result obtained from the Wald test.

Since the (Wald or *LR*) test just described measures the effect of a continuous variable (AGE), its specific purpose is to assess whether the effect of the AGE variable in equation (21) is described by a significant linear relationship between AGE and the log odds of dengue fever, controlling for the other variables in the model. In other words, when no other functions of AGE (e.g., AGE² or AGE $\times$ MOSNET) other than AGE itself are in the model, then we are testing whether a linear effect in AGE is more plausible than no effect of age (i.e., we are testing $H_0 : \beta_1 = 0$) of AGE. Such a test is typically referred to as a (*linear*) *trend* test. Thus, using the

results shown in Table 1, we conclude that a linear trend test for AGE in equation (21) is significant.

Since the AGE variable is a continuous variable in the model in equation (21), the quantity $\exp(\hat{\beta}_1)$ is the following OR estimate:

$$\begin{aligned}\hat{OR}_{(AGE1-AGE0=1 \mid \text{mosnet, sector})} &= \exp(\hat{\beta}_1) \\ &= \exp(0.0.0243) = 1.025\end{aligned}\quad (32)$$

where $\hat{\beta}_1$ is the estimated coefficient of the AGE variable (see Table 1).

The estimated OR in equation (32) describes the OR comparing two persons whose age differs by only 1 year. Since AGE is continuous, a one-unit difference in AGE is rarely of interest. Instead, we need to compare meaningfully (e.g., clinically) different categories of AGE. For example, we might consider two persons whose ages differ by 5 years – say, a 35-year-old person and a 30-year-old person – as having a meaningful difference in age. Using the general OR expression shown in equation (13), we can compute the estimated OR for two persons with a 5-year age difference, controlling for mosquito netting and sector, as follows:

$$\begin{aligned}\hat{OR}_{(AGE1-AGE0=5 \mid \text{mosnet, sector})} &= \exp[5\hat{\beta}_1 + (\text{MOSNET} - \text{MOSNET})\hat{\beta}_2 \\ &\quad + (\text{SECTOR1} - \text{SECTOR1})\hat{\beta}_3 + \dots \\ &\quad + (\text{SECTOR4} - \text{SECTOR4})\hat{\beta}_6] \\ &= \exp[5\hat{\beta}_1] = \exp[5(0.0243)] = 1.13\end{aligned}\quad (33)$$

Since we are controlling for mosquito netting and sector status, we assume that the MOSNET variable has the same value, and that the four sector variables (SECTOR1 through SECTOR4) have the same value for the two persons being compared. The coefficients of these five variables then drop out of the OR expression, and we have only the quantity $5\hat{\beta}_1$ to exponentiate. The estimated adjusted OR is therefore given by $\exp[5\hat{\beta}_1]$ rather than by $\exp[\hat{\beta}_1]$, since we are considering the effect of a 5-year rather than a 1-year difference in age. Thus, for two persons whose age differs by 5 years, the adjusted OR that controls for use of mosquito netting and sector status equals 1.13, which is quite close to the null value of 1.

To obtain a 95% CI for this adjusted OR (i.e., $5\hat{\beta}_1$), we make the following calculation using

values obtained from Table 1:

$$\begin{aligned}95\% \text{ CI for } \hat{OR}_{(AGE1-AGE0=5 \mid \text{mosnet, sector})} : \\ &= \exp[5\hat{\beta}_1 \pm 1.96(5s_{\hat{\beta}_1})] \\ &= \exp[5(0.0.0243) \pm 1.96(5)(0.0091)]\end{aligned}\quad (34)$$

which yields 95% confidence limits of 1.03 and 1.23. Since the confidence limits do not include the null value of 1, the effect of a 5-year difference in age is statistically significant. We can also see that the CI obtained here (i.e., 1.03–1.23) is much narrower than that obtained for the MOSNET variable (i.e., 0.115–16.89), indicating greater precision in the estimation of the OR for the effect of AGE than obtained for the effect of MOSNET.

If, instead of a 5-year difference in age, we consider the effect of a 10-year difference in age, the formulas for the estimated adjusted OR and 95% CI become

$$\hat{OR}_{(AGE1-AGE0=10 \mid \text{mosnet, sector})} = \exp[10\hat{\beta}_1]\quad (35)$$

and

$$\begin{aligned}95\% \text{ CI for } \hat{OR}_{(AGE1-AGE0=10 \mid \text{mosnet, sector})} \\ &= \exp[10\hat{\beta}_1 \pm 1.96(10s_{\hat{\beta}_1})]\end{aligned}\quad (36)$$

respectively. The corresponding computed values are 1.28 for the adjusted OR and 1.07 and 1.52 for the confidence limits.

Thus far, we have only considered models with main effect variables like MOSNET, AGE, and SECTOR; we have not considered product terms such as $\text{MOSNET} \times \text{AGE}$. When the model contains product terms (that include the exposure variable), the simple equation (14) for obtaining the adjusted OR does not work. In such situations, we must use the general formula given by equation (13). Let us consider the model given below, which, in addition to the terms contained in the model shown in equation (21), contains the product term $\text{MOSNET} \times \text{AGE}$:

$$\begin{aligned}\text{logit}[\Pr(Y = 1)] \\ &= \beta_0 + \beta_1(\text{AGE}) + \beta_2(\text{MOSNET}) \\ &\quad + \beta_3(\text{SECTOR1}) + \beta_4(\text{SECTOR2}) \\ &\quad + \beta_5(\text{SECTOR3}) + \beta_6(\text{SECTOR4}) \\ &\quad + \beta_7(\text{MOSNET} \times \text{AGE})\end{aligned}\quad (37)$$

Table 2 Maximum likelihood estimates obtained by fitting the logistic model shown in equation (37) using the LOGISTIC procedure in SAS

Parameter	Estimate	Standard error	Wald > χ^2	Pr > χ^2	Odds ratio
Intercept	−0.8080	1.6311	0.2454	0.6203	0.446
AGE	−0.00434	0.0362	0.0143	0.9048	0.996
MOSNET	−0.8043	1.6433	0.2396	0.6245	0.447
SECTOR1	−2.2929	1.0804	4.5042	0.0338	0.101
SECTOR2	−0.6813	0.5541	1.5118	0.2189	0.506
SECTOR3	0.8153	0.4756	2.9388	0.0865	2.260
SECTOR4	0.5115	0.4515	1.2830	0.2573	1.668
MOSNET × AGE	0.0306	0.0374	0.6689	0.4134	1.031

To obtain an adjusted *OR* for the effect of MOSNET adjusted for AGE and SECTOR, we need to specify the two groups being compared – say $\mathbf{X}_A$ and $\mathbf{X}_B$, as in equation (13). For these two groups we need to specify the values of the predictor variables defined by

$$\begin{aligned} \mathbf{X} = & (\text{AGE}, \text{MOSNET}, \text{SECTOR1}, \text{SECTOR2}, \\ & \text{SECTOR3}, \text{SECTOR4}, \text{MOSNET} \times \text{AGE}) \end{aligned} \quad (38)$$

Then, to compute the *OR* between those who did not use mosquito netting relative to those who did, we need to define the two groups as follows:

$$\mathbf{X}_A = (\text{AGE}, 1, \text{SECTOR1}, \text{SECTOR2}, \text{SECTOR3}, \text{SECTOR4}, 1 \times \text{AGE}) \quad (39)$$

$$\mathbf{X}_B = (\text{AGE}, 0, \text{SECTOR1}, \text{SECTOR2}, \text{SECTOR3}, \text{SECTOR4}, 0 \times \text{AGE}) \quad (40)$$

Substituting $\mathbf{X}_A$ and $\mathbf{X}_B$ into the formula shown in equation (13) for the *OR*, we obtain

$$\begin{aligned} OR_{(\text{MOSNET}=1 \text{ vs } \text{MOSNET}=0 | \text{age, sector})} &= \exp[(\text{age} - \text{age})\beta_1 + (1 - 0)\beta_2 + (\text{sector1} \\ &\quad - \text{sector1})\beta_3 + \dots + (\text{sector4} - \text{sector4})\beta_6 \\ &\quad + (\text{age} - 0)\beta_7] = \exp[\beta_2 + \beta_7(\text{age})] \end{aligned} \quad (41)$$

Thus, rather than involving only β_2 , the preceding expression in equation (41) involves the two coefficients β_2 and β_7 , each of which is a coefficient of a variable in the model that contains the exposure variable, MOSNET, in some form. The coefficient β_7 is included because it is the coefficient of the “interaction term” in the model of equation (37). This model

essentially says that the value of the *OR* for the effect of MOSNET should vary depending on the values of the variable AGE and of the coefficient β_7 . The variable AGE in equation (37) is an example of an *effect modifier* (see **Effect Modification and Interaction**), since the effect of the exposure variable MOSNET, as quantified by the *OR* formula in equation (41) above, changes (i.e., is *modified*) depending on the value of age.

We fit the “interaction model” of equation (37) to the DENGUE dataset using the LOGISTIC procedure in SAS. Selected *ML* estimates from the output are presented in Table 2.

Previously, in equation (24), we considered the estimated *OR* for MOSNET using the “no-interaction model” shown in equation (21), but now we are using the logistic model in equation (37) instead. The estimated *OR* for the interaction model can be calculated using equation (41) as follows:

$$\begin{aligned} \hat{OR}_{(\text{MOSNET}=1 \text{ vs } \text{MOSNET}=0 | \text{age, sector})} &= \exp[\hat{\beta}_2 + \hat{\beta}_7(\text{age})] \\ &= \exp[-0.8043 + 0.0306(\text{age})] \end{aligned} \quad (42)$$

where $\hat{\beta}_2 = -0.8043$ is the estimated coefficient of the MOSNET variable and $\hat{\beta}_7 = 0.0306$ is the estimated coefficient of the (MOSNET × AGE) variable (see Table 2). The adjusted *OR* formula in equation (42) says that the value of the ratio depends on the value we specify for AGE, which is exactly what we mean when we assume (using the model in equation (37)) that age is an effect modifier of the relationship between mosquito net use and risk of contracting dengue fever.

Table 3 shows the computed values for this adjusted *OR* for various specifications of the effect modifier, age.

Table 3 Effect of age on the value of the estimated adjusted odds ratio obtained by fitting the logistic model shown in equation (37)

Age	Adjusted OR
10	0.61
15	0.71
20	0.83
25	0.96
26.28	1.00
30	1.12
35	1.31
40	1.52
45	1.77
50	2.07
55	2.41
60	2.81

Table 3 indicates that, on the basis of the interaction model in equation (37), the adjusted *OR* for the effect of mosquito net usage status varies from values below the null value of 1 when age is 25 or less to values increasingly above 1 (though below 3) when age is over 30. These results suggest, for example, that a person of age 20 who does not use a mosquito net (i.e., MOSNET = 1) is 0.83 times as likely to get dengue fever as a person who uses a mosquito net; in contrast, at age 40, a person who does not use a mosquito net is 1.52 times as likely to get dengue fever as a person who uses a mosquito net. However, this result is not statistically significant (data not shown) and we have not yet evaluated whether the model with the interaction term (MOSNET $\times$ AGE) is a better fit to the data compared to the no-interaction model.

We can obtain *CI*s for any of the adjusted *OR*s presented in Table 3 by using the general formula given below:

$$\exp[\hat{L} \pm 1.96(S_{\hat{L}})] \quad (43)$$

where $\hat{L} = \hat{\beta}_2 + \hat{\beta}_7$ (age) and $S_{\hat{L}} = \sqrt{\hat{\text{Var}}(\hat{L})}$, and where $\hat{\text{Var}}(\hat{L})$ is computed using the formula for the variance of a linear combination of random variables, which for this example turns out to be

$$\begin{aligned} \hat{\text{Var}}(\hat{L}) &= \hat{\text{Var}}(\hat{\beta}_2) + (\text{age})^2 \hat{\text{Var}}(\hat{\beta}_7) \\ &\quad + 2(\text{age}) \hat{\text{Cov}}(\hat{\beta}_2, \hat{\beta}_7) \end{aligned} \quad (44)$$

The numerical values of the variances and the covariances involving the estimated parameter coefficients are typically printed out as an option by the computer program used. For the model in equation (37), the estimated variances and covariances in the preceding variance formula are

$$\hat{\text{Var}}(\hat{\beta}_2) = 2.7004 \quad (45)$$

$$\hat{\text{Var}}(\hat{\beta}_7) = 0.001399 \quad (46)$$

and

$$\hat{\text{Cov}}(\hat{\beta}_2, \hat{\beta}_7) = -0.0435 \quad (47)$$

If, for example, we let age = 40, then

$$\hat{L} = -0.8043 + 0.0306(40) = 0.4197 \quad (48)$$

$$\begin{aligned} \hat{\text{Var}}(\hat{L}) &= 2.7004 + (40)^2(0.001399) \\ &\quad + 2(40)(-0.0435) = 1.4588 \end{aligned} \quad (49)$$

and

$$S_{\hat{L}} = \sqrt{\hat{\text{Var}}(\hat{L})} = 1.2078 \quad (50)$$

The 95% *CI* for the adjusted *OR*, using equation (43), is then given by

$$\begin{aligned} &\exp[\hat{L} \pm 1.96(S_{\hat{L}})] \\ &= \exp[0.4197 \pm (1.96)(1.2078)] \end{aligned} \quad (51)$$

which yields lower and upper confidence limits of 0.14 and 16.23, respectively. This *CI* is very wide, which implies that the point estimate 1.52 when age is 40 (see Table 3), is unreliable. Furthermore, since the *CI* includes the null value of one, a test of significance for the adjusted *OR* is not significant at $\alpha = 0.05$.

Finally, we may wish to evaluate whether the interaction model (equation (37)) forms a better fit to the data as compared to the no-interaction model (equation (21)). We can do this by using the *LR* test of the null hypothesis:

$$H_0 : \beta_7 = 0 \quad (52)$$

The full model is the model in equation (37) and the reduced model, which does not have the interaction term, is the one shown in equation (21). The following *LR* statistic was calculated:

$$\begin{aligned} \text{LR statistic} &= (-2 \log_e L_R) - (-2 \log_e L_F) \\ &= 203.706 - 202.9950.711 \end{aligned} \quad (53)$$

The statistic is approximately χ^2 with 1 *df*. The value 0.711 for the *LR* statistic is nonsignificant indicating that the no-interaction model in equation (21) is preferable to the interaction model in equation (37). In other words, the *OR* for the effect of MOSNET adjusted for age and sector is best expressed by the single value of 1.396 shown in equation (24) based on the fit of the model in equation (21), rather than by the expression in equation (42), which is based on the model in equation (37).

Conclusion

In this chapter, we have provided a brief overview of the logistic regression analysis procedure, with emphasis on describing how to obtain point estimates and *CI*s and how to conduct tests of hypothesis for *OR* parameters of interest derived from estimated regression coefficients in the model.

More detailed discussion on logistic regression, including theoretical considerations, can be found in several textbooks [1, 3–6].

In particular, these references distinguish between two types of *ML* procedures, called *unconditional* and *conditional*. The latter (conditional) procedure is typically used for logistic models in which the number of model parameters is large relative to the number of subjects being studied, as when the study design involves individual matching.

Also, Kleinbaum and Klein [5] provide a detailed discussion of the computer code to be used for logistic regression analysis in the statistical packages SAS, STATA, and SPSS [5].

Finally, in this article, we have not discussed strategies for selecting the variables that should go

into a logistic model. The goal of such a general analysis strategy is to obtain a valid estimate (in the context of epidemiologic validity and not statistical validity) of the relationship between a specified exposure and a particular disease variable, while controlling for other covariates that, if not correctly taken into account, can lead to an incorrect assessment of the strength of the exposure–disease relationship. Readers interested in this important aspect of analysis are referred to Kleinbaum and Klein [5, Chapters 6 and 7] and Kleinbaum *et al.* [6, Chapters 21–24] for a discussion of the aforementioned general analysis strategy [5, 6].

References

- [1] Kleinbaum, D.G., Kupper, L.L., Muller, K.E. & Nizam, A. (1998). *Applied Regression Analysis and Multivariable Methods*, 3rd Edition, Duxbury Press, Washington, DC.
- [2] Dantes, H.G., Koopman, J.S., Addy, C.L., Zarate, M.L., Marin, M.A., Longini Jr, I.M., Gutierrez, E.S., Rodriguez, V.A., Garcia, L.G. & Mirelles, E.R. (1988). Dengue epidemics on the Pacific Coast of Mexico, *International Journal of Epidemiology* **17**(1), 178–186.
- [3] Collett, D. (1991). *Modeling Binary Data*, Chapman & Hall, London.
- [4] Hosmer, D.W. & Lemeshow, S. (1989). *Applied Logistic Regression*, John Wiley & Sons, New York.
- [5] Kleinbaum, D.G. & Klein, M. (2002). *Logistic Regression – A Self-Learning Text*, 2nd Edition, Springer, New York, Berlin.
- [6] Kleinbaum, D.G., Kupper, L.L. & Morgenstern, H. (1982). *Epidemiologic Research – Principles and Quantitative Methods*, Van Nostrand Reinhold, New York.

CHIRANJEEV DASH AND DAVID G. KLEINBAUM

Longevity Risk and Life Annuities

Introduction: What Is Longevity Risk?

According to the Merriam–Webster dictionary, the definition of the word *longevity* is “a long duration or length of individual life” and the word *risk*, in the broadest sense of the term, is defined as “something that creates or suggests a hazard”. Merging these two definitions implies that a long duration for individual life creates a hazard, which might seem odd to anyone not trained in actuarial science or insurance. After all, some might wonder, how can a long life be hazardous?

Indeed, although most readers would agree that a long life should be a blessing, it can be a costly one. From a personal perspective, financing 90, 95, or even 100 years of life can be expensive, especially if it is *unexpected*. This is at the heart of longevity risk. If one thinks in terms of a personal financial balance sheet, longevity risk is related to the liability side of a long life. It is the exact opposite of mortality risk, which relates to the economic cost to survivors and family because of an unexpectedly short life.

From a financial point of view, the practical implications of longevity risk come from its uncertainty as opposed to its length. If everyone knew exactly how long they would live – even if this was a very long time – longevity risk would be nonexistent and, the author’s claim is that there would then be very little interest in, and need for, a market for life annuities or general mortality-contingent claims. The greater the uncertainty around the length of human life, both individually and in aggregate, the greater is the need for a well-developed annuity market. This relatively trivial statement – that the variance is more important than the mean – has critical implications for the future of the insurance industry as it relates to the provision of retirement income, and this aspect is also discussed in this article.

On a more formal level, if the symbol $T_x^z(i) > 0$ denotes the remaining lifetime random variable for individual i , aged x in the year z (for example, $x = 65$ in the year $z = 2006$), then obviously $E[T_0^z]$ denotes general life expectancy at birth in year z . Of critical importance to the topic of longevity risk is the strong disagreement among demographers as to whether

there is a biological upper bound on $E[T_0^z]$ as z goes to 2010, 2020, and beyond.

Well-cited researchers such as Olshansky and Carnes [1] claim that life expectancy at birth is not likely to ever exceed the age of 85 or 90. These rather pessimistic authors are among a group of scholars who believe that the remarkable reductions in mortality rates across all ages during the last two centuries are a “disconnected sequence of unrepeatable revolutions” and that biological barriers create a fixed upper bound on life expectancy.

Others, such as Oeppen and Vaupel [2] take a diametrically opposing view and argue that there are no biological limits to life expectancy. They offer as evidence the fact that female life expectancy in record-holding countries has been increasing at a steady pace of 3 months per year for the last 160 years. Indeed, when (record) life expectancy in any given year is regressed against the last 160 years of data, the trend is remarkably linear with an R^2 value of 99% according to Oeppen and Vaupel [2]. These authors are extremely optimistic that biomedical research will yield unprecedented increases in survival rates.

It is precisely this disagreement among experts and the variance of professional opinion that leads to the richness and fertility of longevity risk as a research field. Likewise, in the language of financial economics, there should be a market price for the undiversifiable portion of longevity risk. This aspect is discussed in a later section.

The focus, in this article, is on three distinct aspects or dimensions of longevity risk. The first dimension deals with relation of longevity risks to individuals and their personal financial decisions. It addresses the question, How should individuals manage longevity risk? The second dimension of this article relates to the role of intermediaries and insurance companies that take on and absorb longevity risk. How can they measure and then manage this risk effectively? Is this risk priced by the market? Can it truly be diversified? The final dimension relates to the role of governments and society, as a whole, in managing longevity risk. If the private sector cannot, for whatever reason, develop a fully functional market for longevity risk, does the government have a role to play? How, if at all, should the market for (unwanted) mortality-contingent claims be regulated?

Owing to a limitation in space, the academic citations and literary references have been kept brief;

however, some key papers in all three – individual, business, and government – dimensions have been highlighted.

The interested reader is referred to [3] for a much more extensive list of relevant technical references related to the market for life annuities. In addition, the December 2006 issue of the *Journal of Risk and Insurance* is devoted entirely to the topic of longevity risk management and the interested reader is encouraged to consult these as well.

Individual Dimension: Why Buy Life Annuities?

One of the stylized facts in the research field of longevity risk and life annuities is the widespread reluctance of individuals to voluntarily purchase life annuities. Recall that life annuities can serve as a hedge against longevity risk. Modigliani [4], in his Nobel Prize lecture referred to this phenomena as the *annuity puzzle* and many papers have since been published. Indeed, in the United States less than 5% of the population of retirees has voluntarily purchased a life annuity, according to a survey conducted by the consulting firm LIMRA International. The survey of consumer finances (SCF) conducted by the Federal Reserve arrives at similar conclusions. However, inclusion of defined benefit (DB) pension income, which is also a type of annuity, increases the number of annuitized individuals. In the United Kingdom, recent changes to pension law weakened the mandatory conversion of certain tax-sheltered savings plans to life annuities at the age of 75, mostly because of public dislike and distrust of annuitization.

To understand why (immediate) life annuities remain unpopular – and the implications of this to the management of personal longevity risk – we must start at the microeconomic foundations of the demand for life annuities.

A simple two-period microeconomic example that illustrates the gains in personal utility from having access to a life annuity market is dealt with, next. Assume that an individual has an initial sum of \$1 which must be allocated and consumed during either of the next two periods. The consumption that must be selected, denoted by C_1 and C_2 , takes place for convenience at the end of the period. There is a p_1 probability that the individual will survive to (consume at) the end of the first period, and a p_2

probability of surviving to (consuming at) the end of the second period. The periodic interest rate – which is also the return on investment in this simple model – is denoted by R .

There is some disagreement as to whether individuals are able to formulate accurate and consistent probability estimates for their own survival rates (p_1, p_2). For example, Smith *et al.* [5] claim that, in general, an individual's estimates are consistent with objective mortality tables (*see Estimation of Mortality Rates from Insurance Data*), and this view is echoed by Hurd and McGarry [6] as well. In contrast, Bhattacharya *et al.* [7] claim that individuals have systematic biases in the way they perceive their own mortality rates. This is clearly an ongoing area of research.

Either way, we assume that the values of the p 's are known, and the objective is to maximize discounted utility of consumption. To make things even easier, we assume logarithmic preferences for consumption utility. This is all standard. In the absence of life annuities or any market for life contingent claims, our objective function and budget constraint would be described by

$$\begin{aligned} \max_{\{C_1, C_2\}} \text{Utility} &= \frac{p_1}{1 + \rho} \ln[C_1] \\ &+ \frac{p_2}{(1 + \rho)^2} \ln[C_2] \end{aligned} \quad (1)$$

$$\text{subject to } 1 = \frac{C_1}{1 + R} + \frac{C_2}{(1 + R)^2} \quad (2)$$

where ρ is our subjective discount rate. This “toy model” is a subset of a classical life cycle model of investment and consumptions that is at the heart of microeconomics and is described in greater length in most economics textbooks. Solving the problem, we obtain the following optimal values for consumption over the two periods:

$$\begin{aligned} C_1^* &= \frac{p_1(\rho R + R + \rho + 1)}{p_2 + p_1\rho + p_1}, \\ C_2^* &= \frac{p_2(1 + 2R + R^2)}{p_2 + p_1\rho + p_1} \end{aligned} \quad (3)$$

The ratio of optimal consumption between period one and period two is $C_1^*/C_2^* = p_1(1 + \rho)/p_2(1 + R)$. Now we arrive at the important qualitative implication.

When the subjective discount rate is equal to the interest rate ($\rho = R$), then $C_1^*/C_2^* = p_1/p_2$, which is the ratio of the survival probabilities, which is strictly greater than one. Stated differently, the optimizer will consume less at higher ages. In fact, one might rationally starve to death! This insight regarding optimal consumption over the life cycle is quite robust and explained in [4], amongst other references that extend this from two periods to continuous time. If one cannot insure or hedge one's longevity risk, which is the probability of living for two periods (as opposed to just one), then one will adapt to this risk by taking one's chances and consuming less as one ages (*see Fair Value of Insurance Liabilities*). There are alternative strategies that provide greater utility and greater consumption possibilities.

In the presence of actuarially fair life annuities – which hedge against personal longevity risk – optimal behavior is quite different. The budget constraint in equation (2) will change to reflect the probability-adjusted discount factor. The optimization problem is now as follows:

$$\begin{aligned} \max_{\{C_1, C_2\}} \quad & \text{Utility} = \frac{p_1}{1 + \rho} \ln[C_1] \\ & + \frac{p_2}{(1 + \rho)^2} \ln[C_2] \end{aligned} \quad (4)$$

$$\text{subject to} \quad 1 = \frac{p_1 C_1}{1 + R} + \frac{p_2 C_2}{(1 + R)^2} \quad (5)$$

We notice that the only difference between this and the previous optimization problem is the constraint itself. If personal longevity risk can be hedged in the insurance market, the present value of any given consumption plan will be lower. It is cheaper to finance the same standard of living if one's assets are lost at death. Annuitization involves a trade-off between utility of consumption and utility of bequest.

It is questionable whether one should use the same survival probability (p_1, p_2) in the objective function and the constraint owing to adverse selection issues in the annuity market. The interested reader is referred to the paper by Finkelstein and Poterba [8] for a discussion of this issue. Nevertheless, in this case, the optimal consumption is denoted by C_1^{**}, C_2^{**} , and is given by

$$C_1^{**} = \frac{\rho R + R + \rho + 1}{p_2 + p_1 \rho + p_1}, \quad C_2^{**} = \frac{1 + 2R + R^2}{p_2 + p_1 \rho + p_1} \quad (6)$$

The important point to notice is that $C_1^{**} = C_1^*/p_1$ and $C_2^{**} = C_2^*/p_2$, which implies that the optimal consumption is greater in both periods, when one has access to life annuities, regardless of whether adverse selection plays a significant role. Specifically, at time zero, the individual would purchase a life annuity that pays C_1^{**} at time 1 and C_2^{**} at time 2. The present value of the two life annuities – as per the budget constraint – is \$1. In this case, the ratio of consumption between period 1 and period 2 is $C_1^*/C_2^* = (1 + \rho)/(1 + R)$. When the subjective discount rate is equal to the interest rate ($\rho = R$), then $C_1^*/C_2^* = 1$.

The point in all of this algebra is as follows. *Buying a life annuity allows rational consumers to eliminate or hedge their personal longevity risk.* Of course, in the absence of life annuities, the best one can hope for is to allocate investment assets in a way that minimizes the lifetime probability of ruin (*see Ruin Theory*), which is an approach taken by Young [9], amongst others. But this is also why Yaari [10, 11], and more recently, Ameriks *et al.* [12] advocate that individuals should allocate a substantial fraction of their retirement wealth to life annuities. In fact, Yaari [10] advocates complete annuitization at any age in the absence of bequest motives. A recent paper by Milevsky and Young [13] looks at annuitization as an optimal timing problem, but all of these papers are based on the same idea. Annuitization hedges against longevity risk.

Business Dimension: Can Companies Hedge This Risk?

When individuals decide to hedge or insure their own longevity risk by transferring their personal exposure to intermediaries, these insurance companies then face aggregate mortality risk. This applies when selling life annuities or any other form of lifetime income such as guaranteed living income benefit (GLiB), which have become very popular in the US retirement market. Sometimes, longevity risk is interlinked with market and interest risk, as in the case of guaranteed annuity options (GOAs).

It is generally thought that insurance companies (or their own reinsurers) can completely eliminate mortality and longevity risk by selling enough of claims to diversify away mortality. The law of large numbers (LLN) is often invoked as a justification.

But, the truth is far more complicated. Indeed, if there is systematic uncertainty regarding aggregate mortality trends – for example, the research cited earlier by Olshansky and Carnes [1] *versus* Oeppen and Vaupel [2] – then selling more annuities and longevity-related insurance will not necessarily eliminate idiosyncratic risk. The following example should help explain the impact of the stochasticity of aggregate mortality.

Assume that an insurance company sells a one-period longevity insurance policy that pays \$2 if the buyer survives to the end of the period, but pays nothing if the buyer dies prior to this date. Assume the insurance company issues N of these policies – at a price of \$1 per policy – to N independent lives, each of whom has an identical probability p of surviving and probability $(1 - p)$ of dying prior to the payout date. Each longevity insurance policy generates an end-of-period Bernoulli liability for the insurance company. The random variable w_i can take on a value \$2 with probability p and a value of \$0 with probability $(1 - p)$. The expected value $E[w_i] = 2p$ and the variance is $\text{var}[w_i] = p(2 - 2p)^2 + (1 - p)(0 - 2p)^2 = 4p(1 - p)$. In the simple case that $p = 0.5$, this collapses to $E[w_i] = 1$ and $\text{var}[w_i] = 1$, as well as $\text{SD}[w_i] = 1$.

Let the random variable, $W_N = \sum_{i=1}^N w_i$ denote the insurance company's aggregate liability at the end of the period owing to selling a portfolio of longevity insurance policies. These independent longevity insurance exposures leave the issuing company facing an aggregate expected payout of $E[W_N] = N2p$, an aggregate variance of $\text{var}[W_N] = N4p(1 - p)$, and an aggregate standard deviation of $\text{SD}[W_N] = 2\sqrt{Np(1 - p)}$. The independent and identically distributed (i.i.d.) nature of the exposures allows me to add up the individual variances. This immediately leads us to the well-known results that the standard deviation per policy (SDP) goes to zero in the limit as $N \rightarrow \infty$. Stated technically,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \text{SD}[W_N] = \lim_{N \rightarrow \infty} 2 \frac{\sqrt{p(1 - p)}}{\sqrt{N}} \rightarrow 0 \quad (7)$$

This is a special case of the LLN, which states that $\lim_{N \rightarrow \infty} \frac{1}{N} W_N \rightarrow p$. If the insurance company sells enough of these longevity insurance policies their risk exposure (per policy) goes to zero. Yet another interpretation of this statement is that longevity risk is completely diversifiable and not compensated by markets in equilibrium, when claims are i.i.d.

What happens when the probability parameter p is unknown? This is a critical aspect of longevity risk. It is an aggregate longevity risk as opposed to a personal longevity risk. This is also equivalent to not knowing the hazard rate or instantaneous force of mortality underlying the probabilities, modeled in various recent papers such as [14], or [15]. Stated differently, What if we have an estimate of p *versus* a certainty for p ?

The underlying payoff function in this simple example is now defined as $w_i^* = 2$ with probability $\tilde{p}$ and $w_i^* = 0$ with probability $(1 - \tilde{p})$. The asterisk on top of the w_i^* reminds the reader that the parameter $\tilde{p}$ itself has its own (symmetric) distribution, which it is assumed takes on a value of $p + \pi$ with probability $1/2$ and a value of $p - \pi$, with a probability of $1/2$.

Obviously a restriction must be imposed on the newly defined uncertainty parameter π , namely, that $\pi \leq 1 - p$ and $p > \pi$. Also, by definition $E[\tilde{p}] = 0.5(p + \pi) + 0.5(p - \pi) = p$. For example, one might assume that $\pi = 0.1$, $p = 0.5$ so that $\tilde{p}$ takes on values of either 0.6 or 0.4. The intuitive interpretation for this would be that while the expected value $E[\tilde{p}] = 0.5$ of the survival probability is 0.5, it is equally likely to take on a value of 0.4 (an improvement in mortality) or 0.6 (a worsening mortality).

Now define the total (aggregate) exposure of the insurance company by the notation $w_N^* = \sum_{i=1}^N w_i^*$, with the immediate implication that $E[W_N^*] = N2p$, which is identical to the traditional (deterministic) case. The key difference between W_N (deterministic mortality) and W_N^* (stochastic mortality) lies in the term for the variance. We are no longer entitled to add up the individual variance terms owing to the implicit dependence created by the common $\tilde{p}$ factor. In this case – and one might argue, in reality – they are not i.i.d. claims. The interested reader is referred to [16] for an extensive discussion on this point, in which it is also shown that

$$\text{var}[W_N^*] = 4Np(1 - p) + 4N\pi^2(N - 1) \quad (8)$$

This collapses to the familiar and intuitive $4Np(1 - p)$ when $\pi = 0$, and we are back to the deterministic mortality world. Note also that when $N = 1$, the variance of the payout is the same $4Np(1 - p)$ it would be under the deterministic case, which means that an individual policy is not any riskier under a stochastic $\tilde{p}$ *versus* a deterministic $p = E[\tilde{p}]$. It is the portfolio aggregation that creates the extra risk.

For example, if an insurance company sells $N = 10\,000$ longevity insurance (also known as *term life annuity*) policies under the parameter set $\pi = 0.1$, so that $\bar{p}$ takes on a value of either 0.6 or 0.4 with equal probability, the variance of the aggregate payout to the insurance company becomes $0.96N + 0.04N^2$ according to equation (8). Notice the N^2 term. The variance grows nonlinearly in N , which means that the SDP $\sqrt{0.96N + 0.04N^2}/N$, will converge to a constant 0.2 instead of zero when $N \rightarrow \infty$. No matter how many longevity insurance policies the company issues, the uncertainty (risk) per policy will never be less than \$0.20 per \$1.00 of expected payoff. The risk never goes away. Diversification can only reduce risk up to a certain point. In general, we have that $\lim_{N \rightarrow \infty} \frac{1}{N}SD[W_N^*] \rightarrow 2\pi$. The SDP converges to the total spread (2π) in the probability (hazard) rate.

What does all of this mean to companies who are trying to manage their own longevity risk? The main take-away is as follows. There is a growing body of research that attempts to model the evolution of aggregate morality, for example the trend in $E[\mathbf{T}_0^z]$, for $z = 2006, 2007, 2008$, etc. Various stochastic processes have been proposed and calibrated, starting with work by Lee and Carter [17], Olivieri [18], Dahl [19], Renshaw and Haberman [20], Carnes, Blake and Dowd [21], as well as Biffis and Millossovich [22]. This is also related to the pricing of GOAs, which is addressed in [23]. All of these papers are effectively based on the same underlying argument. One does not know with certainty the evolution of the dynamics of $\mathbf{T}_x^z$ or even $E[\mathbf{T}_x^z]$.

Government Dimension: Should They Get Involved?

Finally, if longevity risk is faced by individuals and corporations, this becomes a public policy issue as well. If individuals do not value the longevity insurance benefits provided by life annuities and are perhaps myopic in their investment behavior, should government mandate annuitization for tax-sheltered savings plans and other forms of personal pensions? Should governments actively try to correct behavioral biases by subsidizing (entering) the annuity market and making it more appealing to individuals? What can government policy do to encourage a deeper market for annuities? Along these lines, some

researchers have advocated that government issue longevity or survivor bonds as a way for governments to help insurance intermediaries, as well as pension funds cope with longevity risk that cannot be diversified away using the LLN. See, for example, [24], for this line of thinking. Other policy-oriented papers include [25], which examines the role of life annuities in a privatized social security system, or [26], which examines the role of annuities within pension plans, appear to favor the involvement of the private sector to find solutions to this problem. A number of these issues are addressed in the December 2006 issue of the *Journal of Risk and Insurance*, and do not require further elaboration.

To sum up, longevity risk – regardless of how it is defined – is faced by individuals, corporations, as well as society and their governments. In a brief article such as this, it is virtually impossible to review or discuss all papers within this broad field. This article has given a brief overview of the various dimensions and aspects of longevity risk as they relate to the market for life annuities and mortality-contingent claims. It is believed that this field will continue to provide fertile ground for advanced research for many, albeit uncertain, years to come (*see Life Insurance Markets; Equity-Linked Life Insurance; Fair Value of Insurance Liabilities; Pricing of Life Insurance Liabilities; Asset-Liability Management for Life Insurers*).

References

- [1] Olshansky, S.J. & Carnes, B.A. (2001). *The Quest for Immortality: Science at the Frontiers of Aging*, W.M. Norton, New York.
- [2] Oeppen, J. & Vaupel, J.W. (2002). Broken limits to life expectancy, *Science* **296**, 1029–1031.
- [3] Milevsky, M.A. (2006). *The Calculus of Retirement Income: Financial Models for Pension Annuities and Life Insurance*, Cambridge University Press.
- [4] Modigliani, F. (1986). Life cycle, individual thrift and the wealth of nations, *American Economic Review* **76**(3), 297–313.
- [5] Smith, V.K., Taylor, D.H. & Sloan, F.A. (2001). Longevity expectations and death: can people predict their own demise? *American Economic Review* **91**(4), 1126–1134.
- [6] Hurd, M.D. & McGarry, K. (1995). Evaluation of the subjective probabilities of survival in the health and retirement study, *Journal of Human Resources* **30**, 268–291.

- [7] Bhattacharya, J., Goldman, D.P. & Sood, N. (2003). *Market Evidence of Misperceived and Mistaken Mortality Risks*, NBER Working Paper #9863.
- [8] Finkelstein, A. & Poterba, J. (2002). Selection effects in the United Kingdom individual annuities market, *The Economic Journal* **112**, 28–50.
- [9] Young, V.R. (2004). Optimal investment strategy to minimize the probability of lifetime ruin, *North American Actuarial Journal* **8**(4), 106–126.
- [10] Yaari, M.E. (1965). Uncertain lifetime, life insurance and the theory of the consumer, *Review of Economic Studies* **32**, 137–150.
- [11] Richard, S. (1975). Optimal consumption, portfolio and life insurance rules for an uncertain lived individual in a continuous time model, *Journal of Financial Economics* **2**, 187–203.
- [12] Ameriks, J., Veres, R. & Warshawsky, M.J. (2001). Making retirement income last a lifetime, *Journal of Financial Planning* **14**, 1–14.
- [13] Milevsky, M.A., Promislow, S.D. & Young, V.R. (2006). Killing the law of large numbers: the sharpe ratio, *Journal of Risk and Insurance* **73**(4), 673–686.
- [14] Schrager, D. (2006). Affine stochastic mortality, *Insurance: Mathematics and Economics* **38**, 81–97.
- [15] Denuit, M. & Dhaene, J. (2006). *Comonotonic Bounds on the Survival Probabilities in the Lee-Carter Model for Mortality Projection*, Working paper.
- [16] Milevsky, M.A. & Young, V.R. (2007). Asset allocation and annuitization, *Journal of Economic Dynamics and Control* **31**(9), 3138–3177.
- [17] Lee, R.D. & Carter, L.R. (1992). Modeling and forecasting U.S. mortality, *Journal of the American Statistical Association* **87**, 659–671.
- [18] Olivieri, A. (2001). Uncertainty in mortality projections: an actuarial perspective, *Insurance: Mathematics and Economics* **29**, 239–245.
- [19] Dahl, M.H. (2004). Stochastic mortality in life insurance: market reserves and mortality-linked insurance contracts, *Insurance: Mathematics and Economics* **35**, 113–136.
- [20] Renshaw, A.E. & Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors, *Insurance: Mathematics and Economics* **38**(3), 556–570.
- [21] Cairns, A.J.G., Blake, D. & Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty, *Journal of Risk and Insurance* **73**(4), 687–718.
- [22] Biffis, E. & Millossovich, P. (2006). The fair valuation of guaranteed annuity options, *Scandinavian Actuarial Journal* **2006**(1), 23–41.
- [23] Boyle, P.P. & Hardy, M.R. (2003). Guaranteed annuity options, *ASTIN Bulletin* **33**, 125–152.
- [24] Blake, D. & Burrows, W. (2001). Survivor bonds: helping to hedge mortality risk, *The Journal of Risk and Insurance* **68**(2), 339–348.
- [25] Feldstein, M. & Rangelova, E. (2001). Individual risk in an investment-based social security system, *American Economic Review* **91**(4), 1116–1125.
- [26] Brown, J.R. & Warshawsky, M.J. (2001). *Longevity-Insured Retirement Distributions from Pension Plans: Market and Regulatory Issues*, NBER Working Paper #8064.

MOSHE A. MILEVSKY

Low-Dose Extrapolation

Extrapolation, in its most basic statistical sense, is the construction of inferences about a stochastic phenomenon based on data unassociated or only partially associated with the phenomenon. As Hertzberg [1] aptly puts it, the process involves making “inferences about data-poor scenarios . . . from studies of data-rich scenarios”. A variety of applications in quantitative risk assessment involve some form of extrapolation. A well-known, and perhaps infamous, example occurs in health risk assessment, where a major goal is estimation of the severity and likelihood of damage from hazardous stimuli to humans [2]. To identify the damaging effects of a hazardous agent, investigators often perform bioassays on small mammals (*see Dose-Response Analysis*). When the experiment is conducted as a screen for certain toxic effects [3, 4], there is typically a relatively short time span available for the animals to exhibit the toxicity. Thus the design of the assay will require the data be taken at fairly high-dose levels. In these cases, decisions on occupational or environmental exposure standards to the toxic agent must be extrapolated from the high-dose exposure data. Similar issues can occur in select human observational studies of toxic exposures [2, 5].

These types of *low-dose extrapolations* can result in suspect inferences, however, if the lack of observed low-dose information is substantial [6–8]; indeed, considerable research attention has been directed at improving low-dose extrapolations in risk analysis [9–11]. A common assumption is that the model structure is essentially unchanged at low-dose levels. Unfortunately, numerous gaps remain in methodological understanding and implementation of low-dose extrapolations when used to support quantitative risk assessments, owing in part to excess variation and uncertainty in low-dose risk estimates from laboratory or human population studies; see, e.g., [12] or [13], respectively. Because of these gaps, the general problem of low-dose risk extrapolation remains an open issue from both a statistical and a public health policy perspective.

A particular problem connected with this occurs when estimation centers on dose(s) related to a particular, small health effect level. If the health effect/risk is small enough to be regarded as nonadverse or “acceptable”, it is likely to be far below the observed

range of the response data. As a result, the terminology can seem misleading: estimation and inferences are conducted on the dose scale, but the extrapolation is actually conducted from “high” observable effects/levels of risk to much lower effects/levels of risk. (Perhaps the term *low-effect extrapolation* would be more adequate!)

Recent advances in quantitative risk-benchmark analysis [14] (*see Benchmark Analysis*) and estimation of benchmark doses (BMDs) [15–17] have addressed some of these issues, although further developments are needed. Model selection and model quality assessment are particularly important in this respect: it is well recognized that many models can produce reasonably stable point estimates at higher doses, but can break down or diverge greatly at low doses corresponding to very low response rates [18, 19]. As a result, many laboratory studies limit estimation to (benchmark) risk levels in the 1–10% range, and not much lower [20]. When the data relate to a serious health effect such as cancer or developmental damage, this level of risk/response is usually very high. For example, regulations pertinent to carcinogen exposures are often considered at risks at or below 10^{-3} (i.e., one in a thousand) or 10^{-4} (i.e., one in ten thousand); but, toxicology data are often quantal, and as such they rarely produce observed rates of response below 1/50 or 1/100. (This is of less concern when data are taken on a continuous scale [21] and the benchmark effect is based on a percent change in response, e.g., a 5% change in body weight. In these cases, the benchmark is often much more likely to actually have been observed in the data, and as such no “low-dose” extrapolation is required.)

Developments in biological models or other forms of mechanistic modeling for adverse-outcome data may lead to improvements in such low-dose/low-effect extrapolations. This may be an area of future advancement [1, 22, 23], if current limitations in the application of such models can be overcome [10, 24]; this includes issues with the assumption of thresholds [25, 26] and/or approximate low-dose linearity in the dose–response [27, 28]. Connected with this, the increasing use of model averaging [29] and other ways of incorporating model uncertainty [30] may also prove advantageous.

It is also worth emphasizing that the issue of extrapolation in quantitative risk assessment includes and is extended to challenges well past that of

low-dose/low-effect estimation. Issues of inter and intraspecies extrapolation come quickly to mind [31, 32], along with potency estimation [33, 34] and extrapolations necessary in ecological risk assessment [35]. Hertzberg [1], mentioned earlier, gives a useful review of the larger issue; also see Preston [36].

Acknowledgments

Thanks are due Dr. Wout Slob for his helpful suggestions on this topic. Preparation of this material was supported in part by grant #R01-CA76031 from the U.S. National Cancer Institute, and grant #RD-83241901 from the U.S. Environmental Protection Agency. Its contents are solely the responsibility of the author and do not necessarily reflect the official views of these various agencies.

References

- [1] Hertzberg, R.C. (2002). Extrapolation, in *Encyclopedia of Environmetrics*, A.H. El-Shaarawi & W.W. Piegorsch, eds, John Wiley & Sons, Chichester, Vol. 2, pp. 732–739.
- [2] Coherssen, J.J. & Covello, V.T. (1989). *Risk Analysis: A Guide to Principles and Methods for Analyzing Health and Environmental Risks*, Executive Office of the President, Washington, DC.
- [3] Haseman, J.K. (1984). Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies, *Environmental Health Perspectives* **58**, 385–392.
- [4] Haseman, J.K. (1985). Issues in carcinogenicity testing: dose selection, *Fundamental and Applied Toxicology* **5**, 66–78.
- [5] Brenner, D.J. & Sachs, R.K. (2002). Do low dose-rate bystander effects influence domestic radon risks? *International Journal of Radiation Biology* **78**, 593–604.
- [6] Guess, H.A. & Crump, K.S. (1976). Low-dose extrapolation of data from animal carcinogenicity experiments – analysis of a new statistical technique, *Mathematical Biosciences* **32**, 15–36.
- [7] Krewski, D. & Kovar, J. (1982). Low dose extrapolation under single parameter dose–response models, *Communications in Statistics: Simulation and Computation* **11**, 27–45.
- [8] Portier, C.J. (1989). Quantitative risk assessment, in *Carcinogenicity and Pesticides. Principles, Issues, and Relationships*, N.N. Ragsdale & R.E. Menzer, eds, American Chemical Society, Washington, DC, pp. 164–174.
- [9] Krewski, D. & van Ryzin, J. (1981). Dose–response models for quantal response toxicity data, in *Statistics and Related Topics*, M. Csörgö, D.A. Dawson, J.N.K. Rao & A.K.M.E. Saleh, eds, North-Holland, Amsterdam, pp. 201–231.
- [10] Crump, K.S. (1994). Limitations of biological models of carcinogenesis for low-dose extrapolation, *Risk Analysis* **14**, 883–887.
- [11] Piegorsch, W.W., West, R.W., Pan, W. & Kodell, R.L. (2005). Low-dose risk estimation via simultaneous statistical inferences, *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **54**, 245–258.
- [12] Portier, C.J. & Kaplan, N.L. (1989). Variability of safe dose estimates when using complicated models of the carcinogenic process, *Fundamental and Applied Toxicology* **13**, 533–544.
- [13] Nussbaum, R.H. & Köhlein, W. (1994). Inconsistencies and open questions regarding low-dose health effects of ionizing radiation, *Environmental Health Perspectives* **102**, 656–667.
- [14] Crump, K.S. (2002). Benchmark analysis, in *Encyclopedia of Environmetrics*, A.H. El-Shaarawi & W.W. Piegorsch, eds, John Wiley & Sons, Chichester, Vol. 1, pp. 163–170.
- [15] Gaylor, D.W., Kodell, R.L., Chen, J.J. & Krewski, D. (1999). A unified approach to risk assessment for cancer and noncancer endpoints based on benchmark doses and uncertainty/safety factors, *Regulatory Toxicology and Pharmacology* **29**, 151–157.
- [16] Morales, K.H. & Ryan, L.M. (2005). Benchmark dose estimation based on epidemiologic cohort data, *Environmetrics* **16**, 435–447.
- [17] Piegorsch, W.W. & West, R.W. (2005). Benchmark analysis: shopping with proper confidence, *Risk Analysis* **25**, 913–920.
- [18] Guess, H.A., Crump, K.S. & Peto, R. (1977). Uncertainty estimates for low-dose extrapolation of animal carcinogenicity data, *Cancer Research* **37**, 3475–3483.
- [19] Barnes, D.G., Daston, G.P., Evans, J.S., Jarabek, A.M., Kavlock, R.J., Kimmel, C.A., Park, C. & Spitzer, H.L. (1995). Benchmark dose workshop – criteria for use of a benchmark dose to estimate a reference dose, *Regulatory Toxicology and Pharmacology* **21**, 296–306.
- [20] Crump, K.S. (1984). A new method for determining allowable daily intake, *Fundamental and Applied Toxicology* **4**, 854–871.
- [21] Slob, W. (2002). Dose-response modeling of continuous endpoints, *Toxicological Sciences* **66**, 298–312.
- [22] Daston, G.P. (1997). Advances in understanding mechanisms of toxicity and implications for risk assessment, *Reproductive Toxicology* **11**, 389–396.
- [23] Clewell, H.J. (2005). Use of mode of action in risk assessment: past, present, and future, *Regulatory Toxicology and Pharmacology* **42**, 3–14.
- [24] Swenberg, J.A., Richardson, F.C., Boucheron, J.A., Deal, F.H., Belinsky, S.A., Charbonneau, M. & Short, B.G. (1987). High- to low-dose extrapolation: critical determinants involved in the dose–response of carcinogenic substances, *Environmental Health Perspectives* **76**, 57–64.
- [25] Purchase, I.F.H. & Auton, T.R. (1995). Thresholds in chemical carcinogenesis, *Regulatory Toxicology and Pharmacology* **22**, 199–205.

- [26] Slob, W. (1999). Thresholds in toxicology and risk assessment, *International Journal of Toxicology* **18**, 259–268.
- [27] Crump, K.S., Hoel, D.G., Langley, C.H. & Peto, R. (1976). Fundamental carcinogenic processes and their implications for low dose risk assessment, *Cancer Research* **36**, 2973–2979.
- [28] Crawford, M. & Wilson, R. (1996). Low-dose linearity: the rule or the exception? *Human and Ecological Risk Assessment* **2**, 305–330.
- [29] Moon, H., Kim, H.-J., Chen, J.J. & Kodell, R.L. (2005). Model averaging using the Kullback information criterion in estimating effective doses for microbial infection and illness, *Risk Analysis* **25**, 1147–1159.
- [30] Kang, S.-H., Kodell, R.L. & Chen, J.J. (2000). Incorporating model uncertainties along with data uncertainties in microbial risk assessment, *Regulatory Toxicology and Pharmacology* **32**, 68–72.
- [31] Kuo, J., Linkov, I., Rhomberg, L., Polkanov, M., Gray, G. & Wilson, R. (2002). Absolute risk or relative risk? A study of intraspecies and interspecies extrapolation of chemical-induced cancer risk, *Risk Analysis* **22**, 141–157.
- [32] Clayson, D.B. (1991). Trans-species extrapolation in carcinogenesis, in *Statistics in Toxicology*, D. Krewski & C. Franklin, eds, Gordon and Breach, New York, pp. 653–671.
- [33] Allen, B.C., Crump, K.S. & Ship, A.M. (1988). Correlation between carcinogenic potency of chemicals in animals and humans, *Risk Analysis* **8**, 531–544.
- [34] Crouch, E.A.C. (1996). Uncertainty distributions for cancer potency factors: laboratory animal carcinogenicity bioassays and interspecies extrapolation, *Human and Ecological Risk Assessment* **2**, 103–129.
- [35] Chapman, P.M., Fairbrother, A. & Brown, D. (1998). A critical evaluation of safety (uncertainty) factors for ecological risk assessment, *Environmental Toxicology and Chemistry* **17**, 99–108.
- [36] Preston, R.J. (2005). Extrapolations are the achilles heel of risk assessment, *Mutation Research* **589**, 153–157.

Related Articles

Environmental Health Risk

Environmental Risks

WALTER W. PIEGORSCH

Maintenance Modeling and Management

Maintenance management and modeling is, essentially, concerned with describing a system in terms of its constituent parts and associating maintenance actions or rules with these parts. This view was first formalized by Gits [1]. When resolving a system into its constituent parts, it is helpful to use the following definitions of Ascher and Feingold [2]: a *component* is defined as an element of a maintained system for which it is reasonable to investigate a bespoke maintenance action. A *socket* is defined as the interface of a component with the maintained system; together with the component it provides a defined operational function within the system for which failure can be defined in terms of its impact on the system performance. The *system* is the totality of its parts: sockets, components, and their interrelationships that together provide some product or service. A system may be represented as a connected set of sockets into which components are installed. Within this framework, it is the role of maintenance management and modeling to resolve a system into its constituent sockets and components, to assign maintenance actions appropriate to each component, and to combine these actions into maintenance schedules or rules in an optimal manner. The resolution of the system and corresponding set of maintenance rules is the maintenance design.

It is important to recognize that maintenance management of a system may differ depending on one's viewpoint. Thus, a maintenance design is not unique. An organization that uses the operational function of a system may propose a different maintenance design to an organization that supplies the operational function. Differing organizational drivers and resources will imply differing maintenance designs, although the techniques described in this article are applicable regardless of viewpoint and whether maintenance or even the provision of operational function is outsourced. For example, an organization that provides urban, underground transportation may – (a) own and maintain its station escalators, or (b) own its escalators but outsource their maintenance to the original equipment manufacturer (OEM) or a third party, or (c) outsource escalator function to the OEM or a third party. Thus, with respect to (c), the transport

organization is not concerned at all with escalator equipment but merely has functional requirements that it demands (for payment) from the supplier. In each of the three cases, a maintenance design is necessary, but the owners of the maintenance design will differ. In (a), the owner is the user, in (c) it is the supplier, and in (b) we would expect the user and supplier to share ownership. While the techniques used to develop the design should not differ, the objectives that are used to define the system and set maintenance rules may differ.

Appropriate maintenance actions will be a matter of engineering judgment regarding the system of interest. Generally, the following three levels of maintenance intervention can be distinguished: inspection, service (for example, routine adjustment, cleaning), and replacement. Here, by replacement we mean component replacement; where maintenance interventions and replacements are combined into a program of maintenance to be carried out on the system at a particular point in time, then the term *refurbishment* or *overhaul* might be used to describe the program of maintenance depending on its scale. Other actions that are maintenance driven but are not strictly maintenance activities are operator training, and redesign and retrofitting. In the former, functional degradation resulting from inappropriate use of plant may be moderated. As an example of redesign and retrofitting, the Mass Transit Rail Corporation Limited (MTRCL) of Hong Kong is concerned with escalator performance at its stations and found that the redesign and retrofitting of escalator control systems would allow a significant reduction in stoppages owing to transient power dips [3]. Although the retrofitting of new control systems at MTRCL was maintenance driven, funding for the work did not come from the maintenance budget but was part of a large capital expenditure program. Complete system replacement may also be maintenance driven, and therefore, a consideration in the maintenance design. As alluded to, the funding for such action would not normally come from a maintenance budget. However, from a maintenance modeling point of view, system replacement may be regarded as the extreme case of redesign and retrofitting, since with system replacement one would expect a significant improvement in functional capability (which may not be the case with overhaul).

The extent of the resolution of a system into its constituent parts is a key consideration in the

maintenance design. If the resolution is too fine, the collection of all maintenance tasks will be too large and complex to manage effectively. If it is too coarse, the maintenance tasks will be suboptimal. For example, a complete pump might be replaced when renewal of seals and bearings would suffice. The principles of this resolution are an interesting avenue for study. Much of the maintenance modeling that seeks “optimal” policy in relation to a particular component function typically assumes that the system has been resolved at an appropriate level. Also, appropriate definitions of operational function and failure have to be agreed upon by parties that own the maintenance design. As for the escalators, MTRCL would regard an escalator stoppage owing to a power dip as a failure of operational function. However, these events would not, typically, be recorded as escalator failures by the OEM engineers.

Maintenance design in Gits [1] is conceptual. Labib [4] proposes a practical approach. Here a large complex system is resolved into components that possess operational failure characteristics (*see Systems Reliability*), which can be summarized in terms of a small number of key measures. Two such measures are, for example, average downtime per failure (downtime) and number of failures per unit time (frequency) (*see Managing Infrastructure Reliability, Safety, and Security; Common Cause Failure Modeling*). Other alternative measures may be used or the approach can be extended to higher dimensions. Following a standard Pareto analysis (*see Supra Decision Maker*), components are then classified by these measures and plotted on a multidimensional Pareto chart. Such a chart is conceptualized in Figure 1, and classes of maintenance actions that would be appropriate in general are superimposed on this chart. The interpretation of these maintenance actions is as follows. A system which is experiencing frequent, minor stoppages (top left) is likely to be investigated within a program of continuous engineering improvement. This may be considered as a quality management issue, and operator training is frequently a key response. Components incurring frequent, high downtime failures typically require urgent redesign, as such failures have significant impact on system performance. For example, Amasaka and Osaki [5] describe the redesign of an oil seal component in an engine following frequent warranty claims. Low frequency and low downtime components are allowed to operate to failure. Low

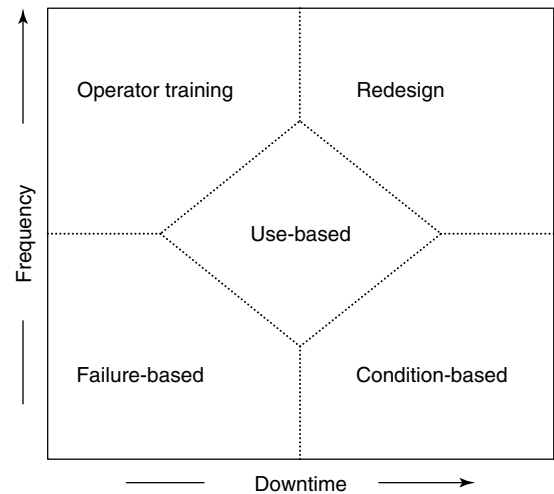


Figure 1 Schematic classification diagram of appropriate maintenance actions based on two decision criteria – average downtime per failure and frequency of failure, with indicative maintenance actions

frequency and high downtime events are likely to be seen by engineers as a prediction problem; condition-based maintenance is expected to focus on such events and the underlying components. A component for which the required maintenance policy is less clear is the subject of traditional ongoing maintenance schedules (use-based maintenance). An example of such a chart is shown in Figure 2 for the principal components of an extrusion press that has been used in the manufacture of copper products for a number of years. It is noticed that no components lie in the maintenance driven redesign region, which is expected – given the maturity of the system here.

With improvements achieved over time, the classification is updated with units previously classified as use-based maintenance becoming the focus of condition-based maintenance, say. In this way, the notion of continuous improvement in the performance of the system can be visualized. Also, higher dimensional Pareto charts may be used when other failure characteristics, such as the cost of lost production, are considered.

One might generalize the above approach within a risk assessment framework. The frequency of failure may be interpreted as a probability, p , and the downtime as the consequences, C (the consequences might equivalently be measured in other terms such as cost or some safety-related measure). With risk defined

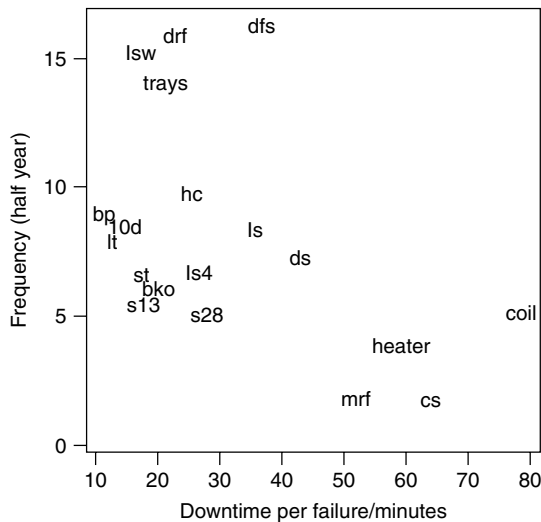


Figure 2 Two-dimensional Pareto diagram for an extrusion press; frequency and downtime for 19 components. Figure labels refer to component faults, e.g., drf, cutting die rotation fault [Reproduced from [6]. © Palgrave Macmillan, 1998.]

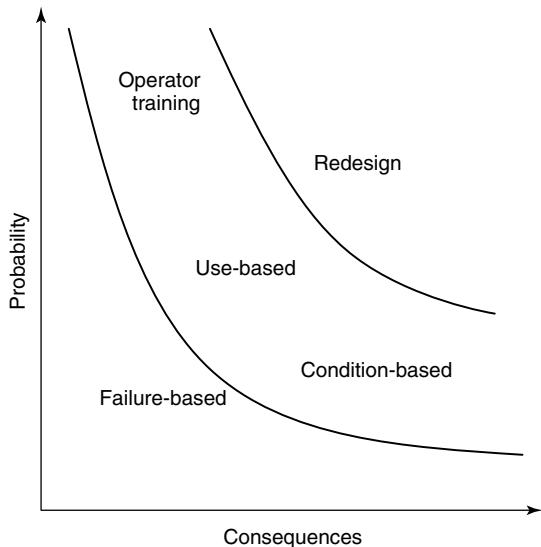


Figure 3 Risk-based maintenance – schematic classification diagram of appropriate maintenance actions based on risk, with lines of equal risk indicated

by $R = p \times C$, Figure 1 may then be redrawn as shown in Figure 3 with lines of equal risk indicated. Low risk components are thus subject to failure-based

maintenance. High risk components are subject to redesign and retrofiting. Components that are in the medium risk region and not subject to improvement through operator training should then be the focus of either use- (or time-) based maintenance or condition-based maintenance (see **Reliability Growth Testing**), and many models have been developed to aid decision making in these contexts.

The simplest of the use-based models are those concerned with component replacement, notably block replacement and age-based replacement. In age-based replacement, the age of a component is monitored and replacement is initiated when the age reaches some critical level, T . This critical level may be naturally expressed in terms of the hazard function [7]. To choose this critical level optimally, information about the time to failure distribution of the component is required; the optimality criterion must also be specified, for example, as the long-run cost per unit time or the mean time between component failures. When component age is not monitored, then replacements may be scheduled periodically as with block replacement [8], or modified block replacement [9], or according to the constraints of maintenance budget. For an extensive discussion of optimality criteria in age-based replacement, see [10]. Component replacement times may also be determined using the techniques of reliability-centered maintenance [11].

A natural extension of age-based replacement is the modeling of the hazard of failure as a function of age and a condition indicator. Proportional hazards modeling (see **Lifetime Models and Risk Assessment; Hazard and Hazard Ratio; Statistical Arbitrage**) is an established method of looking at this problem, and the critical region for replacement can be specified in a relatively straightforward manner [12, 13]. The proportional hazards model assumes

$$h\{t, x(t)\} = h_0(t) \exp \{\beta x(t)\} \quad (1)$$

where $h_0(t)$ is the baseline hazard rate depending on system age only and $\exp\{\beta x(t)\}$ is a multiplicative risk factor that models the effect of system characteristics, with $x(t)$ being a vector of time-dependent covariates (the condition indicator), and β a vector of parameters. If the baseline hazard is unspecified, then equation (1) is the semiparametric proportional hazards model (PHM) [14]. A fully parametric proportional hazards model can be obtained by specifying $h_0(t)$ with Weibull hazard, for example: $h_0(t) = \alpha(t/\eta)^{\beta-1}/\eta$. Given a sample

of lifetime histories, ending in failure or preventive replacement (censored failure time), parameters of this model can be estimated. Given failure and preventive replacement costs and an increasing hazard, an optimal control limit for replacement, d^* , for the hazard (and hence the critical region for replacement) can be determined. Finally, for this model to be useful for the prediction of time to replacement, a model for the development of the condition indicator, $x(t)$, is required such as a discrete Markov process, $\Pr\{X(t+1) \leq x | X(1), \dots, X(t)\} = G\{x | t, X(t)\}$ [12], or a gamma process [15]. This model will then allow the prediction of future condition, $X(t+k)$ given condition history $X(1), \dots, X(t)$.

The approach may be considered natural because optimum policy would reduce to that of age-based replacement when true condition appears unrelated to the condition indicator on the basis of available data. Furthermore, the aging of the system can be accommodated by including a subset of covariates which reflect system age, e.g., number of previous repairs/replacements since new. The key problem, however, is having sufficient data on failures and covariates to estimate model parameters and hence adequately estimate the critical region.

When a failure threshold, c (see **Simulation in Risk Management; Extreme Values in Reliability**), for the component condition, Y , is specified (so that if $Y > c$ then the component fails), an alternative modeling approach may be used. For simple cases, Y may be directly observable (e.g., tyre tread depth), and it is sufficient to model this wear process using a suitable stochastic process, such as the gamma process [16]. This has been referred to as a direct monitoring process [17]. At a particular decision point, given the development of the condition to date, y_{t-} (condition history), the problem is to estimate the residual life or remaining time to failure, that is, the minimum time T , such that $Y_T > c | y_{t-}$. An estimate of the residual life distribution (see **Pricing of Life Insurance Liabilities; Reliability Growth Testing**), $f(T | y_{t-})$ can be used to optimize the time to replacement given a suitable decision criterion. This optimal time to replacement can be updated dynamically as more condition information becomes available. In practice, the decision would be a choice between replacing before the next monitoring check (inspection) or otherwise. Where the process is measured continuously, the replacement problem is trivial involving replacement at the moment $Y = c$. The

current condition may be unobserved, but an indicator of condition, X , is typically observed and used as a surrogate. Provided the condition, Y , can be calibrated in terms of the condition indicator, then in principle the method can proceed as if the condition were directly observable. The calibration is done typically by relating the condition indicator to the condition observed at preventive replacements or on failure, with the latter expected to be rare events. This method was developed for monitoring the refractory thickness in an induction furnace [18]. It has also been used in the vibration analysis of bearings [19].

The specification of a failure threshold may not be feasible. It is necessary to specify a warning threshold to proceed. If the condition indicator is above the threshold, then the component is defective and should be replaced. If the condition indicator is monitored continuously, then the decision problem is again trivial. If the condition indicator is monitored periodically, and the condition indicator only takes the values 0 and 1 (depending on whether the measured condition is above or below the warning threshold), then we will be naturally concerned with how often to monitor. A two-phase or delay-time model [20, 21] (see **Repair, Inspection, and Replacement Models**) can be used to optimize the monitoring interval. In addition to the times between inspection, the warning level may also be regarded as a decision variable [22]. These models balance the cost of monitoring or inspection and preventive replacement against the cost of failure, given the probability distributions of the time at which the condition indicator crosses the warning threshold and the time between the condition indicator crossing of the warning threshold and the consequent failure. The former time is called the *defect initiation time*, with distribution function G , say, and the latter is called the *delay time* with distribution function F , say. Let the costs of failure, monitoring inspection, and preventive replacement at monitoring inspection be C_f , C_i , and C_r , respectively. Suppose, monitoring checks are carried out every Δ time units, and that the component is replaced if the condition indicator is above the warning threshold. Under these assumptions, the long-run cost per unit time is

$$\frac{(C_f - C_r - C_i)\{1 - S(\Delta)\} + C_r G(\Delta) + C_i}{\int_0^\Delta S(t) dt} \quad (2)$$

where $S(t) = 1 - \int_0^t g(u)F(t-u) du$. Here $S(t)$ is the survival distribution (or reliability function) (see **Options and Guarantees in Life Insurance; Risk Classification/Life; Default Correlation**) of the time to failure, and hence the convolution of g , the density of defect initiation time, and the delay-time distribution. The expression (2) can be minimized to determine the optimal value of Δ . We might use the term *warning threshold condition-based maintenance* for the resulting policy. The model can be extended to consider imperfect monitoring and systems for which the delay time is related to the initiation time [23].

Where the cost of monitoring is insignificant, then on-line continuous monitoring will be appropriate, provided the warning threshold is set to allow sufficient time to schedule replacement subsequent to the alarm. The setting of the warning threshold may require modeling. Such on-line continuous monitoring systems are common (e.g., an oil level indicator light).

Given observations on the development of the condition indicator, we may wish to predict the time to the crossing of the warning threshold. Appropriate models are then, as discussed above, with the threshold giving a subtly different interpretation. Care has to be taken with definitions because what the plant operator regards as a failure threshold may be regarded as a warning threshold by modelers, and *vice versa*.

Use-based maintenance policies can be considered as special cases of condition-based maintenance policies. For example, age-based replacement can be considered as a proportional hazards condition-based maintenance policy with an uninformative condition indicator. Routine scheduled inspection maintenance can be considered as a warning threshold policy, as could the modified block replacement with the age being the "condition indicator". Block replacement is itself a special case of modified block replacement, with replacement regardless of age.

The models described thus far assume that the component is renewed on replacement. Where components are not renewed but are merely repaired, then these models may only provide approximate solutions. This assumption is relaxed in models of imperfect maintenance [24, 25].

The cost of data collection is a significant issue in the implementation of a maintenance design. Sufficient data (relating to failures) for estimating the parameters of the component failure models is

unlikely to exist in most, if not all, practical cases. Collecting data may be an option, but few operators are prepared to let components run to failure, and often, replaced components are not in a "failed" state. Also, information most closely related to such condition is likely to be more costly to collect. Oil analysis of a gearbox sump is significantly easier than an internal inspection. Thus, it may be necessary to build the cost of data collection into decision models, and it is important to consider whether condition monitoring, for example, will be effective in reducing overall costs. The costs of data collection need to be balanced against the expected gains as a result of operating a policy, which is nearer to optimal.

Given the replacement and inspection times for the components in a system, these may be combined into schedules that take account of component dependencies. Thus, certain component groups, due to their physical proximity, may be most sensibly maintained simultaneously. Models for scheduling maintenance for multicomponent systems in an opportunistic fashion are also well developed [26]. Wang [27] reviews these models (and also many other models for single-component systems). Sophisticated maintenance policies for two-component systems with failure mechanisms that interact have been developed [28–30]. Given that many systems are large and complex (e.g., oil production platforms, nuclear power plants), there is scope for the use of expert systems in the designing and scheduling of maintenance activity. Such work is the focus of ongoing research [31].

The static policies implied by block replacement and routine preventive and inspection maintenance are relatively straightforward to schedule, even for complex systems. However, where age-based replacement or condition-based maintenance is implemented, such systems will need to schedule maintenance activities in a dynamic way [32], as the execution times of certain activities will be continually updated as components age and degrade. Opportunities for combining the execution of maintenance activities on components that are dependent will also evolve over time. These aspects of maintenance design present challenges for maintenance management information systems.

References

- [1] Gits, C.W. (1992). Design of maintenance concepts, *International Journal of Production Economics* **24**, 217–226.

- [2] Ascher, H. & Feingold, H. (1984). *Repairable Systems Reliability*, Marcel Dekker, New York.
- [3] Scarf, P.A., Dwight, R., McCusker, A. & Chan, A. (2006). Asset replacement for an urban railway using a modified two-cycle replacement model, *Journal of the Operational Research Society* **58**, 1123–1137. DOI: 10.1057/palgrave.jors.2602288.
- [4] Labib, A.W. (1998). World class maintenance using a computerised maintenance management system, *Journal of Quality in Maintenance Engineering* **4**, 66–75.
- [5] Amasaka, K. & Osaki, S. (2003). Reliability of oil seal for transaxle – a science SQC approach, in *Case Studies in Reliability and Maintenance*, W.R. Blischke & D.N.P. Murthy, eds, John Wiley & Sons, pp. 571–588.
- [6] Christer, A.H., Wang, W., Sharp, J. & Baker, R.D. (1998). A case study of modeling preventive maintenance of a production plant using subjective data, *Journal of the Operational Research Society* **49**, 210–219.
- [7] Berg, M. (1995). The marginal cost analysis and its application to repair and replacement policies, *European Journal of Operational Research* **82**, 214–224.
- [8] Barlow, R. & Proschan, F. (1965). *Mathematical Theory of Reliability*, John Wiley & Sons, New York.
- [9] Berg, M. & Epstein, A. (1976). A modified block replacement policy, *Naval Research Logistics Quarterly* **23**, 15–24.
- [10] Scarf, P.A., Dwight, R. & Al-Musrati, A. (2005). On reliability criteria and the implied cost of failure for a maintained component, *Reliability Engineering and System Safety* **89**, 199–207.
- [11] Moubray, J. (1997). *Reliability-Centred Maintenance*, 2nd Edition, Industrial Press, New York.
- [12] Makis, V. & Jardine, A.K.S. (1991). Optimal replacement in the proportional hazards model, *INFOR* **30**, 72–183.
- [13] Vlok, P.J., Coetzee, J.L., Banjevic, D., Jardine, A.K.S. & Makis, V. (2002). Optimal component replacement decisions using vibration monitoring and the proportional hazards model, *Journal of the Operational Research Society* **53**, 193–202.
- [14] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.
- [15] Wang, W., Scarf, P.A. & Smith, M.A.J. (2000). On the application of a model of condition-based maintenance, *Journal of the Operational Research Society* **51**, 1218–1227.
- [16] Park, K.S. (1988). Optimal wear-limit replacement with wear-dependent failure, *IEEE Transactions on Reliability* **37**, 293–294.
- [17] Christer, A.H. & Wang, W. (1995). A simple condition monitoring model for a direct monitoring process, *European Journal of Operational Research* **82**, 258–269.
- [18] Christer, A.H., Wang, W. & Sharp, J. (1997). A state space condition monitoring model for furnace erosion prediction and replacement, *European Journal of Operational Research* **101**, 1–14.
- [19] Wang, W. (2002). A model to predict the residual life of rolling element bearings given monitoring condition information to date, *IMA Journal of Management Mathematics* **13**, 3–16.
- [20] Coolen, F.P.A. & Dekker, R. (1995). Analysis of a 2-phase model for optimization of condition monitoring intervals, *IEEE Transactions on Reliability* **55**, 505–511.
- [21] Christer, A.H. (1999). Developments in delay time analysis for modeling plant maintenance, *Journal of the Operational Research Society* **50**, 1120–1137.
- [22] Grall, A., Berenguer, C. & Dieulle, L. (2002). A condition-based maintenance policy for stochastically deteriorating systems, *Reliability Engineering and System Safety* **76**, 167–180.
- [23] Baker, R.D. & Wang, W. (1992). Developing and testing the delay time model, *Journal of the Operational Research Society* **44**, 361–374.
- [24] Pham, H. & Wang, H. (1996). Imperfect maintenance, *European Journal of Operational Research* **94**, 425–438.
- [25] Jack, N. (1998). Age reduction models for imperfect maintenance, *IMA Journal of Mathematics Applied in Business and Industry* **9**, 347–354.
- [26] Dekker, R., Van der Duyn Schouten, F.A. & Wildeman, R. (1997). A review of multi-component maintenance models with economic dependence, *Mathematical Methods of Operations Research* **45**, 411–435.
- [27] Wang, H. (2002). A survey of maintenance policies for deteriorating systems, *European Journal of Operational Research* **139**, 469–489.
- [28] Murthy, D.N.P. & Nguyen, D.G. (1985). Study of two-component systems with failure interactions, *Naval Research Logistics Quarterly* **32**, 239–247.
- [29] Scarf, P.A. & Dearn, M. (1998). On the development and application of maintenance policies for a two-component system with failure dependence, *IMA Journal of Mathematics Applied in Business and Industry* **9**, 91–107.
- [30] Scarf, P.A. & Dearn, M. (2003). Block replacement policies for a two-component system with failure dependence, *Naval Research Logistics* **50**, 70–87.
- [31] Kobbacy, K.A.H. (2004). On the evolution of an intelligent maintenance optimization system, *Journal of the Operational Research Society* **55**, 139–146.
- [32] Wildeman, R., Dekker, R. & Smit, A.C.J.M. (1997). A dynamic policy for grouping maintenance activities, *European Journal of Operational Research* **99**, 530–551.

PHILIP A. SCARF

Managing Foodborne Risk

Health risks in the food supply often arouse strong political passions and polarized science-policy debates. From fear of genetically modified organisms (GMOs) and irradiated meats to rare but incurable fatal brain-wasting caused by beef contaminated with bovine spongiform encephalitis (BSE, or “mad cow” disease) to *Escherichia coli* outbreaks in retail foods to concern over antibiotic-resistant pathogens in meat, real and perceived risks associated with the food supply – and political and popular reactions to them – make headlines (*see Risk and the Media*).

What makes food safety risks so gripping? And how can methods of quantitative risk assessment (QRA) and risk management help to improve food safety? This article summarizes recent answers to these questions and surveys methods for quantifying and managing foodborne risks.

Foodborne Risks Depend on the Decisions of Many Parties

The risk of human illnesses caused by pathogenic (disease-causing) bacteria such as *Salmonella* or *E. coli* in food depends on the levels of care taken by multiple participants at steps throughout the chain of food production, distribution, and preparation. For meats, these steps can include farming (including practices for hygiene and microbial count control), transportation, slaughter, production and packaging, storage, retail, and food preparation in the kitchen – especially, proper cooking. Consumers with compromised immune systems (for example, patients with cancer, AIDS, or organ transplants) can be hundreds or thousands of times more at risk than consumers with healthy immune systems. The usual effects of foodborne illness include diarrhea and possibly fever, vomiting, and other symptoms of food poisoning. Occasionally, more serious harm, or death, may occur, depending on the pathogen and the victim.

In such a setting, where consumer risk varies greatly among individuals and depends on the actions of many different economic agents, allocating responsibility for maintaining food safety involves overlapping questions of economic efficiency, legal liability,

producer quality control, and consumer negligence or care in kitchen hygiene and in food handling and preparation. Consumers might feel that producers should guarantee a “safe” product (although eliminating all bacteria at retail may be a practical impossibility unless one relies on unpopular techniques such as irradiation); while producers may feel that consumers should exercise prudence and responsibility by cooking meat properly (which eliminates most risks from bacterial pathogens) and by maintaining good kitchen hygiene. Although government inspection and certification can help to assure reasonable microbial safety throughout much of the food chain, the question still remains of how much cost of care, responsibility, and liability should be required from each participant in the chain, given that the final risk per serving of food to a consumer depends on their joint actions.

QRA can help to improve decisions in this complex and contentious area by providing quantitative estimates of how proposed changes in the process will affect human health risks. Human health risks are typically quantified as expected number of illnesses, fatalities, illness-days, or quality-adjusted life years lost to illness per year (for population risks) or per capita-year (for individual risks). Different subpopulations, such as infants, the elderly, and people with weak immune systems (e.g., transplant, leukemia, or AIDS patients) may have distinct risks. A thorough QRA quantifies the individual and population risks for such subpopulations as well as for the whole population of concern (e.g., the whole population of a country).

Farm-to-Fork Models Quantify Microbial Loads in Food, *x*

One approach to quantifying the effects of different interventions on risks created by bacteria in food is to simulate the flow and growth, or reduction, of microbial loads of bacteria throughout the chain of steps leading from production to consumption. This is done by collecting data on the conditional frequency distribution of microbial loads leaving each step (e.g., transportation, processing, storage, etc.), given the microbial load entering that step. Microbial loads are expressed in units such as colony-forming units (CFUs) of bacteria per unit (e.g., per pound, per carcass, etc.) of food.

The discipline of applied microbiology supplies empirical growth curves and kill curves for log increase or log reduction, respectively, in microbial load from input to output of a step. These curves describe the output : input ratio (e.g., a most likely value and upper and lower statistical confidence limits) for the microbial load passing through a stage as a function of variables such as temperature, pH, and time.

A “farm-to-fork” simulation model can be constructed by concatenating many consecutive steps representing stages in the food production process. Each step receives a microbial load from its predecessor. It produces as output a microbial load value sampled from the conditional frequency distribution of the output microbial load, given the input microbial load, as specified by the microbial growth model describing that stage. Measured frequency distributions of microbial loads on animals (or other units of food) leaving the farm provide the initial input to the whole model. The key output from the model is a frequency distribution of the microbial load, x , of pathogenic bacteria in servings of food ingested by consumers.

Risk-reducing factors such as antimicrobial sprays and chilling during processing, freezing or refrigeration during storage, and cooking before serving are often modeled by corresponding reduction factors for microbial loads. (These may be represented as random variables, e.g., with lognormal distributions and geometric means and variances estimated from data.) The complete model is then represented by a product of factors that increase or decrease microbial loads, applied to the empirical frequency distribution of initial microbial loads on which the factors act. Running the complete farm-to-fork model multiple times produces a final distribution of microbial loads on servings eaten by consumers. Some farm-to-fork exposure models also consider effects of cross-contamination in the kitchen, if pathogenic bacteria are expected to be spread to other foods by poor kitchen hygiene practices (e.g., failure to wash a cutting board after use).

In summary, farm-to-fork simulation models can provide an estimate of the frequency distribution of microbial loads ingested by consumers in servings of food, as well as estimates of how these frequency distributions would change if different interventions (represented by changes in one or more of the step-specific factors increasing or decreasing microbial

load) were implemented. For example, enforcing a limit on the maximum time that ready-to-eat meats may be stored at delis or points of retail sale before being disposed of limits the opportunity for bacterial growth prior to consumption. Changing processing steps (such as scalding, chilling, antimicrobial sprays, etc.) can also reduce microbial loads. Such interventions shift the cumulative frequency distribution of microbial loads in food leftward, other things being held equal.

Dose–Response Models for Foodborne Pathogens, $r(x)$

Several parametric statistical models have been developed to describe the relation between quantity of bacteria ingested in food and the resulting probability of illness. One of the simplest is the following exponential dose–response relation:

$$r(x) = \Pr(\text{illness} \mid \text{ingest microbial load} = x \text{ CFUs}) \\ = 1 - e^{-\lambda x} \quad (1)$$

This model gives the probability that an ingested dose of x CFUs of a pathogenic bacterium will cause illness. $r(x)$ denotes this probability. The function $r(x)$ is a *dose–response curve* (see **Dose–Response Analysis**). λ is a parameter reflecting the potency of the exposure in causing illness. Sensitive subpopulations have higher values of λ than the general population.

More complex dose–response models (especially, the widely used beta-Poisson model) have two or more parameters, e.g., representing the population distribution of individual susceptibility parameter values and the conditional probability of illness given a susceptibility parameter. The standard statistical tasks of estimating model parameters, quantifying confidence intervals or joint confidence regions, and validating fitted models can be accomplished using standard statistical methods such as maximum-likelihood estimation (MLE) and resampling methods. The excellent monograph by Haas *et al.* [1] provides details and examples. It notes that “It has been possible to evaluate and compile a comprehensive database on microbial dose–response models.” Chapter 9 of this monograph provides a compendium of dose–response data and dose–response curves along with critical evaluations and results of validation studies for the following: *Campylobacter jejuni*

(based on human feeding study data), *Cryptosporidium parvum*, pathogenic *E. coli*, *E. coli* O157 : H7 (using *Shigella* species as a surrogate), *Giardia lamblia*, nontyphoid *Salmonella* (based on human feeding study data), *Salmonella typhosa*, *Shigella dysenteriae*, *Shigella flexneri*, *Vibrio cholerae*, Adenovirus 4, Coxsackie viruses, Echovirus 12, Hepatitis A virus, Poliovirus I (minor), and rotavirus. Thus, for many foodborne and waterborne pathogens of interest, dose–response models and assessments of fit are readily available.

Quantitative Risk Characterization for Microbial Risk Assessment and Risk Management

Sampling values of x from the frequency distribution predicted by a farm-to-fork model (expressed in units of bacteria per serving) and then applying the dose–response relation $r(x)$ to each sampled value of x produces a frequency distribution of the risk per serving in an exposed population. This information can be displayed in various ways to inform risk-management decision making.

For example, the expected number of illnesses per year in a population, expected illnesses per capita-year in the overall population and for members of sensitive subpopulations, and frequency distributions or upper and lower confidence limits around these expected values are typical outputs of a risk model. If particular decisions are being considered, such as a new standard for the maximum times and/or temperatures at which ready-to-eat meats can be stored before being disposed of, then plotting expected illnesses per year against the decision variables (i.e., maximum times or temperatures, in this example) provides the quantitative links between alternative decisions and their probable health consequences needed to guide effective risk management decision making.

Attribution-Based Risk Assessment and Controversies

In recent years, many efforts have been made to simplify the standard approach to microbial risk assessment just summarized. Because farm-to-fork exposure modeling and valid dose–response modeling can require data that are expensive and time consuming to collect, or that are simply not available when

risk-management decisions must be made, simpler approaches with less burdensome data requirements are desirable. It is tempting to use simple multiplicative models, such as the following:

$$\text{Risk} = \text{exposure} \times \text{dose-response factor} \times \text{consequence} \quad (2)$$

where *risk* is the expected number of excess illness-days per year, *exposure* is measured in potentially infectious meals ingested per year in a population, *dose–response factor* is the expected number of illnesses caused per potentially infectious meal ingested, and *consequence* is measured in illness-days caused per illness.

While such models have attractive simplicity, they embody strong assumptions that are not necessarily valid, and can thus produce highly misleading results. Specifically, the assessment of *dose–response factor* requires attributing some part of the causation of illness-days to *exposure*. Similarly, estimating the change in *dose–response factor* due to an intervention that changes microbial load may require guesswork. There is often no valid, objective way to make such attributions based on available data. The risk assessment model – and, specifically, the attribution of risk to particular food sources – may then become a matter of political and legal controversy.

For example, suppose that the *dose–response factor* is estimated by dividing the observed value of *risk* in a population in one or more years by the contemporaneous values of (*exposure* $\times$ *consequence*). Then, this value will *always* be nonnegative (since its numerator and denominator are both nonnegative). The model thus implies a nonnegative linear relation between *exposure* and *risk*, even if there is no causal relation at all (or is a negative one) between them. In reality, it is a frequent observation that some level of exposure to bacteria in food protects against risk of foodborne illnesses, for example, because of acquired immunity. Thus, the use of a simple multiplicative model implying a necessarily nonnegative linear relation between *exposure* and *risk* may be incorrect, producing meaningless results (or, more optimistically, extreme upper bounds on estimated risks) if the true relation is negative or nonlinear.

For this reason, great caution should be taken when using such simplified risk assessment models. In general, they may be useful in making rapid

calculations of plausible *upper bounds* in certain situations (for example, if the true but unknown dose–response relation between exposure and risk is convex, or upward curving), but should not be expected to produce accurate risk estimates unless they have been carefully validated [2].

Managing Uncertain Food Risks via Quality Principles: HACCP

Even if dose–response relations are uncertain and information needed to quantify microbial loads with accuracy and confidence is not available, it is often possible to apply process quality improvement ideas to control the microbial quality of food production processes. This approach has been developed and deployed successfully (usually on a voluntary basis) using the *hazard analysis and critical control points* (HACCP) approach summarized in Table 1. The main idea of HACCP is to first identify steps or stages in the food production process where bacteria can be controlled, and then to apply effective controls at those points, regardless of what the ultimate quantitative effects on human health risks may be. Reducing microbial load at points where it can be done effectively has proved very successful in reducing final microbial loads and improving food safety.

Discussion and Conclusions

This article has introduced key ideas used to quantify and manage human health risks from food contamination by bacteria. Somewhat similar ideas apply to other foodborne hazards, i.e., risk assessment can be carried out by modeling the flow of contaminants through the food production process (together with any increases or decreases at different steps), resulting in levels of exposures in ingested foods or drinks. These are then converted to quantitative risks using dose–response functions.

The practical successes of the HACCP approach provide a valuable reminder that QRA is not always a prerequisite for effective risk management. It may not be necessary to quantify a risk to reduce it. Reducing exposures at critical control points throughout the food production process can reduce exposure-related risk even if the size of the risk is unknown.

Where QRA can make a crucial contribution is in situations where there is doubt about whether an intervention is worthwhile. For example, QRA can reveal whether expensive risk-reducing measures are likely to produce correspondingly large health benefits. It may be a poor use of resources to implement expensive risk-reducing measures if the quantitative size of risk reduction procured thereby is very small. QRA methods such as farm-to-fork exposure modeling and dose–response modeling [1], or simpler upper-bounding approaches based on multiplicative models [2], can then be valuable in guiding effective

Table 1 Summary of seven HACCP principles^(a)

<i>Analyze hazards.</i>	Potential hazards associated with a food and measures to control those hazards are identified. The hazard could be biological, such as a microbe; chemical, such as a toxin; or physical, such as ground glass or metal fragments.
<i>Identify critical control points.</i>	These are points in a food's production – from its raw state through processing and shipping to consumption by the consumer – at which the potential hazard can be controlled or eliminated. Examples are cooking, cooling, packaging, and metal detection.
<i>Establish preventive measures with critical limits for each control point.</i>	For a cooked food, this might include minimum cooking temperature and time required to ensure the elimination of any harmful microbes.
<i>Establish procedures to monitor the critical control points.</i>	Such procedures might include determining how and by whom cooking time and temperature should be monitored.
<i>Establish corrective actions to be taken when monitoring shows that a critical limit has not been met with</i>	– for example, reprocessing or disposing of food if the minimum cooking temperature is not met with.
<i>Establish procedures to verify that the system is working properly</i>	– for example, testing time-and-temperature recording devices to verify that a cooking unit is working properly.
<i>Establish effective recordkeeping to document the HACCP system.</i>	This would include records of hazards and their control methods, the monitoring of safety requirements, and action taken to correct potential problems. Each of these principles must be backed by sound scientific knowledge: for example, published microbiological studies on time-and-temperature factors for controlling foodborne pathogens.

^(a) Reproduced from <http://www.cfsan.fda.gov/~lrd/bghaccp.html>. Food and Drug Administration (FDA), 2004

risk-management resource allocations by revealing the approximate sizes of the changes in human health risks caused by alternative interventions.

References

- [1] Haas, C.N., Rose, J.B. & Gerba, C.P. (1999). *Quantitative Microbial Risk Assessment*, John Wiley & Sons, New York.
- [2] Cox Jr, L.A. (2006). *Quantitative Health Risk Analysis Methods: Modeling the Human Health Impacts of Antibiotics Used in Food Animals*, Springer, New York.

Related Articles

Benchmark Dose Estimation

Detection Limits

Environmental Health Risk

LOUIS A. COX JR

Managing Infrastructure Reliability, Safety, and Security

Infrastructure and its Social Purpose

Infrastructure typically refers to physical systems and their institutional settings that support social and economic activity, and reliability is essential to meeting this need. Until the early 1980s, infrastructure had received relatively little systematic or comprehensive attention as a unified system. The relative inattention to infrastructure as a distinct area of focus was in part due to the absence of a single political constituency for infrastructure as a whole either in the form of supporters or dissidents. In contrast, many individual infrastructure areas had very active constituencies (e.g., antihighway, antiincinerator, and promass transit groups).

The Meaning of Infrastructure

The way in which infrastructure is defined influences our expectations of its reliability. Ironically, in spite of the inattention to infrastructure, it was reliability that popularized infrastructure as a unified theme in the early 1980s [1]. The term *infrastructure* as it is used today is a relatively recent concept. Prior to the 1980s, the word infrastructure was typically defined in the context of military installations. Tarr [2, p. 4] defines the components of infrastructure as “the sinews of the city: its road, bridge and transit networks; its water and sewer lines, and waste disposal facilities; its power systems; its public buildings; and its parks and recreation areas.” To a large extent infrastructure systems form the “links” and “nodes” that Kevin Lynch identifies as defining the sense of urban places [3]. The common element in these early definitions was a physical facility or system. However, infrastructure goes well beyond a physical dimension, and now has well-recognized economic, social, and political dimensions suitable for a public service. Paralleling the emergence of this broader conceptualization of infrastructure, the field once dominated by the engineering profession now includes many other disciplines [4].

Reliability and Performance

Infrastructure reliability, in its most general sense, refers to whether a system performs as expected; thus performance is central to the concept of reliability and the way in which it is managed (*see Systems Reliability*). The National Research Council (NRC) [5, p. 122] defines reliability as part of performance as “the likelihood that infrastructure effectiveness will be maintained over an extended period of time; the probability that service will be available at least at specified levels throughout the design lifetime of the infrastructure system.” Performance is generally the “carrying out of a task or fulfillment of some promise or claim” with a more specific definition reflective of the specific functions of infrastructure [5, p. 5]. The NRC underscores the relative or contextual nature of performance, that is, “performance must ultimately be assessed by the people who own, build, operate, use, or are neighbors to that infrastructure” [5, p. 5]. Performance is usually measured against benchmarks, guidelines, and standards (*see Benchmark Analysis*).

The Fragmentation of Infrastructure Performance Measures by Sector.

Each type of infrastructure traditionally had its own distinct performance measures for overall reliability. For example, in transportation, roadway performance often emphasized congestion [6] and safety along with many other roadway characteristics [7]. Examples of specific performance measures for congestion and safety are percentage of miles with service levels equal to or less than 0.70 (volume to capacity ratio) and number of fatalities per 100 million vehicle miles of travel, respectively [8, p. 218–219]. Transit measures incorporate operational measures such as mean distance between failures, defined as the number of miles a train travels before it breaks down. Electricity performance was dominated by a long list of operational standards reflected in oversight documents, such as the New York State (NYS) Public Service Commission review of the Queens blackout of 2006 and the special commission report for the massive August 2003 blackout that affected 50 million people in the United States and Canada [9]. Water supply and wastewater infrastructure follow professional guidance as well as regulatory standards established under the Safe Drinking Water Act (SDWA) and the Clean Water Act (CWA), which focus in part on

the capacity or the ability to maintain the required supply and quality of water, and systems that convey water or wastewater often follow professional standards related to leakage rates. Telecommunications is largely measured in terms of a number of different industry and regulatory standards depending on the particular type of telecommunications. The American Society of Civil Engineers (ASCE) maintains an annual scorecard for each individual infrastructure area as a broad measure of performance, and provides an aggregate measure across all of the categories as well. In 2005, the scores ranged from D – for water-related infrastructure sectors to C+ for solid waste, and the aggregate for all sectors was D with an estimated 5-year investment need of \$1.6 trillion (excluding security) [10]. This is an increase over the minimum of \$853 billion estimate (also probably excluding security) cited in 1998 [11]. The ASCE methodology to assess performance uses judgments of an expert panel, based on condition and capacity criteria, and is not risk-based, at least not explicitly.

The Need for Integrating Performance Measures across Infrastructure Sectors. Approaching infrastructure as a unified theme rather than in terms of its component areas is critical to managing reliability, since different types of infrastructures share many things in common. Different areas of infrastructure compete for the same capital investments, labor pools, construction sites, and most of all users or customers. Infrastructure sectors are connected to and are an integral part of economic development in similar ways. For example, unit costs for the provision of infrastructure of any type at a given level of performance are often a function of the type and intensity of land development [12]. Infrastructure technologies share a common problem that abrupt changes can occur as their scale increases and these can influence cost in uneven ways (for example, the transition from septic tanks to wastewater treatment plants; wells to community water supplies). Different infrastructures also have in common “publicness”, that is, they provide public services even though these services might be under private ownership. Finally, different infrastructure sectors often have similar management frameworks, problems, and even metrics in order to achieve reliability though the details may differ.

Another important dimension is that many infrastructures are interrelated both spatially and functionally; e.g., they may be located in the same area.

Since the latter part of the twentieth century, the notion of infrastructure interdependencies, which are fundamental to achieving reliability, security, and safety, has underscored the unity of infrastructure, both functionally and geographically.

Finally, a common framework for all types of infrastructures is needed to incorporate infrastructure characteristics and risks for priority setting and resource allocation. With different metrics for reliability and performance in different infrastructure sectors, this becomes nearly impossible to accomplish. The US Department of Homeland Security (DHS) has underscored the need to use risk-based methods to allocate resources for security, especially with respect to infrastructure [13].

Objectives and Concepts Associated with Reliability

Infrastructure reliability, as a performance concept, has evolved distinctly differently within the different disciplines. Engineers are concerned with the operation of the system. Economists emphasize the efficiency and effectiveness of these operations. Social scientists have underscored public service as a bottom line. Objectives for infrastructure reliability have reflected this wide range of interests, and typically include security, safety, health [14], convenience, comfort, aesthetics, environmental soundness [15], social soundness (including equity, justice, and fairness [16]), and economic and fiscal viability. These objectives often overlap and conflict with one another. They are prioritized differently depending on individual judgments.

Safety and security are considered as a starting point for reliability, since they pertain to basic survival. If the infrastructure reliability objectives listed above have any order at all, safety and security would probably precede and will be a necessary prerequisite for the others. Safety is defined as protection from the threat of death, injury, and loss of property. Security is more subtle than safety, but related to it. Security pertains to the assurance of being protected against catastrophic events such as those posed by natural disasters and terrorism (*see Counterterrorism*). Some concepts fall under both safety and security, such as crime, depending on what the context is. These two concepts will be the focus of the discussion of reliability, since they represent its key features.

Safety

In the decade before *America in Ruins* [1] popularized the condition of US infrastructure and how its reliability was being undermined, many major structures were experiencing massive and often spectacular failures threatening safety that dramatically changed the procedures and processes for measuring and managing infrastructure reliability. The Teton Dam collapse in 1974 initiated the National Dam Inspection Program. Transportation accidents, such as the aviation accidents associated with DC-10s [17, p. 139] provided the foundation for the National Transportation Safety Board (NTSB) as an entity independent of the United States Department of Transportation (DOT) in 1974 (though it had been established in 1967). By the end of 2004, the NTSB had “investigated more than 120 000 aviation accidents and over 10 000 surface transportation accidents – becoming the world’s premier transportation accident investigation agency.” [18, p. vi]. A bridge collapse in 1967 – the Silver Bridge between West Virginia and Ohio killing 46 people [19] led to the establishment of the National Bridge Inspection Program. Several other massive collapses such as the Schoharie Bridge in New York state on April 5, 1987 injuring 10 people fatally [20] and the Mianus Bridge in Connecticut that resulted in the death of three people led to major changes in the way bridges are designed and managed to ensure safety [21]. New, rapidly emerging, and widespread use of infrastructure technologies led to a safety focus. In the area of electric power, safety concerns have arisen in connection with both the availability of electricity and the proximity of facilities to people. Deaths, especially among sensitive populations, have been attributed to blackouts. Deaths have also occurred near distribution systems where stray voltages have caused electrocution.

Security

The management of infrastructure for security was distinct from that of safety. Though safety initially arose as a key focus of reliability, by the early to mid-1990s major terrorist attacks on infrastructure and buildings occurred worldwide, as well as some of the most costly damages to infrastructure and buildings from natural hazards. The 1993 attack on the World Trade Center through its underground parking

facility, the September 11, 2001 attacks *via* aircraft on the World Trade Center that disabled transit and roadways in New York City and the Pentagon in Washington, DC, and a string of terrorist attacks on transit outside of the United States continuing through the twenty-first century all moved infrastructure reliability into a new dimension, that of security. Security meant managing infrastructures with a greater degree of redundancy and flexibility than had been the case before the attacks [22].

As a consequence of the emphasis on security, a subset of infrastructure, “critical infrastructure”, emerged. A number of federal laws, strategies, plans, and programs initiated this effort beginning in the mid-1990s. These arose after a series of rail attacks but well before the September 11, 2001 attacks in New York City and Washington, DC.

1996 Executive Order 13010

1997 President’s Commission on Critical Infrastructure Protection

1997 US Department of Commerce Critical Infrastructure Assurance Office

1998 Presidential Decision Directive (PDD) 63

2001 USA Patriot Act Section 1016

2002 National Strategy for Homeland Security

2003 Homeland Security Presidential Directive (HSPD) 7 and 8

2003 US DHS National Strategy for the Physical Protection of Critical Infrastructures and Key Assets [23]

2004 US DHS National Incident Management System (NIMS) [24]

2005 US DHS National Infrastructure Protection Plan (NIPP) [25] (with Sector-Specific Plans issued in May 2007)

Portions of this history are drawn from Zimmerman [26, p. 527].

Selected Indicators for Security: Infrastructure Concentrations.

Patterns and trends in the location of infrastructure potentially have security implications. Over the past several decades, the extent of infrastructure facilities has grown, paralleled by a concentration of facilities in order to keep the costs down. In order to achieve economies of scale, different types of infrastructure are often colocated. Thus, when one system is disabled, others are also threatened. Examples of these concentrations and their contribution to increasing vulnerability and actual incidence of damage are noteworthy, and occur in

practically every infrastructure sector. At the state level, for example, half of the petroleum facilities are in 4 states, and half of the power plants are in 11 states; half of transit ridership is in 2 states, and half of the auto ridership is in 9 states [26, p. 531–532].

Evidence of the Growing Vulnerability of Infrastructure. In the electricity sector, the number and duration of outage incidents have been increasing in the past decade or so, as per an analysis of incidents in the United States since the 1990s [27]. This phenomenon is largely due to weather-related events. In fact, outages due to weather seem to have surpassed those due to equipment failure in the more recent years, leading one to believe that some of those more conventional equipment-related problems have been solved while some of the more extreme situations related to weather seem to be beyond the ability of infrastructure managers [27].

Infrastructure Interdependencies. Infrastructure systems have grown far more complex. This is in part due to increasing densities in urban areas and increasing demand for infrastructure to meet the needs of suburban area growth. Interdependencies can magnify condition and performance problems beyond what any single system might experience owing to a cascading of impacts [28].

Interdependencies are conceptualized at many different perspectives or scales. Input–output techniques provide interindustry impacts of changes on any one sector such as infrastructure on the rest of the economy at regional and national scales [29]. The development and application of a ratio to the duration of electric power outages relative to duration of the restoration of infrastructure that depends on that electricity in connection with the August 2003 blackout found notable delays in restoration times for infrastructures dependent on electric power [30].

Risk-Based Analytical Approaches

Computation of infrastructure risks from terrorism and natural hazards is a new and emerging field. Some methods are based on risk analysis and decision analysis, while others take alternative approaches. One review of traditional reliability and risk analysis argues that the newer risks require different analytical approaches [31]. A number of analytical tools are

being used, and these can be categorized for the purpose of discussion in terms of whether they aim to identify (a) the likelihood of an attack or alternatively (b) the consequences should such an attack occur (though some studies employ both of these areas).

Likelihood and Form of Attack

The issue of estimating likelihood or form of attack has been approached using analytical techniques, such as probabilistic and statistical forecasting models and game theory, and many involve scenario construction. Game theory, for example, has been applied to the problem of what defenders should do in the face of attackers as a guide to allocating defensive resources [32]. Models incorporating attack scenarios that reflect the form or type of an attack have been based on systems analysis and probability using risk analysis and decision trees, which are then used to identify and rank threats [33] the reference [27].

Consequence Assessment

For consequence assessment, for example, statistical models were developed and applied to historical trends in electric power failures to estimate the expected consequences of potential future outages, on duration, megawatts lost, and customers affected [27]; similarly spatial distributions and time trends were also analyzed for oil and gas pipeline failures and electric power failures, as a basis for inputs to risk management [34], and ongoing research by the same team estimates consequences of oil and gas infrastructure failure such as deaths, injuries, property loss, value of product lost, and cleanup and recovery costs applying statistical models of outage trend data.

Analyses of economic consequences of massive infrastructure disruptions have also been conducted using a variety of modeling techniques. Regional econometric models have been used to estimate the economic impacts of a loss of electric power in a single state, New Jersey [35]. Methods have been developed and applied to estimate the magnitude of business loss, death, injury, and congestion from a statistical model of outage events for individual urban areas, and combined with average dollar estimates for each of the loss categories to obtain overall economic impacts [36]. Input–output analysis has been used to estimate impacts across infrastructure and other economic categories of various infrastructure

disruptions [37]. Computable General Equilibrium (CGE) analysis has been used to estimate economic impacts of outages for a variety of infrastructure areas, for example, water systems [38]. Case-based approaches to combined risk and economic consequence analysis provide important analytical tools. A scenario-based analysis that combined probability of alternative modes of attack, severity of the attack, and its consequences in terms of economic impact was undertaken for a dirty bomb in a large port area [39].

Conclusions

The meaning of infrastructure reliability and two of its key components, safety and security, has changed over time as the goals of infrastructure systems and services have changed and as our understanding of system vulnerabilities has changed, especially in light of the relatively newer threats of natural hazards and terrorism. The integration of information technologies into infrastructure has also altered reliability concepts, adding opportunities to improve reliability as well as another layer of complexity and potential vulnerability [40]. The need for infrastructure to provide critical emergency services for both response during a catastrophe and recovery in the period following these events is an area of growing importance. In order to manage the prevention, recovery, and rebuilding after catastrophes, we need to adapt our understanding and measurement of infrastructure performance in the face of these threats.

Objectives for infrastructure reliability are still highly fragmented and differ widely not only among different infrastructure sectors, many of which are interrelated, but also within a single infrastructure. Different infrastructure sectors evaluate reliability with different levels of information quality and types of standards and protocols. This fragmentation particularly becomes an issue when managing the risks of interdependencies among infrastructures.

Measuring infrastructure reliability, safety, and security are now critical aspects of managing infrastructure, including protecting infrastructure assets by prioritizing them for resource allocation [41]. Risk-based decision making has become a distinct element of security policy and resource allocation [13, 42], yet, a considerable debate exists as to the nature of risks faced by critical infrastructures, how these

risks should be quantified and the tools available to do so, at what geographic scale analyses should be undertaken, and how results are to be used as a basis for prioritization schemes [43]. In particular, the sensitivity of the methodology to changes in assumptions and variables remains a future challenge. The basis for distributing these funds through the Homeland Security Grant Program (HSGP) in part is based on risk assessment methodologies to guide the allocation of funds by urban area as well as for specific types of infrastructure. The funds available under HSGP totaled \$1.7 billion in FY 2006, and according to a recent report by the US Government Accountability Office [44] represented a decline in funding of 14% over the previous fiscal year. Through FY 2007, the DHS has indicated that a total of \$1.5 billion was granted under the Infrastructure Protection Program with \$445 million for FY 2007 alone [45].

Irrespective of safety and security, a growing threat to the continued integrity of infrastructure underscored by persistent low quality levels [10] is declining levels of funding relative to need. In many infrastructure sectors, gap analyses conducted over the years have supported the existence of this divide. The 2008 funding cycle similarly shows that this gap is not likely to be closed in the near future, with critical cuts occurring in transit and water infrastructure [46]. This represents expenditures primarily for standard capital investment, and does not reflect patterns and trends in security funding. These two streams of investment need to be combined. Achieving reliability encompasses many criteria, one of which is security, and solutions to many of the security needs are likely to be solutions to other objectives for reliable infrastructure as well.

Acknowledgments and Disclaimer

This research was supported by a number of research projects: the US DHS through the Center for Risk and Economic Analysis of Terrorism Events (CREATE), grant numbers EMW-2004-GR-0112 and N00014-05-0630; the Institute for Information Infrastructure Protection at Dartmouth College under grant number 2003-TK-TX-0003; and “Public Infrastructure Support for Protective Emergency Services” through the Center for Catastrophe Preparedness and Response at NYU under grant number 2004-GT-TX-0001; however, any opinions, findings, and conclusions or recommendations in this document are those

of the author(s) and do not necessarily reflect views of the US DHS.

References

- [1] Choate, P. & Walter, S. (1983). *America in Ruins*, Duke University Press, Durham.
- [2] Tarr, J.A. (1984). The evolution of the urban infrastructure in the nineteenth and twentieth centuries, in *Perspectives on Urban Infrastructure*, R. Hanson, ed, National Academy Press, Washington, DC, pp. 4–66.
- [3] Lynch, K. (1962). *The Image of the City*, MIT Press, Cambridge.
- [4] Perry, D.C. (1995). Building the public city: an introduction, in *Building the Public City*, D.C. Perry, ed, Sage, Thousand Oaks, pp. 1–20.
- [5] National Research Council (1995). *Measuring and Improving Infrastructure Performance*, National Academy Press, Washington, DC.
- [6] Cambridge Systematics (2006). *Unclogging America's Arteries: Effective Relief for Highway Bottlenecks 1999–2004*, American Highway Users Alliance, Washington, DC, <http://www.highways.org/pdfs/bottleneck2004.pdf>.
- [7] U.S. Department of Transportation (2004). *Status of the Nation's Highways, Bridges and Transit: Conditions and Performance*, Washington, DC, <http://www.fhwa.dot.gov/policy/2004cpr/pdfs/cp2006.pdf>.
- [8] Hendren, P. & Niemeier, D.A. (2006). Evaluating the effectiveness of state departments of transportation investment decisions: linking performance measures to resource allocation, *Journal of Infrastructure Systems* **12**, 216–229.
- [9] US-Canada Power System Outage Task Force (2004). *Final Report on the August 14th 2003 Blackout in the United States and Canada: Causes and Recommendations*.
- [10] American Society of Civil Engineers (ASCE) (2005). *2005 Report Card for America's Infrastructure*, Washington, DC, <http://www.asce.org/reportcard>.
- [11] Rendell, E.G. (1998). A call to pay the U.S. infrastructure price tag, *Public Works Management & Policy* **3**, 99–103.
- [12] Speir, C. & Stephenson, K. (2002). Does sprawl cost us all? *APA Journal* **63**, 56–70.
- [13] U.S. Department of Homeland Security (2005). *Discussion of the FY 2006 Risk Methodology and the Urban Areas Security Initiative*, Washington, DC.
- [14] Zimmerman, R. (2005). Mass transit infrastructure and urban health, *Journal of Urban Health* **82**, 21–32.
- [15] Beatley, T. (2000). *Green Urbanism*, Island Press, Washington, DC.
- [16] Litman, T. (2005). *Lessons from Katrina and Rita: What Major Disasters can Teach Transportation Planners*, Victoria Transport Policy Institute, Victoria, <http://www.vtpi.org/katrina.pdf>.
- [17] Perrow, C. (1999). *Normal Accidents*, 2nd Edition, Princeton University Press, Princeton.
- [18] National Transportation Safety Board (NTSB) (2005). *We Are All Safer: Lessons Learned and Lives Saved 1975-2005*, 3rd Edition, Safety Report NTSB/SR-05/01. Washington, DC, <http://www.nts.gov/publicn/2005/SR0501.pdf>.
- [19] NTSB (1970). *Collapse of U.S. 35 Highway Bridge, Point Pleasant, West Virginia, December 15, 1967*, Highway Accident Report, Washington, DC, at <http://www.nts.gov/publicn/1971/HAR7101.htm>.
- [20] NTSB (1988). *Collapse of New York Thruway (I-90) Bridge over the Schoharie Creek, Near Amsterdam, New York, April 5, 1987*, Highway Accident Report, Washington, DC, at <http://www.nts.gov/publicn/1988/HAR8802.htm>.
- [21] Zimmerman, R. (1999). Planning and administration: frameworks and case studies [integrating risk management and natural hazard management], in *Natural Disaster Management*, J. Ingleton, ed, Tudor Rose, Leicester, pp. 225–227.
- [22] Zimmerman, R. (2003). Public infrastructure service flexibility for response and recovery in the September 11th, 2001 attacks at the World Trade Center, in *Beyond September 11th: An Account of Post-Disaster Research*, Natural Hazards Research & Applications Information Center, Public Entity Risk Institute, and Institute for Civil Infrastructure Systems, University of Colorado, Boulder, pp. 241–268.
- [23] U.S. Department of Homeland Security (2003). *National Strategy for the Physical Protection of Critical Infrastructures and Key Assets*, at http://www.dhs.gov/xlibrary/assets/Physical_Strategy.pdf.
- [24] U.S. Department of Homeland Security (2004). *National Incident Management System (NIMS)*, at http://www.fema.gov/pdf/emergency/nims/nims_doc_full.pdf.
- [25] U.S. Department of Homeland Security (2005). *National Infrastructure Protection Plan (NIPP)*, at http://www.dhs.gov/xlibrary/assets/NIPP_Plan.pdf.
- [26] Zimmerman, R. (2006). Critical infrastructure and interdependency, in *The McGraw-Hill Homeland Security Handbook*, D.G. Kamien, ed, The McGraw-Hill Companies, New York, Chapter 35, pp. 523–545.
- [27] Simonoff, J.S., Restrepo, C.E. & Zimmerman, R. (2007). Risk management and risk analysis-based decision tools for attacks on electric power, *Risk Analysis* **27**(3), 547–570.
- [28] Rinaldi, S.M., Peerenboom, J.P. & Kelly, T.K. (2001). Identifying, understanding and analyzing critical infrastructure interdependencies, *IEEE Control Systems Magazine* **21**, 11–25.
- [29] Crowther, K.G. & Haimes, Y.Y. (2005). Application of the inoperability input-output model (IIM) for risk assessment and management of interdependent infrastructures, *Systems Engineering* **8**, 323–341.
- [30] Zimmerman, R. & Restrepo, C.E. (2006). The next step: quantifying infrastructure interdependencies to improve security, *International Journal of Critical Infrastructures* **2**, 215–230.

- [31] Bier, V. (2006). Game-theoretic and reliability methods in counterterrorism and security, in *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, A.G. Wilson, G.D. Wilson & D.H. Olwell, eds, Springer, New York, pp. 23–40.
- [32] Bier, V., Oliveros, S. & Samuelson, L. (2007). Choosing what to protect: strategic defensive allocation against an unknown attacker, *Journal of Public Economic Theory* 9(4), pp. 563–587.
- [33] Pate-Cornell, E. & Guikema, S. (2002). Probabilistic modeling of terrorist threats: a systems analysis approach to setting priorities among countermeasures, *Military Operations Research* 7, 5–20.
- [34] Simonoff, J.S., Restrepo, C.E., Zimmerman, R. & Naph-tali, Z.S. (2007). Spatial distribution of electricity and oil and gas pipeline failures in the United States: a state level analysis, in *Critical Infrastructure Protection: Issues and Solutions*, E.D. Goetz & S. Sheno, eds, Springer, New York.
- [35] Greenberg, M., Mantell, N., Lahr, M., Felder, F. & Zimmerman, R. (2007). Short and intermediate economic impacts of a terrorist-initiated loss of electric power: case study of New Jersey, *Energy Policy* 35, 722–733.
- [36] Zimmerman, R., Restrepo, C.E., Simonoff, J.S. & Lave, L. (2007). Risks and economic costs of a terrorist attack on the electricity system, in *The Economic Costs and Consequences of Terrorism*, H.W. Richardson, P. Gordon & J.E. Moore II, eds, Edward Elgar Publishers, Cheltenham, pp. 273–290.
- [37] Santos, J.R. & Haimes, Y.Y. (2004). Modeling the demand reduction input-output inoperability due to terrorism of interconnected infrastructures, *Risk Analysis* 24, 1437–1451.
- [38] Rose, A. & Liao, S. (2005). Modeling resilience to disasters: computable general equilibrium analysis of a water service disruption, *Journal of Regional Science* 45, 75–112.
- [39] Rosoff, H. & von Winterfeldt, D. (2005). A risk and economic analysis of dirty bomb attacks on the ports of Los Angeles and Long Beach, Report #05-027, Center for Risk and Economic Analysis of Terrorism Events, University of Southern California, Los Angeles.
- [40] Zimmerman, R. & Horan, T. (eds) (2004). *Digital Infrastructures: Enabling Civil and Environmental Systems through Information Technology*, Routledge, London.
- [41] National Research Council (2002). *Making the Nation Safer*, National Academy Press, Washington, DC.
- [42] National Infrastructure Advisory Council (NIAC) (2005). *Risk Management Approaches to Protection*, Final Report and Recommendations by the Council, Washington, DC, at http://www.dhs.gov/xlibrary/assets/niac/NIAC_RMWG_-_2-13-06v9_FINAL.pdf.
- [43] U.S. Government Accountability Office (2005). *Risk Management. Further Refinements Needed to Assess Risks and Prioritize Protective Measures at Ports and Other Infrastructure*, Washington, DC.
- [44] U.S. Government Accountability Office (2006). *Homeland Security Grants: Observations on Process DHS Used to Allocate Funds to Selected Urban Areas (GAO-07-381R)*, Washington, DC, at <http://www.gao.gov/new.items/d07381r.pdf>.
- [45] U.S. DHS. (2007). *DHS Announces \$445 Million to Secure Critical Infrastructure*, Washington, DC.
- [46] American Planning Association (APA). (2007). *APA Advocate*, Chicago, p. 1.

Related Articles

Protection of Infrastructure

Reliability of Large Systems

Repairable Systems Reliability

RAE ZIMMERMAN

Managing Risks of Consumer Products

Each year consumer products are involved in tens of millions of injuries and tens of thousands of fatalities in the United States [1]. Primary responsibility for risk management rests with the companies that create and manufacture products. Most developed countries have also established government agencies that provide regulatory oversight for consumer products. In the United States, the Consumer Product Safety Commission (CPSC) maintains regulatory jurisdiction over more than 15 000 types of consumer products, including toys, cribs, power tools, cigarette lighters, and household chemicals. The National Highway Traffic Safety Administration provides regulatory oversight for motor vehicles, and the US Food and Drug Administration regulates safety of food, drugs, medical devices, biologics, cosmetics, and radiation-emitting products.

Proper consideration of public safety can improve product design and development, management of product quality during manufacturing, and monitoring of field performance during use by consumers. Formal tools of quantitative risk assessment add value by improving decision making at each of these stages.

Product Design and Development

Product characteristics and statistical analysis of historical data on product-related injuries and deaths can inform safety assessment of a product design before it is manufactured and distributed. Alternative designs can be developed and tested to reduce the risk of particular injury mechanisms. Examples include the design of children's toys and consumer electronics to reduce choking hazards and electric shock risks, respectively.

The most widely used source of US data on product-related injuries is the National Electronic Injury Surveillance System (NEISS) [2] operated by CPSC. NEISS is a national probability sample of hospitals from which patient information is collected for every emergency visit involving an injury associated with consumer products. From this sample, which includes more than 375 000 cases per year, the total

number of product-related injuries treated in hospital emergency rooms nationwide can be estimated.

Each NEISS case provides information on the date of injury, type of product, gender and age of the injured party, diagnosis (e.g., laceration and fracture), body part(s) involved, disposition (e.g., admitted, treated, and released), and locale (e.g., home, work, and school) where the injury occurred. A short narrative describes the incident and associated injuries. Manufacturers can use data sources such as NEISS to understand the types and frequencies of injuries that occur in association with the use of a consumer product.

Estimating not only the frequency but also the risk of injury requires some measure of the opportunity for injury (i.e., the exposure). For example, risk assessments of all-terrain vehicles use data on riding hours from both industry- and government-sponsored surveys of owners [3]. Comparative risk analyses (*see Comparative Risk Assessment*) of other recreational products may use information from the National Sporting Goods Association [4] on levels of participation in related activities. For studies of burn risk from home appliances, the Energy Information Administration's Residential Energy Consumption Survey provides data on the number of US households having these appliances [5].

Risk Management in Product Manufacturing

Unintended declines in the production quality of consumer items can pose safety hazards to consumers, increase failure rates, and shorten product life, which, in turn, causes unexpected financial burdens and potential legal liabilities to manufacturers. At the beginning of the production process, companies may implement acceptance sampling plans to ensure the quality of raw materials or components produced internally or by suppliers [6, 7]. The acceptable quality limits (AQLs) used in such plans can be chosen on the basis of a quantitative risk assessment. Standard tools for statistical process control – such as N_p , $\bar{X}$ -bar, and R charts – are routinely deployed by manufacturers to provide checks on quality throughout their production processes [8, 9]. A motivating principle is that successful business operations require continuous effort to reduce variation in process outputs. End-of-line

testing is typically performed before manufacturers approve products for shipping to customers. In some industries (e.g., pharmaceuticals), such testing may be required by government regulation [10].

Monitoring of Field Performance

Information about the performance of a product after release to consumers comes from field returns, warranty programs (*see Warranty Analysis*), and customer complaints. Records of incidents involving failure of a product during operation, with or without associated injuries to consumers, typically include information such as model and serial number of the unit, production date or date code, incident date, and sales location. In monitoring field performance, a manufacturer needs to determine whether an adverse change has occurred and, if so, assess its implications for product reliability and safety. If a background level of event risk is accepted, statistical methods can determine whether adverse events can be explained by random variation or require attention [11]. Production, shipping, and sales records typically provide the raw data for such analyses.

Confirmation of a problem raises a series of questions: How bad is it? Is it getting worse? Do field data point to a particular production period or facility? A quantitative risk assessment can address these questions using statistical estimation and hypothesis testing. Risk analysts review the available information about the production history and seek changes in time period, plant, process, equipment, design, or supplier that are associated with subsequent problems. Parametric statistical models are frequently used to fit time-to-failure distributions in engineering applications (*see Lifetime Models and Risk Assessment*).

Strategies for Risk Mitigation

When an unacceptable safety hazard is found to exist by manufacturers and/or regulators, the finding prompts corrective action. Product recalls require identification of the units affected, typically by time or source of production, and determination of whether affected units will be repaired or replaced. Increasing globalization of the world economy challenges manufacturers and regulators of consumer products, as a product may contain components from different nations and be sold to markets all over the world.

The process for implementing corrective action varies from country to country. In the United States, CPSC adopted in 1997 a “Fast Track Product Recall Program” [12] for reports filed according to Section 15(b) of the Consumer Product Safety Act (CPSA). This program requires manufacturers, distributors, and retailers of consumer products to notify the commission of certain defects, unreasonable risks, or noncompliance with voluntary or mandatory standards. Under the CPSC Fast Track program, the staff refrains from making a preliminary determination when firms report and, within 20 working days, implement an acceptable corrective action plan. The plan submitted to CPSC must describe the recall action (refund, repair, or replace) that the company will take to eliminate the identified risk, and provide sufficient information on the product design, incident, and testing to allow the CPSC staff to determine whether the proposed action can correct the identified problem.

Tools for Risk Communication

An effective risk assessment must link measured levels of risk to specific corrective actions. Supporting tools have been developed, particularly in the European Union [13].

A community rapid information exchange system known as RAPEX assesses the risk of potentially hazardous consumer products by first considering (a) the probability of health damage from regular exposure, as characterized in Table 1, and (b) the severity of injury from the product, as characterized in Table 2. Depending on the probability and severity of damage, as illustrated in Figure 1, the RAPEX method then classifies the overall gravity of an adverse outcome on a five-point ordinal scale ranging from very low to very high. The final judgment of whether the risk requires corrective action considers three additional factors: (a) the vulnerability of people exposed to the product, particularly children and the elderly, (b) whether the hazard is obvious to nonvulnerable adults, and (c) whether the product has adequate warnings and safeguards. The RAPEX methodology is sometimes viewed as the preferred approach in countries adopting a risk-averse approach to consumer product safety based on such rough scoring [13].

An alternative tool, the Belgian risk matrix, qualitatively rates the level of exposure to the hazardous

Table 1 Risk classification for RAPEX method

Probability of health/safety damage from regular exposure to hazardous product	Probability of hazardous product		
	1%	10%	100%
Hazard is always present and health/safety damage is likely to occur in foreseeable use	Medium	High	Very high
Hazard may occur under one improbable or two possible conditions	Low	Medium	High
Hazard only occurs if several improbable conditions are met	Very low	Low	Medium

Table 2 Severity classification for RAPEX method

Slight	Serious	Very serious
<2% incapacity, usually reversible and not requiring hospital treatment Minor cuts, minor fractures, minor burns, sprains	2–15% incapacity, usually irreversible requiring hospital treatment Serious cuts, loss of a finger or toe, damage to sight, damage to hearing, moderate burns	>15% incapacity, usually irreversible Death, loss of limbs, loss of sight, loss of hearing, serious burns (>25%)

product, the probability of hazard occurrence, and the seriousness of the consequences [13]. Numerical scores are assigned to each level of these factors. The product of the three factors yields a quantitative

risk score, a corresponding characterization of risk on a five-point ordinal scale ranging from slight to very high, and identification of the action required to mitigate the assessed level of risk.

Risk assessment of consumer products for the GPSD

This procedure is proposed to assist companies when deciding whether a specific hazardous situation caused by a consumer product requires notification to the authorities

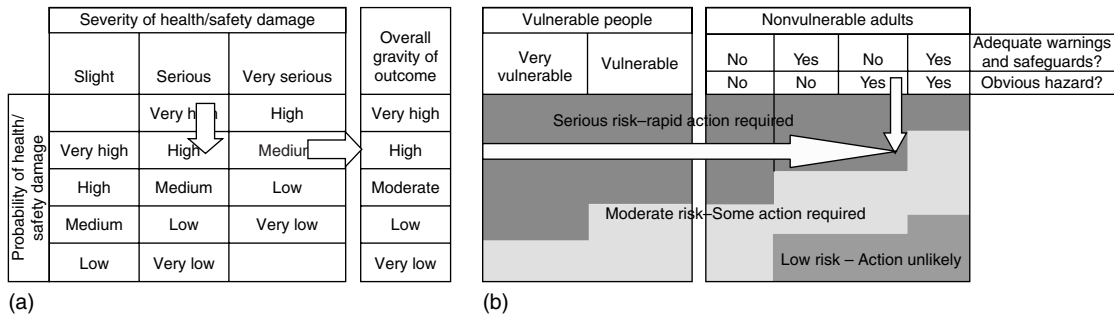


Figure 1(a) is used to determine the gravity of the outcome of a hazard, depending on the severity and probability of the possible health/safety damage (see the text in Example).

Figure 1(b) is used to determine the rating of the gravity of risk depending on the type of user and, for nonvulnerable adults, whether the product has adequate warnings and safeguards and whether the hazard is sufficiently obvious.

Example (indicated by the arrows above)

A chain saw user has suffered a badly cut hand and it is found that the chain saw has an inadequately designed guard that allowed the user's hand to slip forward and touch the chain. The company's assessor makes the following risk assessment.

Figure 1(a)–The assessment of probability is high because the hazard is present on all products and may occur under certain conditions. The assessment of severity is serious so the overall gravity rating is high.

Figure 1(b)–The chain saw is for use by nonvulnerable adults, and presents an obvious hazard but with inadequate guards.

The high gravity is therefore intolerable, so a serious risk exists.

Figure 1 Outline of RAPEX risk assessment methodology: (a) risk estimation and (b) grading of risk

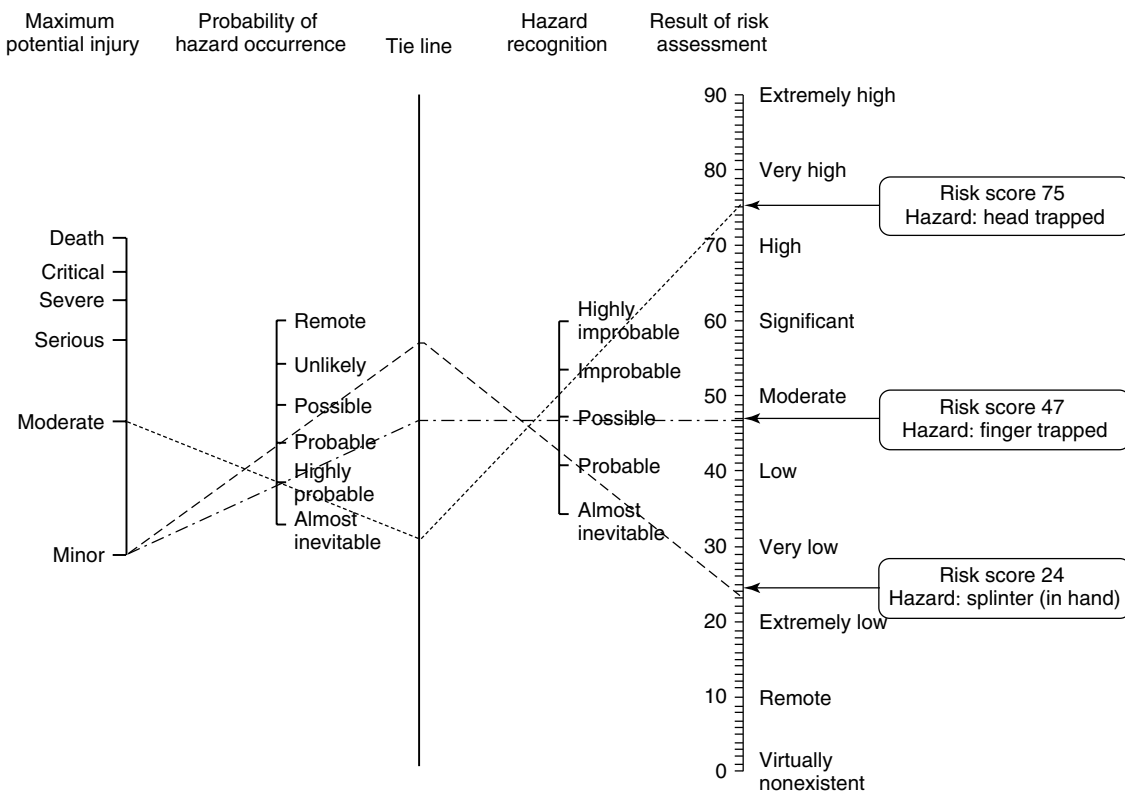


Figure 2 Slovenian nomograph, sample risk assessment of cots

The Slovenian nomograph [13] considers four factors: the availability of the product on the market, the capability of an average adult to recognize product defects and potential misuse, the probability of hazard occurrence, and the maximum potential injury resulting from the hazard. Construction of the nomograph involves developing scales for each factor and placing the scales in positions relative to one another that correspond to a particular weighting of the factors. Figure 2 illustrates the placement of the scales and use of the nomograph to assess various risks posed to consumers from cots. As shown, the risk assessor draws a line through the maximum injury and occurrence probability scales, as well as a tie line, at which point the risk assessor draws an intersecting line through the hazard recognition scale. This line can be extended to intersect a number line from which the assessor obtains a risk score. The inclusion of product availability is optional and would provide an estimate of the societal risk associated with the product.

Summary

Although consumer products are required to be safe, safety does not mean zero risk. A safe product is one that has reasonable risks, given the magnitude of the benefit expected and the alternatives available. In managing consumer product risk, quantitative risk assessment and associated statistical methods are used to frame substantive issues in terms of estimable quantities and testable hypotheses, to extract information on product performance from data on field incidents and manufacturing records, and to communicate findings to upper management, regulatory authorities, and consumers.

References

- [1] United States Consumer Product Safety Commission (2007). *2007 Performance Budget (Operating Plan)*, Washington, DC.

-
- [2] United States Consumer Product Safety Commission (2000). *NEISS, The National Electronic Injury Surveillance System: A Tool for Researchers*, Division of Hazard and Injury Data Systems, Washington, DC.
- [3] United States Consumer Product Safety Commission (2003). *The All-Terrain Vehicle 2001 Injury and Exposure Studies*, Washington, DC.
- [4] National Sporting Goods Association (2001). *Sports Participation Survey*, Mt. Prospect.
- [5] Energy Information Administration (2001). *Residential Energy Consumption Survey, Office of Energy Markets and End Use*, U.S. Department of Energy, Washington, DC.
- [6] American Society for Quality (2003). *Sampling Procedures and Tables for Inspection by Attributes*, ANSI/ASQ Z1.4–2003, Quality Press, Milwaukee.
- [7] American Society for Quality (2003). *Sampling Procedures and Tables for Inspection by Variables for Percent Nonconforming*, ANSI/ASQ Z1.9–2003, Quality Press, Milwaukee.
- [8] Ryan, T.P. (2000). *Statistical Methods for Quality Improvement*, John Wiley & Sons, New York.
- [9] Tennant, G. (2001). *Six Sigma: SPC and TQM in Manufacturing and Services*, Gower, Hampshire.
- [10] United States Pharmacopeial Convention (2007). *United States Pharmacopeia: National Formulary*, Rockville.
- [11] Zhao, K. & Steffey, D. (2007). To recall or not to recall? Statistical assessment of consumer product failure risk, 2007, *Proceedings of the American Statistical Association, Risk Analysis Section CD-ROM*, American Statistical Association, Alexandria.
- [12] United States Consumer Product Safety Commission (1997). Conditions under which the staff will refrain from making preliminary hazard determinations, *Federal Register* **62**(142), 39827–39828.
- [13] Floyd, P., Nwaogu, T., Salado, R. & George, C. (2006). *Establishing a Comparative Inventory of Approaches and Methods Used by Enforcement Authorities for the Assessment of the Safety of Consumer Products Covered by Directive 2001/95/EC on General Product Safety and Identification of Best Practices*, Risk & Policy Analysts Limited, Norfolk.

Related Articles

Competing Risks in Reliability

Reliability of Consumer Goods with “Fast Turn Around”

Repair, Inspection, and Replacement Models

Wear

MADHU IYER, KE ZHAO AND DUANE L.
STEFFEY

Markov Modeling in Reliability

Many life-critical systems are required to operate without a system failure for a given period of time. Examples are nuclear, aerospace, spacecraft, and other such systems. Redundancy (*see Reliability Growth Testing; Reliability Optimization*) is often used to achieve high reliability. Such systems are fault tolerant since they can continue to function even when one or more of the components have failed. The term *reliability* (*see Mathematics of Risk and Reliability: A Select History*) is generally used to express certain degree of assurance that a system will operate successfully in a specified environment, during a certain time period. Reliability is the probability that a system performs a specified service throughout a specified interval of time.

Reliability of a system is related to its lifetime. Let T be the random variable representing the length of the lifetime of a system. Then, the reliability for a system is defined as follows:

Definition Reliability of a system, denoted by $R(t)$, is the probability that the system functions until time t . Then

$$R(t) = P(T > t) = 1 - F(t), \quad t \geq 0 \quad (1)$$

where $F(t)$ is the cumulative distribution function of the random variable T . $F(t)$ gives the probability that the lifetime does not exceed t , i.e.,

$$F(t) = P(T \leq t) = 1 - R(t) \quad (2)$$

Therefore, $F(t)$ is often called unreliability of the system.

Another important function related to life distribution is the instantaneous failure rate or hazard function $h(t)$ (*see Default Risk; Mathematical Models of Credit Risk; Reliability Data; Lifetime Models and Risk Assessment*). This is the instantaneous failure rate of a component that has survived t units of time, i.e.,

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \frac{F(t + \Delta t) - F(t)}{P(T > t)} = \frac{f(t)}{R(t)} \quad (3)$$

where $f(t)$ is the probability density function of T . The product $h(t)\Delta t$ represents the conditional

probability that a component having survived to age t will fail in the interval $(t, t + \Delta t)$.

Finally, we introduce the concept of mean time to failure (MTTF) (*see Imprecise Reliability*). MTTF is the average length of time until failures and is given as

$$MTTF = E[T] = \int_0^\infty t f(t) dt = \int_0^\infty R(t) dt \quad (4)$$

Reliability modeling is a useful technique to abstract the dynamic behavior of computer, communications, aerospace, nuclear and similar systems to predict their reliability. To perform the reliability modeling, we must be familiar with the reliability characteristics of the system and its components, the failure modes and their effect on the system. A reliability model is constructed to represent the reliability characteristics, and predict the system reliability.

Reliability analysis of the systems is typically performed using one of the two methods, combinatorial modeling and Markov modeling (*see Credit Migration Matrices; Imprecise Reliability*). Combinatorial modeling methods such as fault trees (*see Expert Judgment; Latin Hypercube Sampling; Canonical Modeling*), reliability graphs, and reliability block diagrams (*see Systems Reliability*) are applicable to cases where the system can be broken down into series and parallel combination of its components. These model types capture conditions that make a system fail in terms of the structural relationships between the system components. The main limitation of the combinatorial methods is that they cannot be used to model system repairs or dynamic reconfigurations of the system. For more complex systems that cannot be represented as series and parallel combination of events, and also exhibit complicated interactions between the components, the technique of Markov modeling is employed. Various dependencies between the components are captured by Markov models. For the systems with large number of components, the number of states can grow prohibitively large. This results in state space explosion and hence the large Markov models. Hierarchical model (*see Bayesian Statistics in Quantitative Risk Assessment; Geographic Disease Risk; Meta-Analysis in Nonclinical Risk Assessment*) composition is applied to avoid large Markov models [1, 2].

For evaluating the reliability of the complex systems, formulating an analytic model can be quite

cumbersome. Also, in practice, isolating the effect of various parameters on a system, while holding the others as constants, requires exploring a variety of scenarios. It is economically infeasible to build several such systems. Simulation offers an attractive mechanism for reliability evaluation and the study of the influence of various parameters on the failure behavior of the system. In [3], several algorithms are developed for computing the reliability of fault tolerant systems. However, in this article, we emphasize on Markov modeling in reliability for the complex systems.

Markov Modeling

Markov models consider the system states and the possible transitions between these states. The states in a fault tolerant system are distinguished as operational or failed, depending on the available system functionality. The basic assumption in Markov model is that the system is memoryless, that is, the transition probabilities are determined only from the present state and not by the history. For reliability analysis with Markov models, these conditions are specified by the assumption of exponential distributions for failure and repair times. A *state transition diagram* is a directed graph representation of system states, transition between states, and transition rates. These diagrams contain sufficient information for developing the state equations. The effect of changes in the model parameters can be easily investigated. For detailed discussion on Markov processes, readers are referred to [4, 5]. The following examples show how Markov model can be used to obtain the measure of interest such as reliability and MTTF for the parallel redundant system.

Example of Parallel Redundant System

We consider Markov reliability model of a parallel redundant system with two components [4]. Assume

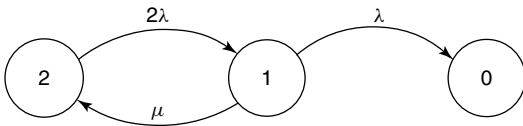


Figure 1 State transition diagram for the parallel redundant system

that the time to failure of a component is exponentially distributed with failure rate λ . The system has a single repair facility. Time to repair of a component is exponentially distributed with parameter μ . The system fails when both the components fail and no recovery is possible. Let the state of the system be defined as the number of components in working condition. The state space is $\{0, 1, 2\}$. State “0” is an absorbing state since the system does not move to other states once it is in state “0”. The state transition diagram is given in Figure 1.

Let $P_j(t)$ be the probability that system is in state j at time t . From the state transition diagram, the system of differential equations can be written as

$$\frac{d}{dt}P_2(t) = -2\lambda P_2(t) + \mu P_1(t) \quad (5)$$

$$\frac{d}{dt}P_1(t) = 2\lambda P_2(t) - (\lambda + \mu)P_1(t) \quad (6)$$

$$\frac{d}{dt}P_0(t) = \lambda P_1(t) \quad (7)$$

Suppose the system is in state 2 at time $t = 0$ i.e., $P_2(0) = 1$; $P_1(0) = P_0(0) = 0$. The above systems of equations can be solved using the technique of Laplace transform. Hence, taking Laplace transform on both sides of the above equations and using the initial condition, we get

$$s\hat{P}_2(s) - 1 = -2\lambda\hat{P}_2(s) + \mu\hat{P}_1(s) \quad (8)$$

$$s\hat{P}_1(s) = 2\lambda\hat{P}_2(s) - (\lambda + \mu)\hat{P}_1(s) \quad (9)$$

$$s\hat{P}_0(s) = \lambda\hat{P}_1(s) \quad (10)$$

On solving the above equations for $P_0(s)$, we get

$$\hat{P}_0(s) = \frac{2\lambda^2}{s[s^2 + (3\lambda + \mu)s + 2\lambda^2]} \quad (11)$$

or

$$\hat{P}_0(s) = \frac{2\lambda^2}{s(s + s_1)(s + s_2)} \quad (12)$$

where

$$s_1, s_2 = \frac{(3\lambda + \mu) \pm \sqrt{\lambda^2 + 6\lambda\mu + \mu^2}}{2} \quad (13)$$

Let the random variable T denote the time to failure of the system. Then $P_0(t)$ gives the probability that the system has failed at or before time t . Therefore,

reliability of the system (that is, probability of nonfailure) is given as

$$R(t) = 1 - P_0(t) \quad (14)$$

The cumulative distribution function (cdf) $F(t)$ of T is given as

$$F(t) = P(T \leq t) = 1 - R(t) \quad (15)$$

Differentiating equation (15), the probability density function of T , denoted by $f(t)$, is given by

$$f(t) = \frac{d}{dt}(-R(t)) = \frac{d}{dt}P_0(t) \quad (16)$$

Taking Laplace transform on both sides of equation (16), we get

$$\hat{f}(s) = s\hat{P}_0(s) = \frac{2\lambda^2}{(s + s_1)(s + s_2)} \quad (17)$$

Taking inverse Laplace transform (see **Ruin Probabilities: Computational Aspects**), we get

$$f(t) = \frac{2\lambda^2}{s_1 - s_2} (e^{-s_2 t} - e^{-s_1 t}) \quad (18)$$

Therefore, reliability $R(t)$ of the system is given by

$$R(t) = \int_t^\infty f(t) dt = \frac{2\lambda^2}{s_1 - s_2} \left[\frac{e^{-s_2 t}}{s_2} - \frac{e^{-s_1 t}}{s_1} \right] \quad (19)$$

Using equation (4), MTTF of the system is obtained as:

$$\begin{aligned} E[T] &= \int_0^\infty (1 - F(t)) dt = \int_0^\infty R(t) dt \\ &= \frac{2\lambda^2(s_1 + s_2)}{s_1^2 s_2^2} \end{aligned} \quad (20)$$

On substituting the values of s_1 and s_2 , we get

$$MTTF = \frac{3}{2\lambda} + \frac{\mu}{2\lambda^2} \quad (21)$$

Example of Parallel Redundant System with Imperfect Coverage

This is a modified form of the example discussed in the section titled "Example of Parallel Redundant System". In this case, it is assumed that not

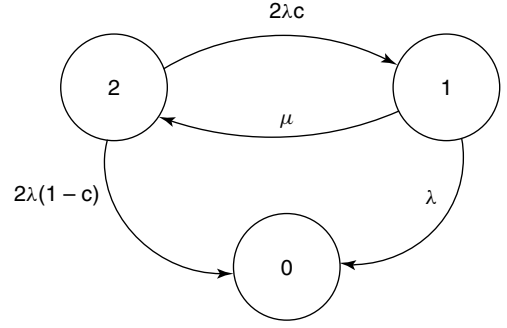


Figure 2 State transition diagram of the system with imperfect coverage

all faults are recoverable and there is a coverage factor "c", denoting the conditional probability that the system recovers, given that a fault has occurred. The state transition diagram of this example is shown in Figure 2. For this state transition diagram, the system of differential equations are written as

$$\begin{aligned} \frac{d}{dt}P_2(t) &= -2\lambda c P_2(t) \\ &\quad - 2\lambda(1-c)P_2(t) + \mu P_1(t) \end{aligned} \quad (22)$$

$$\frac{d}{dt}P_1(t) = -(\lambda + \mu)P_1(t) + 2\lambda c P_2(t) \quad (23)$$

$$\frac{d}{dt}P_0(t) = \lambda P_1(t) + 2\lambda(1-c)P_2(t) \quad (24)$$

Assume that, initially the system is in state 2, so that $P_2(0) = 1$, $P_0(0) = P_1(0) = 0$. Taking Laplace transform on both sides of above equations and applying the initial condition, we get

$$\begin{aligned} s\hat{P}_2(s) - 1 &= -2\lambda c \hat{P}_2(s) \\ &\quad - 2\lambda(1-c)P_2(s) + \mu \hat{P}_1(s) \end{aligned} \quad (25)$$

$$s\hat{P}_1(s) = 2\lambda c \hat{P}_2(s) - (\lambda + \mu) \hat{P}_1(s) \quad (26)$$

$$s\hat{P}_0(s) = \lambda \hat{P}_1(s) + 2\lambda(1-c) \hat{P}_2(s) \quad (27)$$

On solving the above system of equations for $\hat{P}_0(s)$, we get

$$\hat{P}_0(s) = \frac{2\lambda}{s} \left[\frac{s + \lambda + \mu - c(s + \mu)}{s(s + 2\lambda)(s + \lambda + \mu) - 2\lambda\mu c} \right] \quad (28)$$

In this system, the cdf of T , the time to system failure, is given by

$$F(t) = P_0(t) \quad (29)$$

Taking Laplace transform and using the initial condition, we get

$$\hat{f}(s) = s\hat{P}_0(s) = 2\lambda \left[\frac{s + \lambda + \mu - c(s + \mu)}{s(s + 2\lambda)(s + \lambda + \mu) - 2\lambda\mu c} \right] \quad (30)$$

Without using inverse Laplace transformations to obtain distribution of T , $E[T]$ can be obtained directly using the moment generating property of Laplace transforms. We have,

$$E[T] = -\frac{d}{ds} \hat{F}(s) \Big|_{s=0} \quad (31)$$

On solving this, we get the following expression for MTTF,

$$MTTF = E[T] = \frac{\lambda(1 + 2c) + \mu}{2\lambda(\lambda + \mu(1 - c))} \quad (32)$$

The above two examples illustrate Markov modeling technique to predict the system reliability.

Limitations of Markov Models

Even though the Markov modeling method is widely used to analyze the system reliability, it does have several limitations. First, the Markov model becomes too complex if all the possible system states are considered. It leads to the state space explosion. A system having n components may require up to 2^n states in a Markov chain representation. Several approaches such as state truncation, fixed point iteration and hierarchical decomposition techniques are proposed [1, 6]. Second, stiffness is the important problem faced in Markov modeling. For example, in a highly reliable system, with very efficient maintenance, the failure rates are very low and repair rates are high. Their ratio will be of several orders of magnitude and lead to loss of precision. For the stiffness problem, improving the solution results will involve a combination of techniques including better algorithms and acceptable approximations [7, 8]. Finally, the major objection to the use of Markov models in the reliability analysis is the assumption

that sojourn time in any state is exponentially distributed. This does not always realistically represent the observed distribution. One way to deal with non-exponential distributions is the phase type approximation, which consists of modeling a distribution with a set of states and transitions between those states such that the sojourn time in each state is exponentially distributed [4].

Conclusions

The Markov model has been used in several applications such as computer and communication, economics, banking systems, and aviation systems. In this article, we have explained how Markov modeling can be used to evaluate the reliability of the complex systems. Examples are provided to illustrate the technique. However, the application of the Markov model is subject to several limitations such as largeness, stiffness, and the assumption of exponential distributions for failure and repair times. Improving the computer algorithms, use of fast computers, and use of high level modeling formalisms such as stochastic Petri nets tries to fill the gap between Markov modeling and system design specification.

References

- [1] Lanus, M., Liang, Y. & Trivedi, K.S. (2003). Hierarchical composition and aggregation of state based availability and performability models, *IEEE Transactions on Reliability* **52**(1), 44–52.
- [2] Trivedi, K.S., Vasireddy, R., Trindale, D., Swami, N. & Castro, R. (2006). Modeling high availability, *Proceedings of Pacific Rim International Symposium on Dependable Computing (PRDC'06)*, California, pp. 154–164.
- [3] Gokhale, S., Lyu, M. & Trivedi, K.S. (1997). Reliability simulation of fault-tolerant software and systems, *Proceedings of Pacific Rim International Symposium on Fault-Tolerant Systems (PRFTS '97)*, Taipei, pp. 167–173.
- [4] Trivedi, K.S. (2001). *Probability and Statistics with Reliability, Queueing, and Computer Science Applications*, 2nd Edition, John Wiley & Sons, New York.
- [5] Sahner, R.A., Trivedi, K.S. & Puliafito, A. (1996). *Performance and Reliability Analysis of Computer Systems: An Example-Based Approach Using the SHARPE Software Package*, Kluwer Academic Publishers.
- [6] Tomek, L. & Trivedi, K.S. (1991). Fixed-point iteration in availability modeling, in *Informatik-Fachberichte, Volume*

283: *Fehlertolerierende Rechensysteme*, M. Dal Cin, ed, Springer-Verlag, Berlin, pp. 229–240.

- [7] Malhotra, M., Muppala, J.K. & Trivedi, K.S. (1994). Stiffness-tolerant methods for transient analysis of stiff Markov chains, *Microelectronics and Reliability* **34**(11), 1825–1841.
- [8] Bobbio, A. & Trivedi, K.S. (1986). An aggregation technique for the transient analysis of stiff Markov chains, *IEEE Transactions on Computers* **35**(9), 803–814.

Related Articles

Multistate Systems

Repairable Systems Reliability

KISHOR S. TRIVEDI AND DHARMARAJA
SELVAMUTHU

Mathematical Models of Credit Risk

Mathematical modeling of credit risk is concerned with constructing formal models of time evolution of credit ratings (credit migrations, *see Credit Risk Models*) in a pool of credit names, and with studying various properties of such models. In particular, this involves modeling and studying default times and their functionals. Because of the space limitation we do not discuss credit migrations in this article, and we limit the discussion to default events only. As a matter of fact, we only discuss mathematical modeling pertaining to a single obligor, and thus, pertaining to a single default time. Discussion of issues regarding mathematical modeling of multiple default times (*see Default Correlation; Hazard and Hazard Ratio*) and their stochastic dependence, deserves a separate article.

Default time (*see Simulation in Risk Management; Copulas and Other Measures of Dependency*) is a strictly positive random variable, say τ . For simplicity, we assume that τ is continuously distributed on $(0, \infty)$. Typically, one is interested in computing expectations of various functionals of default time, conditioned on available information. Thus, the role of the conditional law of τ , given the information available on the market at current time t , is of main importance. In this context, it is frequently sufficient to study the rate of occurrence of the default, that is, the conditional probability that the default occurs in a small time interval dt , knowing that the default has not occurred before t and perhaps also knowing some other relevant information. However, in the general case when H hypothesis is not satisfied, that knowledge – i.e., the knowledge of the intensity – is not sufficient.

Typical functionals of the default time are defaultable bonds. In what follows, a defaultable zero-coupon (DZC) bond (*see Default Correlation*), with maturity T , pays 1 monetary unit at time T if $\tau > T$, and it pays 0 otherwise. A corporate bond, with recovery process R , pays 1 monetary unit at time T if no default occurs by time T , and it pays R_t at time t if the default occurs at time t , for $t \leq T$. More generally, a generic defaultable claim (X, R, τ) pays X at maturity in case default does not occur prior to or at maturity, and it pays the recovery R_t at default

time, when $\tau = t < T$. We denote by $H_t = \mathbb{I}_{\tau \leq t}$ the default process, by $B(t, T)$ the price of a default-free zero-coupon bond (*see Credit Migration Matrices*) with maturity T , by $\beta_t = \exp(-\int_0^t r_s ds)$ the discount factor, and by $\beta_t^T = \beta_T / \beta_t = \exp(-\int_t^T r_s ds)$. In the case of a deterministic interest rate, $B(t, T) = \beta_t^T$, and in the case of stochastic interest rate $B(t, T) = \mathbb{E}_{\mathbb{Q}}(\beta_t^T | \mathcal{F}_t) = \mathbb{E}_{\mathbb{Q}}(\exp(-\int_t^T r_s ds) | \mathcal{F}_t)$, where $\mathbb{Q}$ is the risk-neutral probability, or the pricing measure, and where $\mathcal{F}_t$ represents conditioning information.

An important class of functionals of default times are credit derivatives (*see Credit Value at Risk*), an example of which is credit default swap (CDS) (*see Credit Risk Models; Simulation in Risk Management*). CDS is a liquid contract referencing a default time τ , and providing a protection against financial loss associated with this default time. The party selling the protection (seller of the CDS) promises to pay to the protection buyer (buyer of the CDS) a protection payment δ_t in case $\tau = t \leq T$, where T is the maturity of the CDS. In return, the protection buyer agrees to pay to the protection seller a premium, which is paid until default time or until maturity of the contract, whichever occurs first. In the financial practice, the premium is paid according to some discrete tenor schedule. For the purpose of mathematical modeling and study of CDSs, it is convenient to assume that the premium is paid continuously in time, at rate κ , which is termed *CDS spread* or *CDS annuity*.

There are two main approaches to modeling default times: the structural approach and the reduced approach, also known as *hazard process approach*.

Structural Models

Classical versions of these models of default assume that the market reveals all needed information about default time *via* prices of related liquid securities, such as prices of equity issued by the default prone obligor. In other words, the default is announced by observation of the prices of traded instruments. Typically, in these models, credit spreads, representing the differentials between yields on defaultable bonds and yields on default-free bonds (treasury bonds), narrow to zero when time approaches the bond's maturity. This, however, is not a phenomenon observed in the market. Another questionable property of classical structural models is that in these models the

default time is predictable, that is, it is announced by a sequence of economic events that can be observed in the market. This property is frequently contradicted by market reality (recall Enron's collapse).

The typical example of a structural model, is the model in which the default time is defined as $\tau = \inf\{t : S_t \leq a\}$, where S represents price process of the underlying equity, assumed to be continuous. There are various mathematical models of the evolution of S , the simplest of which is the geometric Brownian motion model (see **Insurance Pricing/Nonlife; Weather Derivatives**),

$$dS_t = S_t(\mu dt + \sigma dW_t) \quad (1)$$

In this model, pricing and hedging defaultable claims of payoff $h(S_T)\mathbb{1}_{T < \tau}$ reduces to classical hedging of barrier contingent claims. In particular, the price of a DZC is that of a digital barrier option.

Modeling of dependence between several default times is straightforward with the classical structural models, as it amounts to modeling correlation between the processes that trigger the default; typically, this reduces to modeling correlation between Brownian motions. However, closed forms are difficult (or even impossible) to obtain for more than two defaults.

Reduced Models

In these models the default event arrives by surprise; the surprise can be total, or partial. In mathematical terms, this means that the default time is not predictable.

We assume here that there exists a financial market, which we call the full market or the defaultable market. The information flow in this defaultable market is modeled in terms of the so-called full filtration $\mathbf{G} = (\mathcal{G}_t, t \geq 0)$.

Filtration $\mathbf{G}$ consists of two filtrations: $\mathbf{G} = \mathbf{F} \vee \mathbf{H}$. The reference filtration $\mathbf{F}$ corresponds to the default-free component of our market (default-free market or primary market, for short), and is, for simplicity, assumed to be the Brownian filtration. The default filtration $\mathbf{H}$ provides the information regarding occurrence of the default time. Thus, at any time t , the information available in this market consists of the information $\mathcal{F}_t$ provided by the default-free market, and of the knowledge of H_t , that is, knowledge of whether the default has occurred by time t or not.

In what follows, we assume that all random objects are constructed on some probability space $(\Omega, \mathcal{G}, \mathbb{P})$ where $\mathbb{P}$ is the historical probability and where σ -algebra $\mathcal{G}$ is such that $\mathcal{G}_t \subset \mathcal{G}$ for all t . The major modeling issue is modeling of the conditional survival probabilities $\tilde{\mathbb{P}}(\tau > T | \mathcal{F}_t)$, where $\tilde{\mathbb{P}}$ is a probability measure equivalent to $\mathbb{P}$; in particular, we can have $\tilde{\mathbb{P}} = \mathbb{P}$, or $\tilde{\mathbb{P}} = \mathbb{Q}$ – the pricing measure.

Toy Model: τ Independent of F

The total surprise under probability $\tilde{\mathbb{P}}$ occurs when the survival probability does not depend on the conditioning information $\mathcal{F}_t$, that is, for all $t \geq 0$ we have that

$$G^{\tilde{\mathbb{P}}}(t, T) := \tilde{\mathbb{P}}(\tau > T | \mathcal{F}_t) = \tilde{\mathbb{P}}(\tau > T) \quad (2)$$

Since $\mathcal{F}_t$ is the information provided by the default-free market, the above means that the default time is independent of the prices of default-free securities under probability measure $\tilde{\mathbb{P}}$. Thus, we can interpret the total surprise case saying that default-free market provides no information regarding the possible default. Of course, this is not a reasonable hypothesis from the practical perspective, although it needs to be mentioned, that, in practice, it is frequently assumed that default intensity is deterministic, which implies that the default time comes as total surprise. We shall adopt this hypothesis temporarily to illustrate the basic techniques underlying reduced modeling in this very simple set up.

To simplify the notation, we shall write $G^{\tilde{\mathbb{P}}}(t)$ in place of $G^{\tilde{\mathbb{P}}}(t, t) = \tilde{\mathbb{P}}(\tau > t)$. Accordingly, we denote by $F^{\tilde{\mathbb{P}}}$ the cumulative probability distribution of τ , i.e.,

$$F^{\tilde{\mathbb{P}}}(t) = \tilde{\mathbb{P}}(\tau \leq t) = 1 - G^{\tilde{\mathbb{P}}}(t) \quad (3)$$

and we assume that $F^{\tilde{\mathbb{P}}}(t) < 1$ for all $t \geq 0$. Additionally, we assume in the text that follows that $F^{\tilde{\mathbb{P}}}$ is differentiable, and we denote its derivative (the p.d.f. of τ) by $f^{\tilde{\mathbb{P}}}$. Note that for $t < T$, we have

$$\begin{aligned} \tilde{\mathbb{P}}(\tau > T | \tau > t) &= \frac{\tilde{\mathbb{P}}(\tau > T)}{\tilde{\mathbb{P}}(\tau > t)} = \frac{1 - F^{\tilde{\mathbb{P}}}(T)}{1 - F^{\tilde{\mathbb{P}}}(t)} \\ &= \frac{G^{\tilde{\mathbb{P}}}(T)}{G^{\tilde{\mathbb{P}}}(t)} \end{aligned} \quad (4)$$

Hazard Function. Let $\Gamma^{\tilde{\mathbb{P}}}(t) = -\ln(1 - F^{\tilde{\mathbb{P}}}(t))$ be the hazard function (see **Default Risk; Reliability Growth Testing; Lifetime Models and Risk Assessment; Markov Modeling in Reliability**) of τ . In view of our assumptions, we see that it is an increasing function of the form $\Gamma^{\tilde{\mathbb{P}}}(t) = \int_0^t \gamma^{\tilde{\mathbb{P}}}(s) ds$, where

$$\gamma^{\tilde{\mathbb{P}}}(s) ds = \frac{f^{\tilde{\mathbb{P}}}(s)}{1 - F^{\tilde{\mathbb{P}}}(s)} \quad (5)$$

It can be easily shown that the process

$$M_t^{\tilde{\mathbb{P}}} = H_t - \int_0^{t \wedge \tau} \gamma^{\tilde{\mathbb{P}}}(s) ds \quad (6)$$

is a martingale (see **Mathematics of Risk and Reliability: A Select History**) with respect to the filtration generated by process H , under probability $\tilde{\mathbb{P}}$. The superscript $\tilde{\mathbb{P}}$ in the above notations was meant to emphasize that the respective functions and processes depend on the choice of measure $\tilde{\mathbb{P}}$. To simplify the notation, we shall omit the superscript whenever $\tilde{\mathbb{P}} = \mathbb{Q}$. So, for example, we shall write $G(t)$ in place of $G^{\mathbb{Q}}(t)$.

If the primary market is arbitrage free and complete, there exists a unique risk-neutral probability, say $\mathbb{Q}^F$. In the present setting, one cannot hedge defaultable claims using the primary market only, i.e., using the default-free assets traded in the primary market. In particular, measure $\mathbb{Q}^F$ cannot be used to price defaultable securities. For that, we need to know the risk-neutral probability corresponding to the entire defaultable market, that is, a probability measure $\mathbb{Q}$ under which discounted prices of all traded securities, default-free and defaultable, are $\mathbf{G}$ -martingales. Typically, such a measure is calibrated from the observed prices of all underlying securities traded in the defaultable market. Of course, to avoid arbitrage opportunities, the restriction of $\mathbb{Q}$ to the primary market must be equal to $\mathbb{Q}^F$.

Defaultable Zero-Coupon Bonds. We assume here that the default time τ is independent of $\mathbf{F}$ under $\mathbb{Q}$. If a treasury bond of maturity T and a DZC of maturity T are traded in the market at prices $B(t, T)$ and $D(t, T)$ respectively, then the risk-neutral probability of survival can be deduced from these prices. To see this, we first observe that $D(T, T) = \mathbb{1}_{T < \tau}$, and therefore,

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}(D(T, T) | \mathcal{F}_T) &= \mathbb{E}_{\mathbb{Q}}(\mathbb{1}_{T < \tau} | \mathcal{F}_T) = \mathbb{Q}(T < \tau) \\ &= G(T) \end{aligned} \quad (7)$$

Thus, since in view of the first fundamental theorem of asset pricing, the value of a DZC is the conditional expectation of discounted payoff under $\mathbb{Q}$, we obtain

$$\begin{aligned} D(t, T) &= \mathbb{E}_{\mathbb{Q}}(D(T, T) \beta_T^t | \mathcal{G}_t) \\ &= \mathbb{1}_{t < \tau} \frac{\mathbb{E}_{\mathbb{Q}}(D(T, T) \beta_T^t | \mathcal{F}_t)}{\mathbb{E}_{\mathbb{Q}}(\mathbb{1}_{t < \tau} | \mathcal{F}_t)} \\ &= \mathbb{1}_{t < \tau} \frac{\mathbb{E}_{\mathbb{Q}}(\mathbb{E}_{\mathbb{Q}}(D(T, T) | \mathcal{F}_T) \beta_T^t | \mathcal{F}_t)}{\mathbb{E}_{\mathbb{Q}}(\mathbb{1}_{t < \tau} | \mathcal{F}_t)} \\ &= \mathbb{1}_{t < \tau} \frac{G(T)(\beta_T^t | \mathcal{F}_t)}{G(t)} \\ &= \mathbb{1}_{t < \tau} \frac{G(T)B(t, T)}{G(t)} \end{aligned} \quad (8)$$

In particular,

$$\frac{D(0, T)}{B(0, T)} = G(T) \quad (9)$$

Deterministic Interest Rate and Deterministic Recovery. In the particular case of deterministic interest rate, the dynamics of D are

$$\begin{aligned} dD(t, T) &= -\frac{G(T)}{G(t)} B(t, T) dH_t \\ &\quad - (1 - H_t) B(t, T) \frac{G(T)}{G(t)^2} dG(t) \\ &\quad + (1 - H_t) \frac{G(T)}{G(t)} r(t) B(t, T) dt \\ &= r(t) D(t, T) dt - D(t^-, T) (dH_t - \gamma(t) dt) \\ &= r(t) D(t, T) dt - D(t^-, T) dM_t \end{aligned} \quad (10)$$

More generally, the price of a corporate bond with deterministic recovery function R is

$$\begin{aligned} D^{(R)}(t, T) &= \mathbb{1}_{t < \tau} \left(\frac{G(T)}{G(t)} B(t, T) \right. \\ &\quad \left. + \frac{1}{G(t)} \int_t^T B(t, s) R(s) dF(s) \right) \end{aligned} \quad (11)$$

and thus, the risk-neutral dynamics of such a bond are

$$\begin{aligned} dD^{(R)}(t, T) &= (r(t) D^{(R)}(t, T) - R(t) \gamma(t) (1 - H_t)) \\ &\quad \times dt - D^{(R)}(t^-, T) dM_t \end{aligned} \quad (12)$$

The value at time t of a defaultable payoff $\mathbb{1}_{\tau > T} h(S_T)$ is

$$\begin{aligned} & \mathbb{E}_{\mathbb{Q}}(\mathbb{1}_{\tau > T} h(S_T) \beta_T^t | \mathcal{G}_t) \\ &= \mathbb{1}_{t < T} B(t, T) \frac{\mathbb{E}_{\mathbb{Q}}(\mathbb{1}_{\tau > T} h(S_T) | \mathcal{F}_t)}{\mathbb{Q}(\tau > t)} \end{aligned} \quad (13)$$

In this toy model, the independence assumption and the hypothesis that the interest rate is deterministic implies that

$$\begin{aligned} & \mathbb{E}_{\mathbb{Q}}(\mathbb{1}_{\tau > T} h(S_T) \beta_T^t | \mathcal{F}_t) \\ &= B(t, T) \mathbb{E}_{\mathbb{Q}}(h(S_T) \mathbb{E}_{\mathbb{Q}}(\mathbb{1}_{\tau > T} | \mathcal{F}_T) | \mathcal{F}_t) \\ &= B(t, T) G_T \mathbb{E}_{\mathbb{Q}}(h(S_T) | \mathcal{F}_t) \end{aligned} \quad (14)$$

Stochastic Interest Rate. In the case of stochastic interest rate, we assume that the risk-neutral dynamics of the default-free ZC (zero-coupon) bond are

$$dB(t, T) = B(t, T) (r(t) dt + \sigma(t, T) dW_t) \quad (15)$$

Then, the dynamics of D are

$$\begin{aligned} dD(t, T) &= -\frac{G(T)}{G(t)} B(t, T) dH_t \\ &\quad - (1 - H_t) B(t, T) \frac{G(T)}{G(t)^2} dG(t) \\ &\quad + (1 - H_t) \frac{G(T)}{G(t)} dB(t, T) \\ &= D(t, T) (r(t) dt + \sigma(t, T) dW_t) \\ &\quad - D(t^-, T) (dH_t - \gamma(t) dt) \\ &= r(t) D(t, T) dt \\ &\quad + D(t^-, T) (\sigma(t, T) dW_t - dM_t) \end{aligned} \quad (16)$$

Pricing defaultable contingent claims. We now consider a defaultable contingent claim with payoff $\mathbb{1}_{\tau > T} h(S_T)$ and whose the price is

$$\begin{aligned} & \mathbb{E}_{\mathbb{Q}}(\mathbb{1}_{\tau > T} \beta_T^t h(S_T) | \mathcal{G}_t) \\ &= \mathbb{1}_{t < \tau} G(t)^{-1} G(T) \mathbb{E}_{\mathbb{Q}}(h(S_T) \beta_T^t | \mathcal{F}_t) \\ &= \mathbb{1}_{t < \tau} \mathbb{Q}(\tau > T | t > \tau) \mathbb{E}_{\mathbb{Q}}(h(S_T) \beta_T^t | \mathcal{F}_t) \\ &= \mathbb{1}_{t < \tau} \exp\left(-\int_t^T \gamma(s) ds\right) \\ &\quad \times \mathbb{E}_{\mathbb{Q}}(h(S_T) \beta_T^t | \mathcal{F}_t) \end{aligned} \quad (17)$$

Thus, the price of the defaultable contingent claim is the product of the price of the corresponding default-free payoff and of the conditional survival probability. A more interesting way of writing this equality is

$$\begin{aligned} & \mathbb{E}_{\mathbb{Q}}(\mathbb{1}_{\tau > T} \beta_T^t h(S_T) | \mathcal{G}_t) = \mathbb{1}_{t < \tau} \\ & \mathbb{E}_{\mathbb{Q}}\left(h(S_T) \exp\left(-\int_t^T (r_s + \gamma(s)) ds\right) \middle| \mathcal{F}_t\right) \end{aligned} \quad (18)$$

which shows that the price of a defaultable contingent claim is the price of the same promised payoff $h(S_T)$ in a world where the interest rate is $r + \gamma$.

Hedging defaultable contingent claims. It is instructive to give a simple example of hedging of a defaultable claim with payoff $\mathbb{1}_{\tau > T} h(S_T)$. We assume that the primary market is complete. Let $\hat{h}_t = \mathbb{E}_{\mathbb{Q}}(\mathbb{1}_{\tau > T} \beta_T^t h(S_T) | \mathcal{G}_t)$ be the discounted price of the defaultable claim, and $\hat{h}_t^F = \mathbb{E}_{\mathbb{Q}}(h(S_T) \beta_T^t | \mathcal{F}_t)$ be the discounted price of the (default-free) payoff $h(S_T)$. Then, $\hat{h}_t = \mathbb{1}_{t < \tau} \frac{G(T)}{G(t)} \hat{h}_t^F = L_t \hat{h}_t^F$ with $L_t = \mathbb{1}_{t < \tau} \frac{G(T)}{G(t)}$. In particular, if $\hat{D}(t, T)$ and $\hat{B}(t, T)$ denote the discounted prices of DZC and ZC, respectively, then $\hat{D}(t, T) = L_t \hat{B}(t, T)$. Using that $dL_t = L_t^- dM_t$, one gets

$$\begin{aligned} d\hat{h}_t &= -\hat{h}_t^F L_t^- dM_t + L_t d\hat{h}_t^F \\ &= \frac{1}{\hat{B}(t, T)} (\hat{h}_t^F d\hat{D}(t, T) \\ &\quad - L_t \hat{h}_t^F d\hat{B}(t, T)) + L_t d\hat{h}_t^F \end{aligned} \quad (19)$$

Since the primary market is complete, we have $d\hat{h}_t^F = \varphi_t d\hat{S}_t$, where $\hat{S}$ denotes the discounted price of the primary security, and where φ is a predictable process. The following portfolio, constructed on the DZC, the ZC, and the underlying asset

$$\left(\frac{1}{\hat{B}(t, T)} \hat{h}_t^F, -\frac{1}{\hat{B}(t, T)} \hat{h}_t^F L_t, L_t \varphi_t \right) \quad (20)$$

is self-financing (see **Alternative Risk Transfer**) since

$$\begin{aligned} & \frac{1}{\hat{B}(t, T)} \hat{h}_t^F d\hat{D}(t, T) - \frac{1}{\hat{B}(t, T)} \hat{h}_t^F d\hat{B}(t, T) \\ & \quad + L_t \varphi_t d\hat{S}_t = d\hat{h}_t \end{aligned} \quad (21)$$

and thus it is a replicating portfolio for the defaultable claim $\mathbb{1}_{\tau > T} h(S_T)$. This portfolio enjoys the remarkable property that the amount of money invested in the DZC is equal to the value of the claim $h(S_T)$.

CDS. CDS contracts are the most liquid, single name credit derivatives. We illustrate here the mathematical modeling of CDS price process in the simple set up of our toy model.

The (ex-dividend) price at time $t \in [s, T]$ of a CDS contract started at s , with rate κ and protection payment $\delta(\tau)$ at default time, where δ is a function, equals

$$S_t(\kappa) = \mathbb{1}_{t < \tau} \frac{1}{G(t)} \left(\int_t^T B(t, u) \delta(u) G(u) \gamma(u) du - \kappa \int_t^T B(t, u) G(u) du \right) \quad (22)$$

The dynamics of the ex-dividend price $S_t(\kappa)$ on $[s, T]$ are

$$dS_t(\kappa) = -S_{t-}(\kappa) dM_t + (1 - H_t) G_t^{-1} \beta_t^{-1} dm_t + (1 - H_t)(r_t S_t(\kappa) + \kappa - \delta(t) \gamma(t)) dt \quad (23)$$

where the $(\mathbb{Q}, \mathbf{F})$ -martingale m is given by the formula

$$m_t = \mathbb{E}_{\mathbb{Q}} \left(\int_0^T \beta_u^{-1} \delta(u) G(u) \gamma(u) du - \kappa \int_0^T \beta_u^{-1} G(u) du \middle| \mathcal{F}_t \right) \quad (24)$$

Analogous formulae can be obtained in the case of forward credit default swap (CDS) contracts.

General Case

We shall only give here a flavor of mathematical modeling in case of partial surprise. Much more information can be found in the references provided below.

A partial surprise is when one can extract some information from the market, i.e., when $G(t, T) = \mathbb{Q}(\tau > T | \mathcal{F}_t) \neq \mathbb{Q}(\tau > T)$, so that the conditional probabilities $G(t, T)$ form an $\mathbf{F}$ -adapted process.

In fact, the process $G_t = G(t, t) = \mathbb{Q}(\tau > t | \mathcal{F}_t)$ contains all the relevant information about default time, in the sense that

$$G(t, T) = \mathbb{Q}(\tau > T | \mathcal{F}_t) = \mathbb{E}_{\mathbb{Q}}(G_T | \mathcal{F}_t) \quad (25)$$

We now make the assumption that $G_t > 0$. Recall that the full information in our defaultable market is carried by the filtration $\mathbf{G}$. The related, risk-neutral survival probabilities are before time t

$$\begin{aligned} \mathbb{Q}(\tau > T | \mathcal{G}_t) &= \mathbb{1}_{t < \tau} \frac{\mathbb{Q}(\tau > T | \mathcal{F}_t)}{\mathbb{Q}(\tau > t | \mathcal{F}_t)} \\ &= \mathbb{1}_{t < \tau} \frac{\mathbb{E}_{\mathbb{Q}}(G_T | \mathcal{F}_t)}{G_t} \\ &= \mathbb{1}_{t < \tau} \frac{G(t, T)}{G_t} \end{aligned} \quad (26)$$

If τ is an $\mathbf{F}$ stopping time, then $G_t = \mathbb{1}_{\tau > t}$; in this the primary market provides complete information on τ (this is the case of classical structural approach).

The price process of a DZC is now

$$\begin{aligned} D(t, T) &= \mathbb{1}_{t < \tau} \frac{\mathbb{E}_{\mathbb{Q}}(\mathbb{1}_{\tau > T} \beta_T | \mathcal{F}_t)}{\beta_t G_t} \\ &= L_t \beta_T^t \mathbb{E}_{\mathbb{Q}}(G_T | \mathcal{F}_t) \end{aligned} \quad (27)$$

where $L_t = \mathbb{1}_{t < \tau} G_t^{-1}$ is a $\mathbf{G}$ -martingale. It is important to observe that, in full generality, the process G is not decreasing and that its martingale part is of some importance in the pricing methodology. Hedging strategies can be obtained and, obviously, depend strongly on the choice of the dynamics of the traded securities. In particular, the amount of money invested in the defaultable asset is equal to the value of the claim, as soon as the value of the defaultable asset is null after default.

Further Reading

We do not give a long list of references, that would, in any case, be incomplete. We recommend that the reader refers to the following books, where a full up-to-date bibliography is given. More recent works can be found on WWW.DEFAULTRISK.COM.

Bielecki, T.R., Jeanblanc, M. & Rutkowski, M. (2004). Stochastic methods, in *Credit Risk Modelling, Valuation And Hedging, Lecture Notes in Mathematics*, CIME-EMS Summer School on Stochastic Methods in Finance, M. Frittelli & W. Runggaldier, eds, Springer, Bressanone.

- Bielecki, T.R. & Rutkowski, M. (2002). *Credit Risk: Modeling, Valuation and Hedging*, Springer-Verlag, Berlin.
- Bluhm, C., Overbeck, L. & Wagner, C. (2003). *An Introduction to Credit Risk Modeling*, Chapman & Hall.
- Cossin, D. & Pirotte, H. (2001). *Advanced Credit Risk Analysis*, John Wiley & Sons, Chichester.
- Duffie, D. & Singleton, K. (2002). *Credit Risk: Pricing, Measurement and Management*, Princeton University Press.
- Lando, D. (2004). *Credit Risk Modeling*, Princeton University Press.
- Schönbucher, P.J. (2003). *Credit Derivatives Pricing Models*, Wiley Finance, Chichester.

TOMASZ R. BIELECKI AND MONIQUE
JEANBLANC

Mathematics of Risk and Reliability: A Select History

Writing the history about any topic is both challenging and demanding. It is demanding because one needs to acquire a broad perspective about the topic, a perspective that generally comes over time and experience. The challenge of writing history comes from the matter of what to include and what to omit. There is the social danger of offending those readers who feel that their work should have been mentioned, but was not. However, the moral obligation of excluding the works of those who are no more with us is much greater. Writers of history must therefore confront the challenge and draw a delicate line. This task is made easier with the passage of time, because the true impact of a signal contribution is felt only after time has elapsed. By contrast, the impact of work that is incremental or marginal can be judged immediately. It is with the above in mind that the history that follows is crafted. The word “select” in the title of this contribution is deliberate; it reflects the judgment of the authors. Hopefully, the delicate line mentioned before, has been drawn by us in a just and honorable manner. All the same, our apologies to those who may feel otherwise, or those whose works we have accidentally overlooked.

Introduction

From a layperson’s point of view, a viewpoint that predates history, the term *risk* connotes the possibility that an undesirable outcome will occur. However, the modern technical meaning of the term *risk* is different. Here, *risk* is the sum of the product of the *probabilities* of all possible outcomes of an action and the *utilities* (see **Longevity Risk and Life Annuities; Multiattribute Value Functions; Clinical Dose–Response Assessment**) (or consequences) of each outcome. Utilities are numerical values of consequences on a zero to one scale. Indeed, utilities are probabilities and obey the rules of probability [1, p. 56]. They encapsulate one’s preferences between consequences. Thus the notion of risk entails the twin notions of probability and utility. Some adverse outcomes are caused by the failure

or the malfunctioning of certain entities, biological or physical. For such adverse outcomes, the probability of failure of the entity in question is known as the *entity’s unreliability*; its *reliability* (see **Managing Infrastructure Reliability, Safety, and Security; Repair, Inspection, and Replacement Models; Reliability Data; Extreme Values in Reliability; Markov Modeling in Reliability**) is the probability of nonfailure for a specified period of time. In the biomedical contexts, wherein the entity is a biological unit, the term *survivability* (see **Multistate Systems; Multiattribute Value Functions**) is used instead of reliability. Thus, assessing reliability (or survivability) is *de facto* assessing a probability, and *reliability theory* pertains to the methods and techniques for doing such assessments.

The linkage between reliability and risk is relatively new [2]. It is brought about by the point of view that the main purpose of doing a reliability analysis is to make sound decisions about preventing failure in the face of uncertainty. To the best of our knowledge, the first document that articulates this position is Barlow *et al.* [3]. Thus we see that probability, utility, risk, reliability, and decision making are linked, with probability playing a central role, indeed the role of a germinator. Our history of risk and reliability must therefore start with a history of probability.

Probability is a way to quantify uncertainty. Its origins date back to sixteenth-century Europe and discussions about its meaning and interpretation continue until the present day. For a perspective on these, the review articles by Kolmogorov [4] and Good [5] are valuable. The former wholeheartedly subscribes to probability as an objective *chance*, and the latter makes the point that probability and chance are distinct concepts. The founding fathers of probability were not motivated by the need to quantify uncertainty; they were more concerned with action than with interpretation. This enables us to divide the history of probability into three parts: until 1750, 1750–1900, and from 1900. These reflect, in our opinion, three reasonably well-defined periods of development of the mathematics of uncertainty which we label as foundations, maturation, and expansion of applicability.

Some excellent books on the history of probability are by Hald [6, 7], Stigler [8], and von Plato [9]. Since the history of probability is the background for the history of risk and reliability, a reading of these and the exhaustive references therein should provide risk

and reliability analysts a deeper appreciation of the foundations of their subject.

Until 1750: The Foundations of Probability

Insurance was the first field in which the traditional notion of risk had to be quantified. Its use can be traced back four millennia to ancient China and Babylonia, where traders took on the risks of the caravan trade by taking out loans that were repaid if the goods arrived. The ancient Greeks and Phoenicians used marine insurance, while the Romans had a form of life insurance that paid for the funeral expenses of the holder. However, there is no evidence that insurance was a common practice and indeed it disappeared with the fall of the Roman Empire. It took the growth of towns and trade in Renaissance Europe, where risks such as shipwreck, losses from fire, and even kidnap ransom worried the wealthy, for insurance to develop once again.

It was the development of probability in the seventeenth century that finally saw the foundation for the mathematics of risk, and where our brief history can really begin. We should mention first that the mathematization of uncertainty can be traced back to Gioralimo Kardano (1501–1575). However, it was the short correspondence between Pierre de Fermat (1608–1672) and Blaise Pascal (1623–1662) that began the development of modern probability theory. Their correspondence concerned a gambling question called *The Problem of Points*, which is to determine the fair bet for a game of chance in which each player has an equal chance of winning, and the bet is won as soon as either player wins the game a predetermined number of times. The difficulty arises if the number of games to win is different for each player; Fermat’s and Pascal’s correspondence led to a solution. Meanwhile, a contemporary of both, Christiaan Huygens (1629–1695), was one of the earliest scientists to think mathematically about risk. He was motivated by problems in annuities, which at that time were common means for states and towns to borrow money. Huygens wrote up the solution of Fermat and Pascal, and is thus credited with publishing the first book on probability theory [10].

Without the benefit of Fermat’s and Pascal’s theory, Graunt produced the first mortality table by decade [11], from which he concluded that only 1% of the population survived to 76 years. Table 1 shows

Table 1 The table records the percentages of people to die in the first 6 years of life and each decade after to 76 years

Of 100 [“quick conceptions”]	
there dies within the first 6 years	36
The next 10 years, or <i>decad</i>	24
The second decad	15
The third decad	9
The fourth	6
The next	4
The next	3
The next	2
The next	1
[“perhaps but one surviveth 76”]	

Reproduced from [11]. Thomas Roycroft, 1662

this brilliant if unsophisticated effort; see Seal [12] for a discussion of its use.

Graunt’s work happened at the time when property insurance as we know it today began. Following the Great Fire of London in 1666, which destroyed about 13 000 houses, Nicholas Barbon opened an office to insure buildings. In 1680, he established England’s first fire insurance company, “The Fire Office”, to insure brick and frame homes; this also included the first fire brigade. Edmond Halley constructed the first proper mortality table, based on the statistical laws of mortality and compound interest [13]. The table was corrected by Joseph Dodson in 1756, making it possible to scale the premium rate to age; previously, the rate had been the same for all ages.

The idea of a fair price was linked to probability by Jacob Bernoulli (1654–1705), work that was published posthumously by his nephew Nicholas [14]. This work is important because it was the first substantial treatment of probability, and contained the general theory of permutations and combinations, the weak law of large numbers, as well as the binomial theorem. What interested Bernoulli was to apply the Fermat–Pascal idea of a fair bet to other problems where the idea of probability had meaning. He argued that opinions about any event occurring or not were analogous to a game of chance where betting on a certain outcome led to a fair bet. The fair bet then represents the certainty that one attaches to an event occurring. This analogy between games of chance and one’s opinions also appears to have been made at the time of Fermat and Pascal [15]. The law of large numbers was particularly important for this argument because Bernoulli realized that, in practical problems, fair prices could not be deduced exactly

and approximations would have to be found. This allowed him to justify approximating the probability of an event by its relative frequency. Thus, in Bernoulli's ideas we see parts of the two currently dominant interpretations of probability: subjective degree of belief and relative frequency.

The relative frequency idea was further developed by de Moivre [16], who proposed the ideas of independent events, the summation rule, the multiplication rule, and the central limit theorem. This connection between fair prices and probability is the basis for insurance pricing. Bernoulli's and de Moivre's work came during a period of rapid development of the insurance market, spurred on by the growth of maritime commerce in the seventeenth and eighteenth centuries. We have seen that fire insurance had been available since the Great Fire of London, but up to the eighteenth century, most insurance was underwritten by individual investors who stated how much of the loss risk they were prepared to accept. This concept continues to this day in Lloyd's of London, having its beginnings in Edward Lloyd's coffeehouse around 1688 in Tower Street, London, which was a popular meeting place for the shipping community to discuss insurance deals among themselves. Soon after the publication of Bernoulli's work, corporations began to engage in insurance. They were first chartered in England in 1720, and in 1735, the first insurance company in the American colonies was founded at Charleston, SC, by 1750 all the basic ideas of probability necessary for quantifying risk – probability distributions, expected values and the idea of fair price, and mortality – were in place, and were in use in insurance.

1750–1900: Probability Matures

Post-1750, the first notable name is that of Thomas Bayes (1702–1761) and his famous essay on inverse probability [17, 18]. His main contribution was to articulate on the multiplication rule that allows conditional probabilities to be computed from unconditional ones; vitally, this permitted Laplace (1749–1827) to derive the law of total probability and Bayes' law. In contrast to de Moivre, Laplace thought that probability was a rational belief and the rules of probability and expectation followed naturally from this interpretation [19, 20].

Poisson (1741–1840) did much work on the technical and practical aspects of probability, and

greatly expanded the scope and applications of probability. His main contribution was a generalization of Bernoulli's theorem; his seminal work [21] also introduced the Poisson distribution. While Poisson agreed with Laplace's rational belief interpretation of probability, criticisms of this view were raised during second half of the nineteenth century. John Venn (1834–1923) revived the frequency interpretation of probability, hinted at by Bernoulli; however, he took it further to state that frequency was the starting point for defining probability [22].

We note that little attempt had been made so far to quantify the consequences of adverse events through utility and hence to manage risks in a coherent manner. However, we note two developments. First, the idea of utility did arise through Daniel Bernoulli in 1738 and utilitarian philosophers such as Bentham (1748–1832). They proposed rules of rationality that stated individuals desire things that maximize their utility, where positive utility is defined as the tendency to bring pleasure, and negative utility is defined as the tendency to bring pain [23]. Second, the industrial revolution meant that manufacturing and transport carried far graver risks than before, and we do see the first attempts at risk management through regulation. In the United Kingdom, the Factory Act of 1802 (known as the *Health and Morals of Apprentices Act*) started a sequence of such acts that attempted to improve health and safety at work. Following a rail accident that killed 88 people in Armagh, Northern Ireland, the Regulation of Railways Act 1889 made fail-safe brakes as well as block signaling mandatory.

All the main areas of insurance – life, marine, and fire insurance – continued to grow throughout this period. After 1840, with the decline of religious prejudice against the practice, life insurance entered a boom period in the United States. Many friendly or benefit societies were founded to insure the life and health of their members. The close of the nineteenth century finally allows us to say something about mathematical reliability theory; Pearson [24] named the *exponential distribution* for the first time.

From 1900 to the Present: Utility and Reliability Enter

The first half of the twentieth century saw the beginning of the modern era of probability;

Kolmogorov (1903–1987) axiomized probability and in doing so freed it from the confusions of interpretation [25]. This period also saw many developments in the frequency interpretation of probability, and several advances in subjective probability. Von Mises (1883–1953) wrote a paper extolling the virtues of the frequentist interpretation of probability [26]. Together with the work of Karl Pearson (1857–1936) and Fisher (1890–1962), methods of inference under the frequency interpretation of probability became the dominant approaches to data analysis and prediction.

However, at about the same time there were breakthrough developments in the subjective approach to statistical inference and decision making. Noteworthy among these were the work of Ramsey [27] who proposed that subjective belief and utility are the basis of decision making and the nonseparability of probability from utility. Jeffreys' (1891–1986) highly influential book on probability theory combined the logical basis of probability with the use of Bayes' law as the basis of statistical inference [28]. At about this time, de Finetti (1906–1985), unaware of Ramsey's work, adopted the latter's subjectivistic views to produce his seminal work on probability [29], later translated into English [30]. De Finetti is best remembered for the above writings, and his bold statement that *Probability Does Not Exist!*

The period 1900–1950 also saw the laying of the foundations of modern utility theory, from which a prescription for normative decision making comes about. The mathematical basis of today's quantitative risk analysis is indeed normative decision theory. Impetus for a formal approach to utility came from von Neumann and Morgenstern [31] with its interest in rational choice (*see Group Decision*), game theory (*see Managing Infrastructure Reliability, Safety, and Security; Risk Measures and Economic Capital for (Re)insurers; Game Theoretic Methods*), and the modeling of preferences (*see Group Decision; Multiattribute Value Functions*). This was brought to its definitive conclusion by Savage [32], who proposed a system of axioms that linked together the ideas of Ramsey, de Finetti, and von Neumann and Morgenstern. Readable accounts of Savage's brilliant work are in DeGroot [33] and Lindley [1], two highly influential voices in the Bayesian approach to statistical inference (*see Bayesian Statistics in Quantitative Risk Assessment*) and decision making. Not to be overlooked is the 1950 treatise

of Wald (1902–1950) whose approach to statistical inference was decision theoretic [34]. However, unlike that of Savage, Wald's work did not entail the use of subjective prior probabilities (*see Longevity Risk and Life Annuities; Reliability Data; Bayesian Statistics in Quantitative Risk Assessment; Subjective Probability*) on the states of nature.

Hardly mentioned up to now is the mathematical and the statistical theory of reliability. This is because it is only in the 1950s and the 1960s that reliability emerged as a distinct field of study. The initial impetus of this field was driven by the demands of the then newer technologies in aviation, electronics, space, and strategic weaponry. Some of the landmark events of this period are: Weibull's (1887–1961) advocacy of the Weibull distribution for metallurgical failure [35, 36], the statistical analysis of failure data by Davis [37], the proposal of Epstein and Sobel [38] that the exponential distribution be used as a basic tool for reliability analysis, the work of Grenander [39] on estimating the failure rate function and the book of Gumbel [40] on the application of the theory of extreme values for describing failures caused by extremal phenomena (*see Large Insurance Losses Distributions; Risk Measures and Economic Capital for (Re)insurers; Extreme Value Theory in Finance*) such as crack lengths, floods, hurricanes, etc., the approach of Kaplan and Meier [41] (*see Detection Limits; Individual Risk Models; Estimation of Mortality Rates from Insurance Data*) for estimating the survival function under censoring, and the introduction in Watson and Wells [42] of the notion of burn-in. Some, though not all, of this work was described in what we consider to be the first few books on reliability – Bazovsky [43], Lloyd and Lipow [44], and Zelen [45]. Initially, the statistical community was slow to embrace the Weibull distribution as a model for describing random failures; indeed the *Journal of the American Statistical Association* rejected Weibull's 1951 paper. This is despite the fact that the Weibull distribution is a member of the family of extremal distributions [46]. Subsequently, however, the popularity of the Weibull grew because of the papers of Lieblien and Zelen [47], Kao [48, 49], and later the inferential work of Mann [50–52]. Today, along with the Gaussian and the exponential distributions, the Weibull is one of the most commonly discussed distributions in statistics.

Whereas the emphasis of the works mentioned above has been on the statistical analysis of lifetime data, progress in the mathematical and probabilistic aspects was also made during the 1950s and 1960s. A landmark event is the work by Drenick [53] on the failure characteristics of a complex system with the replacement of failed units. It started a line of research in reliability that focused on the probabilistic aspects of components and systems; in a similar vein is a book by Cox [54]. The next major milestone was the paper by Birnbaum *et al.* [55] on the structural representation of systems of components; inspiration for this work can be traced to the classic paper of Moore and Shannon [56] on reliable relays. This was followed by the paper of Barlow *et al.* [57] on monotone hazard rates (*see Statistics for Environmental Toxicity*). This work was highly influential in the sense that it spawned a generation of researchers who explored the probabilistic and statistical aspects of monotonicity from different perspectives. Fault trees (*see Systems Reliability; Markov Modeling in Reliability; Expert Judgment*) also appeared during this decade [58]. Much of this work is summarized in the two books of Barlow and Proschan [59, 60]. There were other notable developments during the late 1960s and mid-1970s, some on the probabilistic aspects, and the others on the statistical aspects. With regard to the former, Marshall and Olkin [61] proposed a multivariate distribution with exponential marginals for describing dependent lifetimes. The noteworthy features of their work are that the distribution was motivated using arguments that are physically plausible, and that its properties bring out some subtle aspects of probability models. At about the same time, Esary *et al.* [62] proposed a notion of dependence that they called *association*. This notion was motivated by problems of system reliability assessment, and the generality of the idea was powerful enough to attract the attention of mathematical statisticians and probabilists to develop it further.

During this period, and perhaps earlier than that, there was important work in reliability also done in the Soviet Union. Indeed, Kolmogorov [4] in his expository papers on statistics, often used examples from reliability and life length studies to motivate his material. Important stochastic process methods were developed, such as in Markov processes and queueing theory, under Gnedenko (1912–1995). The book by Gnedenko *et al.* [63], and the more recent review by

Ushakov [64], gives a perspective on the Soviet work in reliability.

Some other developments in that period were the papers of Cox [65], and of Esary *et al.* [66] and the book by Mann *et al.* [67]. Cox's highly influential paper provided a means for relating the failure rate with covariates. A similar strategy was used in Singpurwalla [68], in the context of *accelerated testing*. The paper by Esary *et al.* on shock models and wear processes was remarkable in two respects. The first is that it addressed a phenomenon of much interest to engineers and produced some elegant results. Second, it paved the way for using stochastic processes to obtain probability models of failure (*see Structural Reliability; Reliability Integrated Engineering Using Physics of Failure; Probabilistic Risk Assessment*) [69]. The book by Mann *et al.* integrated the probabilistic and statistical techniques used in reliability that were prevalent at that time, and by doing so it created a template for the subsequent books that followed. The book was also the first of its kind to make a case for using Bayesian methods for reliability assessment.

Subsequent to the mid-1970s, interest in reliability as an academic discipline took a leap and several books and papers began to appear, and are continuing to appear today. Notable among the former are the books by Lawless [70], Martz and Waller [71], Nelson [72, 73], Gertsbakh [74], Crowder *et al.* [75], Meeker and Escobar [76], Aven and Jensen [77], Singpurwalla and Wilson [78], Hoyland and Rousand [79], and Saunders [80]. With the exception of Martz and Waller [71] and Singpurwalla and Wilson [78], the statistical paradigm guiding the material in the above books has been sample theoretic (i.e., non-Bayesian). In terms of signal developments during the period, two notable ones seem to be Natvig's [81] suggestion to consider multistate systems, and the consideration of subjective Bayesianism in reliability. The latter was triggered by Barlow's interpretation of decreasing failure rates caused by subjective mixing [82], and brought to its conclusion by Gurland and Sethuraman [83]; also see the discussion in Lynn and Singpurwalla [84] of Block and Savits [85]. The book by Spizzichino [86] is an authoritative treatment of the generation of subjective probability models for lifetimes based on *exchangeability* (*see Combining Information; Repair, Inspection, and Replacement Models*).

Some other developments in reliability have come about from the biostatistical perspective of survival analysis. Notable among these are Ferguson's [87] advocacy of the *Dirichlet process* for survival analysis, and Aalen's [88] point process perspective and the martingale approach to modeling lifetimes. The former has been exploited by Sethuraman [89], and the latter by Pena and Hollander [90] and Hollander and Pena [91] in a variety of contexts that are germane to reliability.

To conclude, the last 60 years have seen two trends in risk. First of all, the idea of risk has spread to many other fields outside the traditional areas of insurance and actuarial science. It is now an important idea in medicine, public health, law, science, and engineering. Secondly, driven by its increasing use and by the growth of computing and data-collecting power, increasingly complex quantifications of risk and reliability have been made to make better use of increasing quantities of data; reliability and risk models, inference and prediction with those models, and numerical methods have all advanced enormously. Since the 1960s, in particular, the literature on reliability, risk, and survival analysis has grown in journals that cover statistics, philosophy, medicine, engineering, law, finance, environment, and public policy. Annual conferences on risk in all these subject areas have been held for the last 30 years. In addition to these two trends, we might add that the magnitude of the risks being quantified and managed has increased over the last century; environmental pollution, intensive food production, and the nuclear industry being examples. The same trends in reliability theory can be discerned as those in risk: the spread of application into new fields and the impact of increasing computing power and availability of data. It is worth comparing seminal books on statistical reliability of the 1960s such as Bazovsky [43] and Barlow and Proshan [59] with that of the current decade [2] to see how much the field has changed.

The debate over the interpretation of probability, and uncertainty quantification more generally, continues. The important work of Savage [32], DeGroot [33], and de Finetti [30] publicized the justifications for the laws of probability through their interpretation as a subjective degree of belief. This, along with the practical development of the necessary numerical tools, has increased the use of subjective probability and Bayesian inference in the last 30 years. The strong link between risk,

reliability, and the mathematical tools of probability and decision making, that has existed for 400 years, looks set to continue.

Acknowledgment

The work of Nozer D. Singpurwalla was supported by The Office of Naval Research Grant N00014-06-1-037 and by The Army Research Office Grant W911NF-05-1-2009.

References

- [1] Lindley, D.V. (1985). *Making Decisions*, 2nd Edition, John Wiley & Sons, London.
- [2] Singpurwalla, N.D. (2006). *Reliability and Risk: A Bayesian Perspective*, John Wiley & Sons, Chichester.
- [3] Barlow, R.E., Clarotti, C.A. & Spizzichino, F. (1993). *Reliability and Decision Making*, Chapman & Hall, London.
- [4] Kolmogorov, A.N. (1969). *The Theory of Probability in Mathematics, Its Content, Methods and Meaning, Part 3*, MIT Press, Cambridge, Vol. 2.
- [5] Good, I.J. (1990). Subjective probability, in *The New Palgrave: Utility and Probability*, J. Eatwell, M. Milgate & P. Newman, eds, W. W. Norton, New York.
- [6] Hald, A. (1990a). *A History of Probability and Statistics and Theory Applications Before 1750*, John Wiley & Sons, New York.
- [7] Hald, A. (1990b). *A History of Mathematical Statistics from 1750 to 1930*, John Wiley & Sons, New York.
- [8] Stigler, S.S. (1990). *The History of Statistics: the Measurement of Uncertainty Before 1900*, Harvard University Press, Cambridge.
- [9] von Plato, J. (1994). *Creating Modern Probability*, Cambridge University Press, Cambridge.
- [10] Huygens, C. (1657). Tractatus, de ratiociniis in aleæ ludo, in *Excercitationum Mathematicarum Libri Quinque*, F. Schooten, ed, Elsevier, Lugduno Batava.
- [11] Graunt, J. (1662). *Natural and Political Observations... Made Upon the Bills of Mortality*, Roycroft, London. Reproduced in a more modern format in *Journal of the Institute of Actuaries* **90**(1).
- [12] Seal, H.L. (1980). Early uses of Graunt's life table, *Journal of the Institute of Actuaries* **107**, 507–511.
- [13] Halley, E. (1693). An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the City of Breslaw; with an attempt to ascertain the price of annuities upon lives, *Philosophical Transactions of the Royal Society of London* **17**, 596–610. Reproduced in *Journal of the Institute of Actuaries* **112**, 278–301.
- [14] Bernoulli, J. (1713). *Ars Conjectandi*, Thurnisiorum Fratrum, Basileæ.
- [15] Arnauld, A. & Nicole, P. (1662). *L'art de Penser*, Paris.

- [16] de Moivre, A. (1718). *The Doctrine of Chances*, London. Reprinted by Chelsea, New York in 1967, published in 1990 by Wiley & Sons, and online by Wiley from 2005.
- [17] Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society of London* **53**, 370–418. Reprinted in 1958; see Barnard (1958).
- [18] Barnard, G.A. (1958). Thomas Bayes' essay towards solving a problem in the doctrine of chances, *Biometrika* **45**, 293–315.
- [19] Laplace, P.-S. (1812). *Théorie Analytique des Probabilités*, Paris. Reprinted in *Oeuvres* **10**, 295–338, 1894.
- [20] Laplace, P.-S. (1814). *Essai Philosophique Sur les Probabilités*, Paris. 6th edition translated by F. W. Truscott and F. L. Emory as *A Philosophical Essay on Probabilities*, 1902.
- [21] Poisson, S.-D. (1837). *Recherches Sur la Probabilité des Jugements en Matière Criminelle et Matière Civile*.
- [22] Venn, J. (1866). *The Logic of Chance*, McMillan, London. Third edition published by McMillan & Co., London, 1888. Fourth edition published by the Chelsea Publishing Company, New York, 1962.
- [23] Bentham, J. (1781). *An Introduction to the Principles of Morals and Legislation*. Latest edition published by Adamant Media Corporation, 2005.
- [24] Pearson, K. (1895). Contributions to the mathematical theory of evolution II: skew variation in homogeneous material, *Philosophical Transactions of the Royal Society of London* **186**, 343–414.
- [25] Kolmogorov, A.N. (1956). *Foundations of the Theory of Probability*, 2nd Edition, Chelsea Publishing Company, New York. Translation edited by Nathan Morrison.
- [26] von Mises, R. (1919). Grundlagen der wahrscheinlichkeitsrechnung, *Mathematische Zeitschrift* **5**, 52–99.
- [27] Ramsey, F.P. (1931). Truth and probability, in *The Foundations of Mathematics and Other Logical Essays*, R.B. Braithwaite, ed, Kegan Paul, London, pp. 156–198.
- [28] Jeffreys, H. (1939). *Theory of Probability*, Oxford University Press, Oxford.
- [29] de Finetti, B. (1937). Calcolo delle probabilità, in *Atti dell'XI Convegno, Torino-Aosta*, Associazione per la Matematica Applicata alle Scienze Economiche Sociali. Typescript for the Academic Year 1937–1938, University of Padua.
- [30] de Finetti, B. (1974). *Theory of Probability*, John Wiley & Sons, New York, **2**.
- [31] von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Modern Behavior*, Princeton University Press, Princeton.
- [32] Savage, L.J. (1954). *The Foundations of Statistics*, 1st Edition, John Wiley & Sons, New York.
- [33] DeGroot, M.H. (1970). *Optimal Statistical Decisions*, McGraw-Hill, New York.
- [34] Wald, A. (1950). *Statistical Decision Functions*, John Wiley & Sons, New York.
- [35] Weibull, W. (1939). A statistical theory of the strength of materials, *Ingeniörs Vetenskaps Akademiens, Handlingar* **151**, 45–55.
- [36] Weibull, W. (1951). A statistical distribution function of wide applicability, *ASME Journal of Applied Mechanics* **18**, 293–297.
- [37] Davis, D.J. (1952). An analysis of some failure data, *Journal of the American Statistical Association* **47**, 113–150.
- [38] Epstein, B. & Sobel, M. (1953). Life testing, *Journal of the American Statistical Association* **48**, 486–502.
- [39] Grenander, U. (1956). On the theory of mortality measurement, I and II, *Skandinavisk Aktuarietidskrift* **39**, 70–96, 125–153.
- [40] Gumbel, E.J. (1958). *Statistics of Extremes*, Columbia University Press, New York.
- [41] Kaplan, E.L. & Meier, P. (1958). Non-parametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.
- [42] Watson, G.S. & Wells, W.T. (1961). On the possibility of improving the mean useful life of items by eliminating those with short lives, *Technometrics* **3**, 281–298.
- [43] Bazovsky, I. (1961). *Reliability Theory and Practice*, Prentice-Hall, Englewood Cliffs.
- [44] Lloyd, D.K. & Lipow, M. (1962). *Reliability: Management, Methods and Mathematics*, Prentice-Hall, Englewood Cliffs.
- [45] Zelen, M. (1963). *Statistical Theory of Reliability*, The University of Wisconsin Press, Madison. As editor, Proceedings of an advanced seminar by the Mathematical Research Center, U.S.A. Army.
- [46] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire, *Annals of Mathematics* **44**, 423–453.
- [47] Liebliin, J. & Zelen, M. (1956). Statistical investigation of the fatigue life of deep-groove ball bearings, *Journal of Research of the National Bureau of Standards* **57**, 273–316.
- [48] Kao, J.H.K. (1958). Computer methods for estimating Weibull parameters in reliability studies, *Transactions of IRE – Reliability Quality Control* **13**, 15–22.
- [49] Kao, J.H.K. (1959). A graphical estimation of mixed Weibull parameters in life testing electron tube, *Technometrics* **1**, 389–407.
- [50] Mann, N.R. (1967). Tables for obtaining best linear invariant estimates of parameters of Weibull distribution, *Technometrics* **9**, 629.
- [51] Mann, N.R. (1968). Point and interval estimation procedures for 2-parameter Weibull and extreme-value distributions, *Technometrics* **10**, 231.
- [52] Mann, N.R. (1969). Optimum estimators for linear functions of location and scale parameters, *Annals of Mathematical Statistics* **6**, 2149–2155.
- [53] Drenick, R.F. (1960). The failure law of complex equipment, *Journal of the Society for Industrial and Applied Mathematics* **8**, 680–690.
- [54] Cox, D.R. (1962). *Renewal Theory*, Methuen, London.
- [55] Birnbaum, Z.W., Esary, J.D. & Saunders, S.C. (1961). Multi-component systems and structures and their reliability, *Technometrics* **3**, 55–77.

- [56] Moore, E.F. & Shannon, C. (1956). Reliable circuits using less reliable relays I, *Journal of Franklin Institute* **262**, 191–208.
- [57] Barlow, R.E., Marshall, A.W. & Proschan, F. (1963). Properties of probability distributions with monotone hazard rate, *Annals of Mathematical Statistics* **34**, 375–389.
- [58] Watson, H.A. (1961). *Launch Control Safety Study*, Bell Labs, Murray Hill, Vol. 1 Section VII.
- [59] Barlow, R. & Proschan, F. (1965). *Mathematical Theory of Reliability*, John Wiley & Sons, New York.
- [60] Barlow, R. & Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing*, 1st Edition, Holt, Rinehart and Winston, New York.
- [61] Marshall, A.W. & Olkin, I. (1967). A multivariate exponential distribution, *Journal of the American Statistical Association* **62**, 30–44.
- [62] Esary, J.D., Proschan, F. & Walkup, D.W. (1967). Association of random variables, with applications, *The Annals of Mathematical Statistics* **38**, 1466–1474.
- [63] Gnedenko, B.V., Belyaev, Y.K. & Soloyev, A.D. (1969). *Mathematical Models of Reliability Theory*, Academic Press, New York.
- [64] Ushakov, I. (2000). Reliability: past, present and future, in *Recent Advances in Reliability Theory: Methodology, Practice and Inference*, Birkhäuser, Berlin, pp. 3–22.
- [65] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.
- [66] Esary, J.D., Marshall, A.W. & Proschan, F. (1973). Shock models and wear processes, *Annals of Probability* **1**, 627–649.
- [67] Mann, N.R., Schafer, R.E. & Singpurwalla, N.D. (1974). *Methods for Statistical Analysis of Reliability and Life Data*, John Wiley & Sons, New York.
- [68] Singpurwalla, N.D. (1971). Problem in accelerated life testing, *Journal of the American Statistical Association* **66**, 841–845.
- [69] Singpurwalla, N.D. (1995). Survival in dynamic environments, *Statistical Science* **10**, 86–113.
- [70] Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.
- [71] Martz, H.F. & Waller, R.A. (1982). *Bayesian Reliability Analysis*, John Wiley & Sons, New York.
- [72] Nelson, W. (1982). *Applied Life Data Analysis*, John Wiley & Sons, New York.
- [73] Nelson, W. (1990). *Accelerated Testing*, John Wiley & Sons, New York.
- [74] Gertsbakh, I.B. (1989). *Statistical Reliability Theory*, Marcel Dekker, New York.
- [75] Crowder, M.J., Kimber, A.C., Smith, R.L. & Sweeting, T.J. (1991). *Statistical Analysis of Reliability Data*, Chapman and Hall, London.
- [76] Meeker, W.Q. & Escobar, L.A. (1998). *Statistical Methods for Reliability Data*, John Wiley & Sons, New York.
- [77] Aven, T. & Jensen, U. (1999). *Stochastic Models in Reliability Series: Stochastic Modeling and Applied Probability*, Springer, New York, Vol. 41.
- [78] Singpurwalla, N.D. & Wilson, S.P. (1999). *Statistical Methods in Software Reliability: Reliability and Risk*, Springer, New York.
- [79] Hoyland, A. & Rousand, M. (2004). *System Reliability Theory: Models and Statistical Methods*, John Wiley & Sons, New York.
- [80] Saunders, S.C. (2006). *Reliability, Life Testing and Prediction of Service Lives*, Springer, New York.
- [81] Natvig, B. (1982). Two suggestions of how to define a multistate coherent system, *Advances in Applied Probability* **14**, 434–455.
- [82] Barlow, R.E. (1985). A Bayes explanation of an apparent failure rate paradox, *IEEE Transactions on Reliability* **34**, 107–108.
- [83] Gurland, J. & Sethuraman, J. (1995). How pooling data may reverse increasing failure rates, *Journal of the American Statistical Association* **90**, 1416–1423.
- [84] Lynn, N.J. & Singpurwalla, N.D. (1997). Burn-in makes us feel good, *Statistical Science* **12**, 13–19.
- [85] Block, H.W. & Savits, T.H. (1997). Burn-in, *Statistical Science* **12**, 1–13.
- [86] Spizzichino, F. (2001). *Subjective Probability Models for Lifetimes*, Chapman & Hall/CRC, Boca Raton.
- [87] Ferguson, T.S. (1973). A Bayesian analysis of some non-parametric problems, *Annals of Statistics* **1**, 209–230.
- [88] Aalen, O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.
- [89] Sethuraman, J. (1994). A constructive definition of Dirichlet priors, *Statistica Sinica* **4**, 639–650.
- [90] Pena, E.A. & Hollander, M. (2004). Models for recurrent events in reliability and survival analysis, in *Mathematical Reliability: an Expository Perspective*, R. Soyer, T.A. Mazzuchi & N.D. Singpurwalla, eds, Kluwer Academic Publishers, Boston, pp. 105–123.
- [91] Hollander, M. & Pena, E.A. (2004). Nonparametric methods in reliability, *Statistical Science* **19**, 644–651.

NOZER D. SINGPURWALLA AND SIMON P.
WILSON

Mercury/Methylmercury Risk

The toxicity of mercury vapor has been known since ancient times, and Bernadino Ramazzini described the occupational health risks 300 years ago. The risks of environmental contamination were dramatically illustrated in the 1960s and 1970s by poisoning incidents in Minamata (Japan) and in Iraq. The current focus is on the risks of neurotoxic damage due to the presence of methylmercury in freshwater fish and seafood. Recent risk assessments include those of the UN Environment Program [1] and the US Environmental Protection Agency [2], and these sources may be consulted for further information and additional references.

Mercury in the environment originates from both natural and anthropogenic sources. Emissions from volcanoes and other natural sources constitute about 1000 tons/year. The total annual emission from anthropogenic sources is about 2500 tons, the major source being energy production from fossil fuels, especially coals with high mercury contents. Exposure to mercury vapor occurs in diminishing industrial operations, while the general population is primarily exposed to amalgam fillings.

In sediments, inorganic mercury is methylated by microorganisms to methylmercury, which is accumulated in fish. Methylmercury attains its highest concentrations in large predatory species at the top of the aquatic and marine food chains, e.g., pike, swordfish, and tuna, as well as shark, seals, and toothed whales. Increased levels of exposure have been documented in Japan, the Mediterranean region, and especially in Arctic populations. Methylmercury is readily absorbed from the gastrointestinal tract and also passes the blood–brain barrier and the placenta. In adults, the earliest symptoms are paresthesias of the fingers, the tongue, and the face, followed by problems in motor coordination. Later, tunnel vision and hearing problems occur. This syndrome was called *Minamata disease* from its appearance in a Japanese fishing village, where the local fish had been contaminated with mercury effluents from a local factory. Similar symptoms were observed in Iraq and other several countries after farmers had eaten seed grain treated with methylmercury fungicides.

Children and the fetus are more susceptible to the toxic effects of methylmercury than adults. Thus, Minamata mothers appeared in good health, but gave birth to children with a spastic paresis syndrome and severe neurological deficits. The main target organ for methylmercury toxicity is the developing brain. Epidemiological studies of prenatal exposures from maternal seafood consumption have shown more subtle neuropsychological and neurophysiological deficits that remained detectable through adolescence [3–5]. Detectable developmental delays seem to appear at maternal hair mercury concentrations of $1\text{--}3\text{ }\mu\text{g g}^{-1}$ [3] (i.e., cord blood concentrations of $4\text{--}12\text{ }\mu\text{g l}^{-1}$). In adults, the earliest clinical effects, such as paresthesias, appear to occur when blood concentrations are above $200\text{ }\mu\text{g l}^{-1}$. There is no effective antidote.

Recent epidemiological studies suggest that adverse cardiovascular effects may also occur at low-level exposures prevalent among fish consumers [6]. Although the mechanism is unknown, this evidence supports the notion that methylmercury exposure should be avoided by nonpregnant adults too.

The epidemiology evidence has inspired downward revision of methylmercury exposure limits and limits for environmental mercury releases [1, 2]. Also, governmental agencies have issued advisories regarding consumption of freshwater fish and seafood. Still, the occurrence of essential nutrients suggests that fish should be included in the diet, but maximizing the benefits would depend on the choice of fish low in mercury, i.e., smaller species, low in the food chain and from uncontaminated waters. An international agreement has been reached that mercury releases to the environment should be decreased.

The central nervous system is also the main target organ in exposures to mercury vapor. Adverse effects are well known from occupational health reports, but documented adverse effects from using amalgam fillings are apparently rare. Nonetheless, the use of amalgam in dentistry is being phased out as a preventable source of environmental mercury releases.

References

- [1] UNEP (United Nations Environment Programme) (2002). *Global Mercury Assessment Report*, Geneva, at <http://www.chem.unep.ch/mercury/Report/Final%20Assessment%20report.htm> (access Mar 12, 2006).

- [2] U.S. EPA (Environmental Protection Agency) (2001). *Water Quality Criterion for the Protection of Human Health: Methylmercury*, Publication EPA-823-R-01-001, Washington, DC, At <http://www.epa.gov/waterscience/criteria/methylmercury/document.html> (access Mar 1, 2007).
- [3] Grandjean, P., Weihe, P., White, R.F., Debes, F., Araki, S., Yokoyama, K., Murata, K., Sørensen, N., Dahl, R. & Jørgensen, P.J. (1997). Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury, *Neurotoxicology Teratology* **19**, 417–428.
- [4] Murata, K., Weihe, P., Budtz-Jørgensen, E., Jørgensen, P.J. & Grandjean, P. (2004). Delayed brainstem auditory evoked potential latencies in 14-year-old children exposed to methylmercury, *Journal of Pediatrics* **144**, 177–183.
- [5] Debes, F., Budtz-Jørgensen, E., Weihe, P., White, R.F. & Grandjean, P. (2006). Impact of prenatal methylmercury toxicity on neurobehavioral function at age 14 years, *Neurotoxicology Teratology* **28**, 363–375.
- [6] Virtanen, J.K., Voutilainen, S., Rissanen, T.H., Mursu, J., Tuomainen, T.P., Korhonen, M.J., Valkonen, V.P., Seppanen, K., Laukkanen, J.A. & Salonen, J.T. (2005). Mercury, fish oils, and risk of acute coronary events and cardiovascular disease, coronary heart disease, and all-cause mortality in men in eastern Finland, *Arteriosclerosis, Thrombosis, and Vascular Biology* **25**, 228–233.

ESBEN BUDTZ-JØRGENSEN AND PHILIPPE
GRANDJEAN

Meta-Analysis in Clinical Risk Assessment

Meta-analysis is a statistical approach that provides a logical structure and a quantitative methodology for the review, evaluation, and synthesis of information from independent studies. Combining the results of multiple independent studies in a well-conducted meta-analysis provides an objective summary assessment of the overall results of a series of independent studies. With the use of meta-analytic techniques, a series of clinical trials (*see* **Randomized Controlled Trials**) for example, can be summarized, and the overall treatment effect can be estimated with increased precision so that the risks of reaching incorrect conclusions from individual trials are reduced. In this article, we introduce the process of meta-analysis, issues in the conduct of a meta-analysis, and some statistical models that are commonly used to conduct a meta-analysis. The goals are to provide guidance for the interpretation of published meta-analyses as well as a framework and introduction to the tools for the conduct of a meta-analysis of clinical trials. We note that meta-analysis itself is observational in that we are summarizing experience without any experimental manipulation and deciding which studies are comparable for combinability. While meta-analyses can be conducted to summarize observational studies including case-control (*see* **Case-Control Studies**) and cohort studies (*see* **Cohort Studies**), our focus in this article is on the meta-analysis of randomized clinical trials. The material in this article is based on the paper by Friedman and Goldberg [1].

In medicine, there is a long tradition of literature review to assess the current state of knowledge with respect to the treatment of a given disease entity. The purposes of the literature review include efforts to determine the studies that have been done to address the question at hand; the results of these studies; how to compare and combine the results of the individual studies; how to interpret the results to make treatment recommendations; and what other studies need to be carried out.

The tradition of literature reviews is by its nature retrospective. When the results of published studies are inconsistent or inconclusive, it is difficult to reach a conclusion and recommendation for treatment. The inconsistencies or inconclusiveness of published

reports can result from small study sizes, varying study designs, different patient populations, varying quality of the published studies, changing technology for diagnosis and outcome measurements over time, publication bias, or the absence of negative studies in the published literature, among other reasons.

Meta-analysis methods attempt to provide a quantitative structure to address these questions. There is no one method of meta-analysis. It “is not a statistical method, *per se*, but rather an orientation toward research synthesis that uses many techniques of measurement and data analysis” [2].

The incorporation of quantitative methods into the review and synthesis of information represents an opportunity to make explicit the judgments that inevitably influence the conclusions of any summary of research in a substantive area. Many of the statistical methods used in meta-analysis predate the term, which was introduced by Glass [3]. Cochran [4] developed statistical principles for combining studies that are analogous to those for summarizing the overall results of a single study. A fundamental issue in statistics is how to assess the risk or benefit associated with the use of a new treatment. For example, how to obtain a single overall estimate of the risk of mortality in patients with a heart attack, who are treated with aspirin compared with placebo in a single randomized trial conducted in multiple centers, requires that we make several assumptions. In particular, the homogeneity of the population under study in the single trial is assumed. We can then combine the observations into a single estimate of risk and obtain its associated estimate of variability [5]. These same ideas provide a framework for meta-analysis.

In the context of using statistical methods in meta-analysis, the unit of analysis is a “study” or “trial”. Judgments are needed to decide what constitutes a study; which studies to collect; which studies to include; and then, to determine how comparable the studies are with respect to patients, design, conduct; what are the strengths and weaknesses of each study; how to define the results; what are comparable results; how to combine results; how to cope with disparity among results; and finally, how to interpret and communicate the results. Judgments about a study are made from the available reports of that study. Although there may be multiple reports of a study, no study should be included more than once in any summary. The judgments regarding the

selection of studies for inclusion and the relative importance of each study are as important as the choice of methods to combine them. The resulting meta-analysis across studies provides an increased power compared to a single trial to detect potentially important treatment effects.

Example

In 1980, six randomized multicenter clinical trials that were conducted during the 1970s, some concurrently, had been completed to evaluate the effect of aspirin compared with placebo on mortality after a heart attack. The results of these six trials are shown in Table 1 [1, 6, 7]. In contrast to each multicenter trial, there was no *a priori* plan for combining the six studies; the six studies were not planned to be conducted under a common protocol, although they overlapped in time. What could be concluded about the effects of aspirin on mortality? Was there sufficient evidence for a physician to decide to recommend aspirin for an individual patient? From a meta-analysis of these six aspirin trials, we seek to combine information to obtain a useful summary of the trial experience.

Outcome Measures

For a single study, the effect of the treatment or intervention can be measured in different ways that depend on the nature of the outcome measurement. Two treatment groups can be compared in several ways: difference between means, difference between proportions, relative risks, relative odds or odds ratios, logarithms of the odds ratio, and logarithms of the relative risks. All of these measures can be derived from the binary data for a single trial displayed as a 2×2 table. In the aspirin example shown in Table 1, the outcome variable is the total mortality rate and the treatment effect is defined as the difference between the mortality rates in the treatment and control groups. The standardized treatment effect is this difference divided by its standard error. Meta-analysts call this latter quantity the “effect size”. It is used for summaries across studies when the outcome measures are different. In the aspirin example, the outcome measures are the same for each study.

Combining Results of Studies

One of the questions to be addressed as part of a meta-analysis is can the results of the studies be combined to provide an overall estimate of treatment effect and its variability? A first step in the analysis phase of any meta-analysis is to develop and display the data to allow for a comparison of the estimated treatment effects and their associated variability from the multiple studies.

Figure 1 is an example of a commonly used graph to display this information.

Each study from Table 1 is shown with its treatment effect and associated 95% confidence limit centered at the effect (difference in mortality rates ± 2 SE). From Figure 1, we observe (a) that all of the 95% confidence limits include zero and (b) that the estimated treatment effect from AMIS is in the opposite direction. The two most commonly used methods for combining results are the simple fixed- and random-effects models. Results from these two methods, which we discuss later, are also shown. From these analyses, we note that although none of the studies were conclusive (because all of the 95% confidence limits included zero), five studies appear to favor aspirin and one does not. However, the overall result does not conclusively favor aspirin because of the dominant effect of AMIS, the largest study. In fact, without AMIS, the combined results of the remaining five studies would conclusively favor aspirin regardless of the method used to combine the information.

An alternative view of the data is to look at the relative risk of mortality rates for treatment compared with control. These estimates are also shown in Table 1. Figure 2 provides a graphical display based on the relative risks and the associated 95% confidence intervals.

Even with carefully designed protocols to conduct studies under similar guidelines, difficulties and surprises invariably occur. Centers in a single multicenter trial often are heterogeneous, complicating interpretation of the results. Indeed, in a multicenter trial, the results at the individual centers sometimes differ [5]. This problem becomes even more complex when we attempt to summarize across studies. Summary statistics, even from a single multicenter study, can be misleading when patients are heterogeneous or interactions exist among patient characteristics, treatment, and outcome. Careful attention to

Table 1 Numbers of patients and all cause mortality rates from six randomized clinical trials comparing the use of aspirin and placebo by patients after a heart attack^(a)

Study	Aspirin		Placebo		Aspirin – placebo			Aspirin/placebo		
	Number of patients	Mortality rate (%)	Number of patients	Mortality rate (%)	Difference (%)	Standard error (s_i)	Standardized difference (%/standard error)	Relative risk of mortality	Ln relative risk	Odds ratio
UK-1	615	7.97	624	10.74	-2.77	1.65	-1.68	0.74	-0.298	0.72
CDPA	758	5.8	771	8.3	-2.5	1.31	-1.91	0.7	-0.358	0.68
GAMS	317	8.52	309	10.36	-1.84	2.34	-0.79	0.82	-0.196	0.806
UK-2	832	12.26	850	14.82	-2.56	1.67	-1.54	0.83	-0.19	0.803
PARIS	810	10.49	406	12.81	-2.31	1.98	-1.17	0.82	-0.2	0.798
AMIS	2267	10.85	2257	9.7	1.15	0.903	+1.27	1.12	0.112	1.133
										0.125

^(a) Adapted from [1, 6, 7]

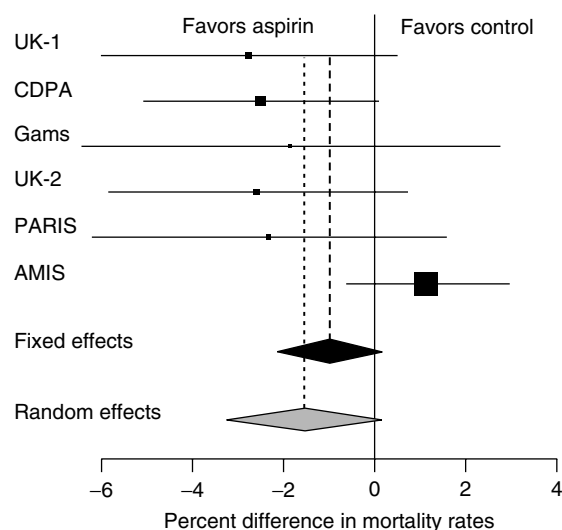


Figure 1 Effect of treatment with aspirin on total mortality rate compared with placebo with 95% confidence intervals: difference in mortality rates (aspirin/placebo). Size of symbol for estimate is proportional to study size; line lengths for individual studies are 95% confidence intervals. Summary estimates are widest at point estimate and width is 95% confidence interval

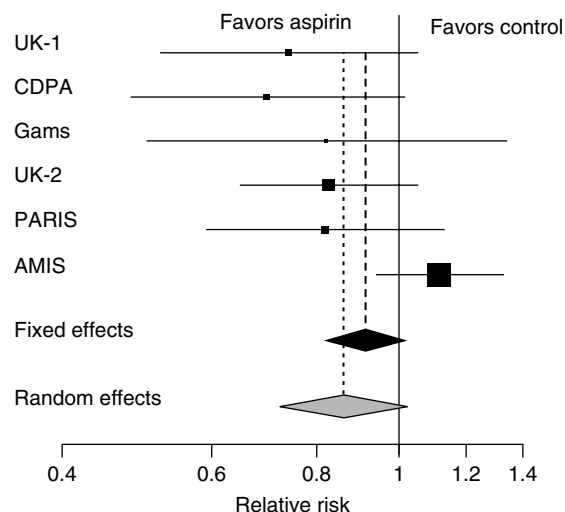


Figure 2 Effect of treatment with aspirin on total mortality rate compared with placebo with 95% confidence intervals: relative risk of mortality (aspirin/placebo). Size of symbol for estimate is proportional to study size; line lengths for individual studies are 95% confidence intervals. Summary estimates are widest at point estimate and width is 95% confidence interval

design and conduct issues and to operational definitions can yield insight into the interpretation of disparate results. Thus, if we are faced with these data *de novo*, we would attempt to reach some understanding of the disparity of results in these apparently similar long-term randomized clinical trials by introducing a viable framework to interpret the results and to assess the plausibility of combining the results. Our approach would be to define comparable subgroups within each study that would be predictive of outcome and to investigate the consistency of the results within strata defined by these subgroups. More specifically, the strategy would be to determine influential variables related to outcome. To do this, we would examine baseline characteristics in the different studies; effects of clinical characteristics on outcome; and patterns of interaction with treatment. Multiple methods of analysis should also be used to assess the consistency and interpretability of results.

Then we would revisit the question of which trials to combine. The ability to carry out this exercise depends on the availability of the required data in each of the publications.

A key aspect of this approach involves looking within each multicenter trial to assess the homogeneity of the centers that were combined to form the estimate of the overall treatment effect for the multicenter study. Thus, in attempts to interpret disparity among studies, we have to first look within studies to disaggregate if there is heterogeneity, and then to recombine.

Canner [6] in his overview of the six aspirin trials, carried out some adjustments for baseline characteristics among studies. Although he was unable to examine the details within studies from the published reports, he attempted to assess comparability across the studies. From his review, we note that the maximum length of patient follow-up was 12–30 months, with a minimum of 2 months except for PARIS and AMIS, where all patients were followed for 35–48 months. Thus, long-term total mortality on study may have a different meaning in the various studies. In fact, if we examine only year 1, we note that relative mortality was comparable across all six studies. A more recent meta-analysis [8] of aspirin and antiplatelet drugs that combined the results of 300 trials involving 140 000 patients has developed recommendations for subgroups of patients.

Statistical Models in Meta-Analysis

In this section, we focus on the use and interpretation of some commonly used methods of analysis that include fixed- and random-effects models.

Fixed-Effects Model

Statistical models for combining effects fall into two major classes. The simplest, or fixed-effect model, assume that there is a single “true” effect and that each study to be combined provides an unbiased estimate of this true effect. This assumption is called the *homogeneity of studies assumption*. The estimates from the individual studies may have different variability. In mathematical notation, we describe this model for N independent studies by

$$d_i = \delta + e_i \quad (1)$$

where d_i is the observed effect for the i th study and δ is the unknown true effect. The e_i is assumed to be an independent, normally distributed random variable with average value zero, no bias, and variance v_i . The variance of the i th study is estimated by s_i^2 , where s_i is the standard error of the mean treatment effect d_i in study i . For the aspirin example shown in Table 1, we see the observed values of the d_i (difference between aspirin and placebo mortality rates) and their estimated standard errors s_i . Given the fixed model assumptions, statistical theory says that the “best” estimate of the true effect is given by a weighted average of the estimated effects, where the weights are inversely proportional to the estimated variances of the d_i . In this case, the weights, w_i , are $1/s_i^2$.

Thus, the estimate of $\delta = \sum_{i=1}^6 (w_i d_i) / \sum_{i=1}^6 w_i = -0.98$ for the fixed-effect model for the aspirin example. The fixed-effect model assumptions allow for an estimate of the standard error of the estimate, which is computed as $(1 / \sum_{i=1}^6 w_i)^{1/2}$. In this instance, the standard error of the estimated true effect is 0.58 (95% confidence interval for $-0.98 \pm 1.96 (0.58) = (-2.11, 0.16)$).

Thus, the fixed-effects model-based analysis leads to a result that is consistent with no statistically significant treatment effect at the 5% level of significance because the confidence interval includes zero. If, however, we analyze only the five studies without AMIS, with the same model assumptions, we obtain an estimated overall effect of -2.47 with

standard error of 0.7567, giving 95% confidence limits of $(-3.95, -0.99)$, suggesting that aspirin usage provides a statistically significant and practical reduction on mortality at the 5% level of significance. But there was no reasonable evidence on the basis of the available knowledge at that time to assume that there was something inherently different about AMIS [6]. Table 2 provides a summary of the results of meta-analyses using the various models and different estimates of treatment effect. Notice that the fixed- and random-effects models yield identical results without AMIS, since the source of heterogeneity has been removed. We also note that year-1 mortality is significantly lower with aspirin when the six studies are combined for year 1 using this model. Therefore, the choice of endpoint as well as the choice of studies to be included can have a major effect on the conclusions of the analysis.

Estimates of the overall treatment effect measured as the relative risk from the fixed-effects model along with associated 95% confidence limits are shown in Figure 2 for total mortality. Figure 2 illustrates the results using the relative risk as the measure of treatment effect.

Tests for Homogeneity of Studies

Because the studies that are identified for inclusion in a meta-analysis are generally not designed under a common protocol, there is no *a priori* rationale for the use of the fixed-effects model with its assumption of a common treatment effect. There are statistical tests of homogeneity to examine this assumption. The Q -statistic [9] is widely used to test the null hypothesis of a common true treatment effect across all studies by a comparison of the variability within studies to the variability between studies. Failure to reject this null hypothesis because the between-study variability is not statistically significant suggests that it is plausible to combine the results of a group of studies. Rejection of the null hypothesis indicates heterogeneity among the studies and that the between-study variability must be accounted for in combining the results of the studies.

For the aspirin example, $Q = \sum_{i=1}^6 w_i (d_i - \hat{\delta})^2 = 9.57$. This statistic has a χ^2 distribution with $N - 1 = 5$ degrees of freedom under the assumption that the individual treatment effects are normally distributed with a common mean (observed $P = 0.089$). This level casts a reasonable statistical doubt on the

Table 2 Summary of meta-analysis results by model and effect measure for six randomized controlled clinical trials comparing aspirin and placebo after a heart attack

Studies available	Method of analysis	Estimate of the effect measure (95% confidence interval)		
		Difference (%)	Odds ratio	Ln odds ratio
All	Fixed effects	−0.98 (−2.12, 0.16)	0.90 (0.80, 1.02)	−0.10 (−0.23, 0.02)
	Random effects	−1.51 (−3.20, 0.19)	0.84 (0.697, 1.023)	−0.17 (−0.36, 0.02)
Without AMIS	Fixed effects	−2.47 (−3.95, −0.99)	0.76 (0.65, 0.90)	−0.27 (−0.43, −0.10)
	Random effects	−2.47 (−3.95, −0.99)	0.76 (0.65, 0.90)	−0.26 (−0.43, −0.10)
All	Bayesian fixed effects ^(a)	−0.98 (−2.14, 0.20)	0.89 (0.80, 1.0)	−0.11 (−0.24, 0.004)
	Bayesian random effects ^(a)	−1.54 (−4.12, 0.51)	0.839 (0.63, 1.05)	−0.18 (−0.458, 0.044)

^(a) Priors are $d \sim N(0, 10\,000)$ and $\sigma \sim U(0, 100)$

validity of the assumptions of the fixed-effects model because such tests of homogeneity have low statistical power. In contrast, without AMIS, $Q = 0.12$, with 4 degrees of freedom and observed significance level of $P = 0.998$. This failure to reject lends credibility to the assumption that these other five studies may be measuring a common treatment effect. However, if we include AMIS in the analysis, we need to account for the between-study variability.

Additional approaches to assess heterogeneity among studies are discussed in Higgins and Thompson [9].

Random-Effects Models

Rejection of the null hypothesis of homogeneity does not tell us whether the between-study differences are systematic or random. The random-effects model for meta-analysis assumes that the individual treatment effects are random. In the context of the aspirin example, the random-effects assumptions are that each study is estimating a different true effect, δ_i . The simple random-effects model incorporates the heterogeneity of the studies through the assumption that the studies are a random sample from a population of studies. The mean of the population of studies is the “true effect”. An estimate of the population variance is then added to the within-study variances, and a weighted average of individual study effects is

used to estimate the “true effect” in the population. In the simplest random-effects model, this population of studies has true mean effect δ , and the distribution of $\delta - \delta_i$ is assumed to be normal, mean zero, and variance τ^2 .

Thus, in this model, by assumption, between-study variability is measured by τ^2 . The effect of this model is to increase the variability associated with the d_i by τ^2 . Under this random-effects model, the overall treatment effect is estimated, as in the fixed-effects model, by a weighted average of the individual effects with weights inversely proportional to the variance of the d_i 's. In the random-effects case, the variance of the $d_i = s_i^2/\tau^2$ and the weights are $w_i^* = [1/(s_i^2 + \tau^2)]$. The overall estimate of the treatment effect is given by the same formula as above with the w_i replaced by the w_i^* . To apply this procedure, one must obtain an estimate of τ^2 . There are a number of technically complex methods to estimate τ^2 ; these methods depend on the s_i^2 , the within-study variances. We do not pursue these details here. The estimate of τ^2 attributable to DerSimonian and Laird [10] used in the following computations is 2.048 [7, 11]. Based on the w_i^* , the estimate of the overall treatment effect is −1.505 (approximate 95% confidence interval for the true treatment effect of (−3.20, 0.20)). This larger treatment effect and corresponding larger standard error is also shown in Figure 1. However, the confidence limits that result

from the random-effects model still include zero, and the conclusions of the fixed-effects model are unchanged. If we do not believe (and we do not) that these six studies are a random sample from a population of studies as required for the simple random-effects model, then we need to continue to explore the sources of heterogeneity among these studies.

Recently, Berkey *et al.* [12] proposed a random-effects regression model for meta-analysis that allows for the inclusion of other available information that may be used to explain heterogeneity among studies. If the studies are homogeneous, then both the simple random-effects model and this type of regression model reduce to the fixed-effects model. Even more complicated hierarchical models can be used with the available information to model differences among the treatment effects within studies and between studies [13].

The introduction of complications into the analysis opens up another level of judgment and decision making to ensure that statistical assumptions do not dominate the data. When there is a large disparity among effects, good measurements and expert judgments about explanatory variables are needed to provide credibility to the statistical models. The R software package is useful for the computations required for these approaches [14].

Bayesian Methods

More recently, Bayesian methods (*see Bayesian Statistics in Quantitative Risk Assessment*) that can incorporate additional information through the specification of prior distributions have become feasible for the conduct of a meta-analysis [7, 15, 16]. The development of Gibbs sampling and Markov chain Monte Carlo (MCMC) methods and the availability of software such as WinBUGS have facilitated the use of these approaches [17–19]. For the aspirin example, the results of these types of analyses with the frequentist approaches are shown in Table 2 for the different measures of treatment effect. We note that the credible intervals generated from these methods are usually wider than the corresponding confidence intervals generated from the frequentist fixed- or random-effects models.

Berry *et al.* [20] conducted a Bayesian meta-analysis to compare the dose response relationships of lipid-lowering agents that explicitly compared the

results of a study with historical data that illustrate the utility of these methods.

We note that all of the different statistical models and tests of homogeneity apply to all of the different measures of treatment effect, including the $\ln$ (relative risk) and $\ln$ (relative odds).

Other Issues in the Conduct of Meta-Analysis

Historical Perspective

There are many excellent references that review quantitative approaches to research reviews. *Summing Up* [21] is one such reference. Olkin discusses the history and goals of meta-analysis along with some examples in his introductory chapter in the Wachter and Straf publication based on a 1986 workshop [1]. The literature of meta-analysis developed along two separate paths in the social sciences and in the medical sciences. The development in the social sciences is well described in [1, 22, 23]. In the medical sciences, the 1986 Workshop on Overviews in Clinical Research was published, with discussions in *Statistics in Medicine* (1987). Many of the references cited elsewhere in this article are found in that issue, along with discussions [24–30]. In 1993, an issue of *Statistical Methods in Medical Research* devoted to meta-analysis in clinical medicine described current thinking and controversies [31–35]. The issue of *Statistical Science* devoted to meta-analysis brings together technical issues and medical and social science applications [35–40].

Statistical Methods for Combining Evidence in Studies.

The need to combine observations is ubiquitous in statistical analysis. Any time two or more observed quantities are combined into a single summary value, we are faced with a statistical problem that requires a close look at the validity of model assumptions as well as the soundness of the theoretical and philosophical basis underlying the process. The emergence of meta-analysis has fueled the interest of statisticians in combining results from different studies. This interest is reflected in the volume *Combining Information*, a publication of the National Academy of Sciences [7] that has been reissued as the first issue of *Contemporary Statistics* (vol. 1), which examines statistical issues and cross-disciplinary opportunities for research. This report

highlights the similarity of techniques in different disciplines under different names for combining information and provides an exposition of methods, free of the jargon of any one discipline, including statistics.

The published literature of meta-analyses frequently contains tests of significance of the overall treatment effect that is based on methods that we have not described. These include Fisher's method for summing the logarithms of the P values associated with each individual study, summing the P values themselves, and testing the average of the P values. Approaches such as these include each study with equal weight regardless of sample size or variability; because each study is given equal weight, these methods have fallen into disfavor [26]. Standardized treatment effects can be added using different weighting schemes to produce a weighted z -statistic as well.

Peto [28] proposes the use of the log rank statistic (or Cochran's statistic), which provides a test statistic based on summing the differences between the observed and expected numbers of events in each study weighted by their variances; this is the (O-E) method and essentially weights the $\ln$ (relative odds) by study size in a fixed-effects model framework. The Mantel-Haenszel χ^2 test [41] is also used for this same situation; the overall estimated log odds provided by this approach is preferable to the Peto estimate, unless the sample sizes in the treatment and control groups are comparable. These approaches essentially consider each study as a separate stratum in the fixed-effect model framework. Fleiss [33] and DeMets [26], among others, provide careful expositions of these methods and their application.

Other reviews of statistical approaches for combining similar experiments can be found in Hedges and Olkin [22], Whitehead and Whitehead [42], Goodman [43], Zucker and Yusuf [44], and Berlin *et al.* [45], who, among others, discuss alternative methods of meta-analysis. A critical review of the use and application of meta-analytical results may be found in Thompson and Pocock [46] and Boden [47]. Mann [48] reviews the diversity of applications of meta-analysis. A more recent review and exposition of methods can be found in the volume edited by Cooper and Hedges [23].

The book by Bryk and Raudenbush [13] provides a good introduction to some of the more sophisticated modeling techniques that we just discussed.

This type of hierarchical modeling will become more prevalent in meta-analysis as the view of Rubin [49] becomes more accepted. He makes the distinction between meta-analysis for "literature synthesis" and meta-analysis for "understanding the science". In the former, the concern is with sampling from the finite population of studies that have been done and summarizing the results of these studies by some average effect with some appropriate weighting, as in the fixed-effects approaches we described previously.

The references provided in the body of this paper provide further examples of the attempt to use statistical methods to aid in the understanding of disparate results.

Guidelines for Planning and Carrying out Meta-Analysis

To successfully carry out a meta-analysis, the techniques of statistical analysis must be integrated with the subject matter. To accomplish this task, expert medical knowledge of the subject matter must be brought to bear on the key decision points in the process. There is no theory or formalism to direct the process of carrying out a meta-analysis. At best, there are informal guidelines for the practice of meta-analysis. These informal guidelines are best understood in the context of case studies.

The best opportunity for carrying out a successful meta-analysis would arise from the prospective planning of a sequence of studies for combinability along the lines of a multicenter trial. Short of prospective planning, standards within subject matter areas for the reporting primary results of clinical studies need to be developed. Suggestions along these lines have been made by Mosteller [50]. Other useful sources of guidelines for the conduct of meta-analyses include Wachter and Straf [2] and Cooper and Hedges [23]. These issues address the identification of studies for inclusion in the meta-analysis, retrieval techniques, selection of studies, quality of studies, and analysis. The Cochrane Collaboration [51] is a major source of guidance for the conduct of meta-analyses or overviews and is a repository for overviews based on the patient level data from clinical trials in a variety of areas of medicine.

Issues in Meta-Analyses in Medicine

Most of the issues that were identified in the publication of the 1986 Workshop on Methodologic Issues

in Clinical Trials [6, 24–30] are still being debated. The issue of *Statistical Methods in Medical Research* in 1993 [31–35] was devoted to these same issues. These issues, critical to the conduct and interpretation of a meta-analysis, include publication bias or study selection problems, assessment of study and data quality, criteria for inclusion and exclusion of studies in the meta-analysis, statistical methods for combining, and the interpretation and generalizability of the results. The collection of “all” studies is difficult and time consuming [23], and involves identification and assembly of published and unpublished literature identified through indices, Medline, registries, and files. Further, many studies appear in several different published formats, as abstracts, as proceedings, as interim reports, and as final reports. Because the assumption in the applicable statistical methods requires that studies be independent, only one report of any study should be included – the final report.

The opposite problem to be faced is “what else is there”; i.e., what studies have not been published? What is the effect of publication bias on the results of a meta-analysis? Negative studies are not as likely as positive studies to be published; are there enough such studies to change the results of a meta-analysis of published literature? The “file-drawer problem” has been addressed by several investigators who have developed methods to estimate the number of negative studies [22, 52]. Other approaches include methods of weighting in the analysis [22, 36–38].

The issue of which studies to include requires that criteria be developed. The decisions to consider here are whether to restrict the analysis only to randomized, controlled trials or to include other kinds of studies. We restricted ourselves to randomized, controlled trials for the reasons discussed earlier. Even with this restriction, many decisions need to be made explicit. Study design, treatment regimens, control regimens, patient inclusion and exclusion criteria, study size, duration of patient observation, outcome assessments, and study quality need to be explicitly documented and discussed. Chalmers *et al.* [24] and Sacks *et al.* [53, 54] recommend dual, blinded abstraction and review, including quality assessments of studies, to develop the databases for analysis. Alternatively, however, one can consider the equivalent of the “intent-to-treat” notion in the reporting of results of individual randomized clinical

trials. In the context of clinical trials, intent-to-treat refers to the notion that all randomized patients are included in the analysis; that is, there are no *post hoc* exclusions. An intent-to-treat design for a clinical trial requires detailed *a priori* operational definitions to enable inclusion and exclusion criteria before randomization to be consistently applied by all investigators. The spirit of this view would lead to the inclusion of almost all identified randomized trials. This is consistent with the views of Peto [28], who suggests minimal exclusion criteria and eliminates many of the so-called objective judgments that are themselves subjective. Quality and other aspects of the studies can be addressed in the process of meta-analysis by the use of sensitivity analysis to assess the impact of any questionable studies on the results of the meta-analysis. For example, the effects of the inclusion of abstracts can also be assessed in this way. Sensitivity analysis would consist of the recomputation of estimated treatment effects with the study or studies in question removed and examination of the influence of the removed observations on the results and conclusions.

Others go beyond the selection of studies for inclusion to suggest guidelines for the inclusion of patient data within studies. This approach requires that data from individual patients or subgroups of patients be obtained. The meta-analysis is then based on the individual data within studies in the analysis [25, 28, 51]. The purpose of such an exercise is to put all of the data from multiple sources into a single database with common conventions and procedures surrounding the processing of the information. This is, of course, only possible to the extent that the procedures and measurements in the trials are similar. The opportunities available to the researcher who obtains such data are presented in our discussion of the aspirin data.

Why Meta-Analysis?

The purpose of meta-analysis is itself a source of controversy. Mosteller and Chalmers [39], Antman *et al.* [11], and Chalmers and Lau [35] have proposed that small, inconclusive studies can be combined in an ongoing manner so that large trials may be avoided if the accumulating evidence from small trials becomes conclusive. That is, before a new large study is undertaken, prior evidence should be evaluated. Peto [28] proposes meta-analysis primarily to

assess the presence of moderately small effects on endpoints such as mortality, essentially in lieu of large, simple trials. He and his colleagues also argue that individual patient data should be included in meta-analyses or overviews to enable common operational definitions to be used. Such large, simple trials have minimal entry criteria and data collection requirements, and are closest to an overview in spirit in that the results obtained are average results over large, possibly heterogeneous patient groups.

Concluding Remarks

The concerns of meta-analysis touch on fundamental issues in science, including what is evidence; how does it accumulate; and how do we learn from experience? The statistical methods applicable to meta-analysis are evolving and becoming more complex; information in speciality areas of medicine is growing rapidly. Many of the issues that we have raised in this article are still being debated. While this debate continues, any report of a meta-analysis should make explicit all judgments regarding the issues that we identified in the earlier section on guidelines for planning and carrying out a meta-analysis. A meta-analysis should include the peer-reviewed published reports of all identified randomized studies only once. This is the primary meta-analysis. If any other types of reports or abstracts of studies of any type are included in the review, the results should be reported separately for the various types of reports to enable the reader to assess the possible effects on the conclusions of the primary meta-analysis. All other judgments and descriptions of analysis should also be made explicit, and adequate information should be provided to allow the reader to make his or her own judgments.

There is a serious requirement to plan for meta-analysis in a manner analogous to the multicenter clinical trial model. Furthermore, one should not underestimate the inherent difficulties in searching for studies and in finding useful information in the studies. We recommend that the process of meta-analysis be carried out collaboratively between subject matter experts and statisticians.

Although meta-analysis is often said to increase power to detect a statistically significant treatment

effect, such an exercise can be misleading if the judgments concerning combinability are incorrect that is, if the studies included are seriously flawed or heterogeneous. Real disparity among studies is not resolved by statistical models alone. Model assumptions are not an adequate substitute for the detective work required to uncover the causes of disparity. Furthermore, the sensitivity and robustness of the results to judgments about the subject matter and to statistical assumptions should be analyzed and reported. Although we have shown that different methods of analysis can give rise to some differences in the results of a meta-analysis (and its interpretation), the choices made by the meta-analyst with respect to all of the issues that we have discussed can dominate any analysis.

As Mosteller [50] points out, a meta-analysis still may be considered successful if one can document that the available information is inadequate to answer the question at hand. In such a case, if the question is important enough, a well-designed and well-conducted primary research study should be performed.

In the process of summarizing overall results and treatment effects, one should be careful not to overlook the anomalies that have been uncovered in the process. These anomalies may be very real contributions to the subject area.

References

- [1] Friedman, H.P. & Goldberg, J.D. (1996). Meta-analysis: an introduction and point of view, *Hepatology* **23**, 917–928.
- [2] Wachter, K.W. & Straf, M.L. (eds) (1990). *The Future of Meta-Analysis*, Russell Sage Foundation, New York.
- [3] Glass, G.V. (1976). Primary, secondary, and meta-analysis of research, *Educational Research* **6**, 3–8.
- [4] Cochran, W.G. (1954). The combination of estimates from different experiments, *Biometrics* **10**, 101–129.
- [5] Goldberg, J.D. & Koury, K.J. (1992). Design and analysis of multicenter trials, in *Statistical Methodology in the Pharmaceutical Sciences*, D.A. Berry, ed, Marcel Dekker, New York, pp. 201–237.
- [6] Canner, P.L. (1987). An overview of six clinical trials of aspirin in coronary heart disease, *Statistics in Medicine* **6**, 255–263.
- [7] National Research Council (1992). *Combining Information: Statistical Issues and Opportunities for Research*, National Academy Press, Washington, DC.
- [8] Aldous, P. (1994). A hearty endorsement for aspirin, *Science* **263**, 24.

- [9] Higgins, J.P.T. & Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis, *Statistics in Medicine* **21**, 1539–1558.
- [10] DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials, *Controlled Clinical Trials* **7**, 177–188.
- [11] Antman, E.M., Lau, J., Kupelnick, B., Mosteller, F. & Chalmers, T.C. (1992). A comparison of results of meta-analysis of randomized control trials and recommendations of clinical experts, *Journal of the American Medical Association* **268**, 240–248.
- [12] Berkey, C.S., Hoaglin, D.C., Mosteller, F. & Colditz, G.A. (1995). A random effects regression model for meta-analysis, *Statistics in Medicine* **14**, 395–411.
- [13] Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models*, Sage Publications, Newbury Park.
- [14] R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*, Version 2.5.1, R Foundation for Statistical Computing, Vienna, at <http://www.R-project.org>.
- [15] Carlin, B.P. & Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd Edition, Chapman & Hall/CRC Press, Boca Raton.
- [16] Parmigiani, G. (2002). Meta-analysis, in *Modeling in Medical Decision Making: A Bayesian Approach*, G. Parmigiani, ed, John Wiley & Sons, West Sussex, pp. 123–163.
- [17] Spiegelhalter, D.J., Thomas, A. & Best, N.G. (1999). *WinBUGS with DoodleBUGS Version 1.4.2 User Manual*, Imperial College and Medical Research Council, at <http://www.mrc-bsu.cam.ac.uk/bugs>.
- [18] Lunn, D.J., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility, *Statistics and Computing* **10**, 325–337.
- [19] Warn, D.E., Thompson, S.G. & Spiegelhalter, D.J. (2002). Bayesian random effects meta-analysis of trials with binary outcomes: methods for absolute risk difference and relative risk scales, *Statistics in Medicine* **21**, 1601–1623.
- [20] Berry, D.A., Berry, S.M., McKellar, J. & Pearson, T.A. (2003). Comparison of the dose response relationships of 2 lipid-lowering agents: a Bayesian meta-analysis, *American Heart Journal* **145**, 1036–1045.
- [21] Light, R. & Pillemer, D.B. (1984). *Summing Up*, Harvard University Press, Cambridge.
- [22] Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*, Academic Press, Orlando.
- [23] Cooper, H. & Hedges, L.V. (1994). Potentials and limitations of research synthesis, in *The Handbook of Research Synthesis*, H. Cooper & L.V. Hedges, eds, Russell Sage Foundation, New York, Chapter 32.
- [24] Chalmers, T.C., Levin, H., Sacks, H.S., Reitman, D., Berrier, J. & Nagalingam, R. (1987). Meta-analysis of clinical trials as a scientific discipline. I: control of bias and comparison with large co-operative trials, *Statistics in Medicine* **6**, 315–325.
- [25] Collins, R., Gray, R., Godwin, J. & Peto, R. (1987). Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews, *Statistics in Medicine* **6**, 245–250.
- [26] DeMets, D.L. (1987). Methods for combining randomized clinical trials: strength and limitations, *Statistics in Medicine* **6**, 341–348.
- [27] Hennekens, C.H., Buring, J.E. & Hebert, P.R. (1987). Implications of overviews of randomized trials, *Statistics in Medicine* **6**, 397–402.
- [28] Peto, R. (1987). Why do we need systemic overviews of randomized trials? *Statistics in Medicine* **6**, 233–240.
- [29] Wittes, R.E. (1987). Problems in the medical interpretation of overviews, *Statistics in Medicine* **6**, 269–276.
- [30] Yusuf, S. (1987). Obtaining medically meaningful answers from an overview of randomized clinical trials, *Statistics in Medicine* **6**, 281–286.
- [31] Pocock, S. (1993). Guest editorial, *Statistical Methods in Medical Research* **2**, 117–119.
- [32] Oakes, M. (1993). The logic and role of meta-analysis in clinical research, *Statistical Methods in Medical Research* **2**, 147–160.
- [33] Fleiss, J.L. (1993). The statistical basis of meta-analysis, *Statistical Methods in Medical Research* **2**, 121–145.
- [34] Thompson, S.G. (1993). Controversies in meta-analysis: the case of the trials of serum cholesterol reduction, *Statistical Methods in Medical Research* **2**, 173–192.
- [35] Chalmers, T.C. & Lau, J. (1993). Meta-analytic stimulus for changes in clinical trials, *Statistical Methods in Medical Research* **2**, 161–172.
- [36] Dear, K.B.G. & Begg, C.B. (1992). An approach for assessing publication bias prior to performing a meta-analysis, *Statistical Science* **7**, 237–245.
- [37] Hedges, L.V. (1992). Modelling publication selection effects in meta-analysis, *Statistical Science* **7**, 246–255.
- [38] Iyengar, S. & Greenhouse, J.B. (1988). Selection models and the file drawer problem, *Statistical Science* **3**, 109–135.
- [39] Mosteller, F. & Chalmers, T.C. (1992). Some progress and problems in meta-analysis of clinical trials, *Statistical Science* **7**, 227–236.
- [40] Olkin, I. (1992). Meta-analysis: methods for combining independent studies, *Statistical Science* **7**, 226.
- [41] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.
- [42] Whitehead, A. & Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials, *Statistics in Medicine* **10**, 1665–1677.
- [43] Goodman, S.N. (1989). Meta-analysis and evidence, *Controlled Clinical Trials* **10**, 188–204.
- [44] Zucker, D. & Yusuf, S. (1989). The likelihood ratio versus the p-value in meta-analysis: where is the evidence? Comment on the paper by S. N. Goodman, *Controlled Clinical Trials* **10**, 205–208.
- [45] Berlin, J.E., Laird, N.M., Sacks, H.S. & Chalmers, T.C. (1989). A comparison of statistical methods for combining event rates from clinical trials, *Statistics in Medicine* **8**, 141–151.

- [46] Thompson, S.G. & Pocock, S.J. (1991). Can meta-analysis be trusted? *Lancet* **338**, 1127–1130.
- [47] Boden, W.E. (1992). Meta-analysis in clinical trials reporting: has a tool become a weapon? *The American Journal of Cardiology* **69**, 681–686.
- [48] Mann, C. (1990). Meta-analysis in the breech, *Science* **249**, 476–480.
- [49] Rubin, D. (1990). A new perspective, in *The Future of Meta-Analysis*, K.W. Wachter & M.L. Straf, eds, Russell Sage Foundation, New York, pp. 55–165.
- [50] Mosteller, F. (1990). Summing up, in *The Future of Meta-Analysis*, K.W. Wachter & M.L. Straf, eds, Russell Sage Foundation, New York, pp. 185–190.
- [51] J.P.T. Higgins & S. Green (eds) (2005). *Cochrane Handbook for Systematic Reviews of Interventions 4.2.5*, The Cochrane Library, Issue 3, John Wiley & Sons Chichester.
- [52] Rosenthal, R. (1979). The “file drawer problem” and tolerance for mere results, *Psychological Bulletin* **86**, 638–641.
- [53] Sacks, H.S., Berrier, J., Reitman, D., Ancona-Berk, V.A. & Chalmers, T.C. (1987). Meta-analyses of randomized controlled trials, *The New England Journal of Medicine* **316**, 450–455.
- [54] Sacks, H.S., Berrier, J., Reitman, D., Pagano, D. & Chalmers, T.C. (1992). Meta-analyses of randomized control trials: an update of the quality and methodology, in *Medical Uses of Statistics*, 2nd Edition, J.C. Builar & F. Mosteller, eds, NEJM Books, Boston, pp. 427–442.

Related Articles

Meta-Analysis in Nonclinical Risk Assessment

JUDITH D. GOLDBERG, HEATHER N. WATSON
AND HERMAN P. FRIEDMAN

Meta-Analysis in Nonclinical Risk Assessment

Background

Systematic review and meta-analysis methods are widely used to summarize and combine results of clinical trials (*see* **Randomized Controlled Trials**), forming the basis for evidence-based medicine [1]. They are also increasingly popular in epidemiology, but they remain relatively rare in toxicology despite some early signs of their uptake there [2, 3]. They have been used for some time in disciplines other than health research, such as education and psychology. Increasing adoption of such techniques reflects the emphasis on “evidence-based” approaches, as well as explosion of publication rates of primary studies beyond the capacity of many researchers. Systematic reviews use explicit, transparent, repeatable criteria to identify, review, and evaluate evidence while seeking to minimize possible biases [4, 5] in identifying, evaluating, and summarizing all the available and relevant evidence concerning a clearly focused, predefined research question. Where appropriate, systematic reviews can be extended into meta-analyses: the quantitative synthesis of results from individual studies.

Systematic reviews and meta-analyses have the potential to help improve the objectivity and transparency of the current risk assessment process through contribution to several aspects of risk assessment, principally in the risk estimation phase. Potentially, they can contribute whenever more than one estimate of a parameter in a risk assessment or decision analysis model (*see* **Multiaattribute Utility Functions**) is available, provided these estimates can legitimately be combined. They will most obviously be useful in summarizing evidence on exposure–disease relationships, from animal toxicology experiments (*see* **Cancer Risk Evaluation from Animal Studies**) or from epidemiological studies in humans, or perhaps at a later stage in combining these two (and potentially other) types of evidence.

Use of systematic criteria force the process to be more systematic, thus allowing readers and users of systematic reviews to have a better understanding of

the results and recommendations given the assumptions and decisions made in the review process [6]. Advantages of using systematic review methods to review and summarize evidence are well documented in several health and related research fields [4, 5, 7, 8]. In a systematic review (but not necessarily in a traditional or narrative review), the target research question; methods used to identify, obtain and select relevant evidence of an appropriate quality; and the results obtained, including any assumptions and judgments made in synthesizing primary data are set out in a standardized and explicit way [4, 5]. This provides transparency of methodology, allows reproducibility, and facilitates updating as new evidence becomes available. Deficiencies in the existing literature may be highlighted, so future research priorities can be identified.

Quantitative synthesis of results from several studies (meta-analysis) may be advantageous and appropriate. Meta-analyses offer greater statistical power and more precise estimates than the primary studies, if it is appropriate to combine the results of a set of studies investigating a common intervention, exposure, or feature. They also provide a framework for investigation of sources of heterogeneity; combining several sources of evidence, derived from a number of studies possibly with differing designs, for example, may increase the generalizability of the results, and allow a full exploration of the effects in subgroups. In the broad risk assessment context, extensions of the standard meta-analytical techniques into comprehensive decision modeling approaches begin to merge with probabilistic risk assessment modeling approaches (*see* **Comparative Risk Assessment**).

Areas of application thus include risk assessment procedures for setting standards for human exposure to environmental pollutants (*see* **Environmental Hazard**) and safety testing and dose establishment in development of therapeutic drugs. In these fields, review and combination of evidence often rely on expert but informal methods, but sometimes on quantitative methods based on biomathematical models. Assumptions and judgments are implicit throughout, whichever approach is adopted, but are typically more transparent when systematic review and meta-analysis, or other explicit modeling approaches, are employed. Transparency regarding the various assumptions made, and identification and quantification of uncertainty in the resultant exposure limit estimates are essential for those involved in the review

process and for professional users and “consumers” of the results alike.

Outline of Systematic Review and Meta-Analysis Methods

There are a number of important steps in any systematic review, and these are described and discussed in the procedural documentation from the Cochrane Collaboration [1] and the Centre for Reviews and Dissemination [9]. The steps include

1. formulating the research question to be addressed in the systematic review/meta-analysis;
2. development of an explicit protocol for the systematic review/meta-analysis;
3. identifying and selecting primary studies to be included in the systematic review/meta-analysis;
4. assessing the quality of the primary studies;
5. extraction of the results of the primary studies;
6. synthesis (if appropriate) of primary study results through meta-analysis;
7. interpreting results of systematic reviews/meta-analyses and formulating recommendations based on them.

These aspects of a systematic review are important regardless of the area of application. Several sets of guidelines on the review process have been available for some time for the combination of randomized controlled trials [1, 9]; for epidemiological studies [10, 11], where meta-analysis often needs to be used in an exploratory rather than definitive analysis mode because of the dangers of bias in primary studies [12]; and latterly for animal toxicology experiments [3, 13]. Methodology for systematic reviews encompasses many topics, from prereview issues, such as methods of literature searching, through mainstream methods of description and synthesis, to the dissemination of the reviews findings. A brief resumé of some of the principal statistical methods available for formally combining several study estimates is given below. Methods specific to various outcome measures and data types are also available, and are described elsewhere [4, 5].

Search Methods

The procedural methodology for all stages of a systematic review has been clearly defined in two sets of

guidelines [1, 9]. A study protocol should be drawn up to establish the procedures used *before* a review is undertaken to avoid bias due to decisions made when the results of the review are known. Multiple methods [14], usually (multiple) electronic database searching complemented by handsearching of key journals by “explosion” of relevant entries in the reference lists of articles identified, are needed. Searches must be documented sufficiently to allow others to replicate them, and in due course update them. After the potentially relevant study reports have been identified, their relevance for inclusion in the systematic review/meta-analysis must be assessed. (For example, have they been performed in circumstances or contexts that allow their results to be regarded as valid in the target situation?) Not surprisingly, the results of systematic reviews may be sensitive to methods of searching and subsequent inclusion criteria [15].

Study Quality Assessment

It is common to assess the quality of the primary studies eligible for inclusion in the review. Exclusion of poor quality studies, or at least a downweighting of them in a formal pooled analysis, may appear sensible, but in fact there is little consensus on the most appropriate strategy. Although scales for scoring study quality exist for some study types [16], the validity of these instruments is not established [17], and guidelines for even the reporting of observational studies lag behind those for clinical trials [18]. Exploration of the influence of components of study design and conduct associated with study quality may be more appropriate. Sensitivity analysis of the impact on the subsequent decisions made concerning quality and of pooled estimates obtained as a result is essential; typically this sensitivity analysis will entail repeating the analysis on subsets of the data meeting increasingly stringent quality criteria.

Meta-Analysis and Evidence Synthesis

As indicated in step 6 above, in some cases a systematic review may be extended to a meta-analysis, allowing quantitative estimation of effects across similar studies. Whether this is advisable and justifiable will depend on the combinability of the results from the individual primary studies – for example, are they measuring similar outcomes or will it be a case of combining “apples and oranges”

[19] – and on the impact of potential problems including publication bias discussed below.

The most commonly reported pooled estimate is calculated as a weighted average of the results of primary studies, the weight given to each study depending on the precision of the estimate from it, so that studies with large precision are given more weight in the meta-analysis [4]. Other methods such as vote counting (tallying “positive” and “negative” studies relating to a particular issue) or combination of p values of effects in each study are occasionally necessary owing to limitations in information available from study reports, but neither method yields a pooled estimate of effect size [20]. Methods that do provide a pooled estimate of effect size are generally known as *inverse-variance weighted* meta-analysis models.

Heterogeneity

The variability in the true underlying effects between studies in a meta-analysis [21], known as *between-study heterogeneity*, is an important feature of any meta-analysis that must be considered and explored. Failure to allow for such heterogeneity may lead to inappropriate results. If heterogeneity is “substantial”, for example, as assessed with the I^2 criterion (the proportion of total variation in study estimates that is due to heterogeneity [21]), combining study estimates may make little sense; more generally, if present, heterogeneity should be accounted for as far as possible in the analysis, in terms of sources such as different populations or contexts studied in the primary studies, differing primary study methodologies, varying endpoint definitions, and study quality markers. Generally, no consensus has been reached concerning the best strategy for dealing with heterogeneity, although the use of mixed modeling section titled “Meta-Regression and Mixed Models”, is recommended.

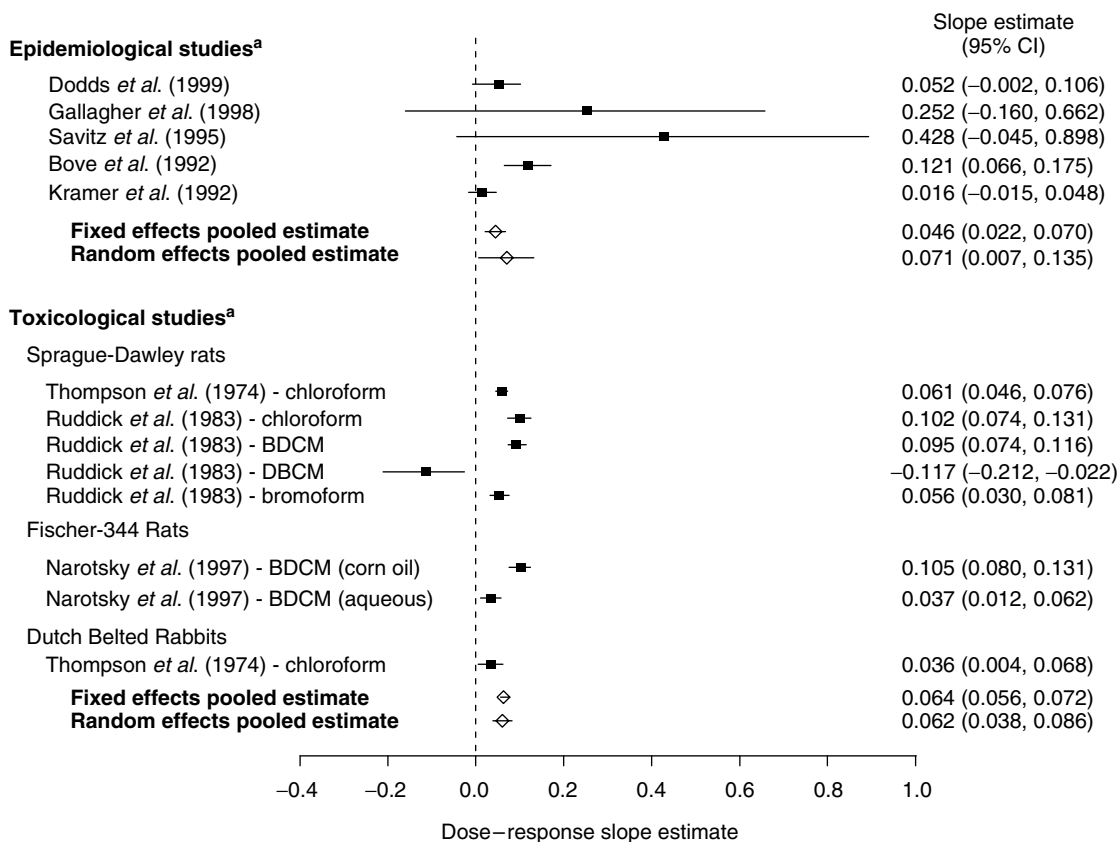
Fixed Effect Models

Fixed and random-effects models are commonly used in meta-analysis and choice of one over the other depends on the variability of the study-specific estimates assumed within the meta-analysis. The fixed effect method assumes all the studies are estimating the same, underlying treatment effect, and hence that no heterogeneity is present, which is often unrealistic. For continuous outcome measures

a weighted average of the primary study estimates (on a standardized scale) is calculated, giving a weighting to each that is proportional to the inverse of its variance. Several estimation methods have been proposed when the outcome is an odds ratio [22] as it often will be in risk assessment contexts. Although the results of the different methods will often be similar, in some circumstances serious differences can arise [23].

Random Effect Models

A random effect model [24], which can incorporate some aspects of heterogeneity when pooling estimates, is a popular alternative to the fixed effect method. Such models assume that the true effect size values for each study are not identical, but are drawn from an underlying distribution (usually assumed to be a normal distribution) of effect sizes on an appropriate scale, and that it is the mean of this distribution that is of interest. As a result of incorporating extra variation (due to heterogeneity present between studies), the confidence interval obtained for the pooled estimate is wider than that of the fixed effect method, so it is more conservative. The point estimate may also differ, as this method reduces the weighting of the larger studies and increases the relative weight given to the smaller studies, compared to the fixed effect model. These points are illustrated in the forest plots in Figure 1, which show dose–response estimates from primary studies of exposure to trihalomethanes and low birth weight with within-discipline fixed and random-effects meta-analysis estimates, which were subsequently synthesized using more complex models [25]. The uncertainty involved in estimating the between-study variance needs to be incorporated as well [26]; this widens the resulting confidence interval further. Random effects models have been criticized on grounds that unrealistic or unjustified distributional assumptions may have to be made in implementing them [27]. Neither fixed nor random effect analysis can be considered ideal, and the choice between them remains a source of controversy [28]. As noted below, combination of evidence of diverse types in risk assessment may in any case require *generalized synthesis of evidence* methods, extending conventional meta-analysis techniques.



^a The details of all the references cited in this figure are available in [25].

Figure 1 Study-specific dose-response estimates and 95% confidence intervals, with fixed and random effects meta-analysis estimates and 95% confidence intervals

Meta-Regression and Mixed Models

Although random effect models allow for between-study heterogeneity, if all or part of this variation is systematic, it may be possible to extend the model to include variables that explain it. Such analyses should usually be treated as exploratory, since associations between characteristics and the outcome can occur purely by chance or may be due to the presence of confounding factors. If study level covariates are included in fixed and random effect models, meta-regression and mixed models, respectively, result [29]. The mixed-modeling approach may be recommended in instances where covariates explain part of the heterogeneity, but a random effect term is required to “accommodate” the remainder [30].

Publication and Related Biases

Publication biases arise because statistically significant, interesting or “welcome” results are potentially more likely to be submitted, published or published quickly, and published in English. There is no consensus [31] on how best to proceed when publication bias is suspected, and methodological research in this area is active at present. “Funnel” plots of some measure of sample size *versus* effect size are often inspected for asymmetry, which may arise as a result of publication bias, but may also been seen for other reasons [32]. More formal tests or investigations are also performed [33, 34]. Even if a study is published, selection biases may operate in respect of the particular endpoints and covariate effects reported [35]. This is arguably more likely

in observational epidemiological studies than when the primary studies are experimental, and perhaps also in risk assessment contexts in which government, non-governmental agency, and industry data are not always published or made widely available.

Using Data on Individual Subjects

An alternative approach to combining aggregated study results is to gather individual subject data from the study investigators (if they are willing to provide it), and carry out a meta-analysis on data at the individual subject or animal level [36]. This is particularly useful for the synthesis of certain types of study designs such as survival or repeated measures studies, and allows detailed data checking, and the appropriateness of the analyses to be ensured. Such individual subject level meta-analyses are, however, usually both very time consuming and costly, so they are comparatively rarely carried out [37].

Applications of Systematic Review and Meta-Analysis in Risk Assessment

Methods for human health risk assessment of environmental exposures (*see **Environmental Health Risk Assessment***) are frequently based on conventional narrative reviews of usually diverse evidence (including both human and animal data). When limits for human exposure to these chemicals need to be determined, the limits may be based on the results from animal experiments, the dose level observed to cause no or little adverse effects in animals being divided by a number of safety or uncertainty factors in obtaining a limit for human exposure [38, 39]. More complex models based on biological processes are also used in risk assessment [40] but may require evidence that is rarely available. Thus current strategies for reviewing and incorporating evidence in risk assessments are often not systematic and transparent (with the prevalence results in the gray – not formally published – literature being a special problem in this area), and may lack some degree of quantification [41]. The nature and degree of uncertainty that exist in the estimation of these exposure limits need to be clear to both setters and users of the limits.

A review of a number of comparable risk assessments (of lung cancer and exposure to chromium

(*see **Hexavalent Chromium***)) comments that different risk assessments often result in different exposure levels being recommended because of differences in the evidence base, the methods used, and the assumptions made, the basis for which is not always clear [41]. For example, if a choice of pivotal study has been made, it may well be important to ask how sensitive the conclusions of the risk assessment are to that choice. Further, use of just one study as the basis for calculation (of an exposure limit, for example) may be inefficient when there is evidence also available from other relevant studies.

Examples of Systematic Review and Meta-Analysis in Risk Assessment

Although the use of meta-analysis in risk assessment was considered some time ago [42], adoption of the technique has been limited to date. Examples of use of systematic review and meta-analysis methods in risk assessment contexts include a meta-analysis of epidemiological evidence in a risk assessment of dimethoate [43], modeling outcome severity in terms of tetrachloroethylene exposure concentrations and durations and other covariates [44], and synthesis of evidence from more than one species exposed to hydrogen sulfide [45].

Other approaches relevant to the quantitative synthesis of evidence for a risk assessment have been developed. These methods include hierarchical Bayesian models for combination of evidence (on dose–response relationships) in several species [25] related to generalized synthesis methods considered below, direct modeling of the relevance of cross-species evidence in a Bayesian framework [46], and other modeling approaches [47]. Bayesian hierarchical models have also been used to combine estimates of potency from mutagenesis arrays of toxic environmental agents [48].

However, although examples of use of meta-analytic techniques can be found, there is little evidence to date of the widespread use of systematic review methods in the risk assessment process even though it could be argued that systematic review methods are more important than, or at least more widely applicable than, the quantitative synthesis. Further, meta-analyses not preceded by systematic review are in danger of providing quantitative estimate of risk, with associated levels of uncertainty, on the basis of incomplete data sets, with problems

analogous to those arising from publication bias (in which study reports would not be found even if a search had been performed).

We now consider in turn some characteristic features of applications of systematic review, meta-analysis, and related approaches in contexts characteristic of risk assessment. In particular, use of the techniques to summarize the results of *animal experiments* is in its infancy. When the studies to be summarized and combined are *observational*, as in the case of epidemiological studies, some further considerations arise from the potential for bias in such studies. Most risk assessments will require combination of evidence of diverse types or from diverse study types – for example, data on dose–response relationships from animal toxicology experiments and from observational epidemiological studies in humans. This may require *generalized synthesis of evidence* methods, extending conventional meta-analysis techniques, and often using *Bayesian methods*; in turn these may be developed into *comprehensive decision models* [49].

Animal Studies

Despite calls to make the risk assessment process more systematic and transparent [41, 50], the uptake for review and summary of animal experiments is still low [3, 13], with a few notable exceptions including articles in the radiation biology field [51] and on the genotoxicity of pesticides [52] demonstrating explicit description and specification of literature search terms and procedures, and of inclusion and exclusion criteria. Furthermore there is considerable room for improvement in the quality of systematic reviews and meta-analyses of animal experiments [3, 13].

Observational Epidemiological Studies

Since observational studies, including almost all epidemiological studies of the risks of disease associated with exposure in humans, are prone to a greater range and potential degree of bias than good experimental studies (toxicological experiments in animals or randomized clinical trials in humans), caution is needed in combining their results in meta-analyses [53]. Otherwise misleading precision may be achieved through pooling of primary studies suffering from substantial (and varied) biases, so that including

studies with poor designs may weaken the analysis by increasing bias [12, 54, 55]. This issue may be partially addressed by conducting sensitivity analyses under various credibility assumptions [56]. A set of guidelines for meta-analysis applied to environmental health issues [10] is more broadly applicable to observational studies in other similar areas and another report considers meta-analysis of epidemiological studies in risk assessment [57].

Generalized Synthesis of Evidence and Bayesian Methods

The Confidence Profile approach [58] was a forerunner to other cross-design synthesis methods [59], which allowed adjustments in the analysis to be made in principle for different study characteristics and biases. Development of synthesis approaches incorporating direct modeling of biases is currently an active area of research [60]. Methods for combination of studies with disparate designs into a single synthesis have often used Bayesian hierarchical modeling approaches [25, 61]. The hierarchical nature of the model specifically allows quantitatively, for both within and between design sources of heterogeneity, whilst the Bayesian approach can accommodate *a priori* beliefs regarding qualitative differences between the various sources of evidence, and levels of between-study heterogeneity [62]. Prior distributions may represent subjective beliefs elicited from experts (with explicit acknowledgement that different individuals may hold different *a priori* beliefs), or other data-based evidence, which though pertinent to the issue in question is not of a form that can be directly incorporated. This model can be viewed as an extension of the standard random-effects model described above, but with an extra level of variation to allow for variability in effect sizes between different sources. Bayesian approaches also allow probability statements to be made directly regarding quantities of interest, and predictive statements to be made, conditional on the current state of knowledge. They lead naturally into a decision analysis framework, which can also take into account utilities when making health care or policy decisions [63] and underlie many probabilistic risk assessment models [64].

Conclusions

Systematic review and meta-analysis methods have useful contributions to make to current risk assessment approaches. The use of systematic review methods means that clear criteria for searching for and identification of relevant studies can be defined and reported, allowing ease in updating and repeating the search process. These methods also provide a structured and systematic approach to the review of diverse human and animal evidence, maximizing efficient use of already available evidence [65], and facilitating identification of “gaps” in the knowledge base – perhaps that more epidemiological studies are needed to help answer the question of interest – the strategy for filling which could potentially be formalized by application of value-of-information methods [66].

Using meta-analysis methods in risk assessment means that *quantitative* summaries of evidence can be obtained, making it possible to estimate the amount by which risk is increased in those exposed to the substance compared to those not exposed, for example, rather than just being able to say that evidence of an increase in risk exists. A meta-analytical framework allows investigation of between-study heterogeneity and of publication bias, which may in turn provide information and insights not obtainable through reliance on just one (pivotal) study, and which facilitates sensitivity analyses [56] to help assess the impact of assumptions made in the procedures. In meta-analysis of observational studies in particular, where a variety of biases, some unrecognized or unmeasured, may operate in the primary studies, meta-analysis can prove to be an invaluable exploratory tool. The potential increase in power of a meta-analysis compared to that of an individual study makes it especially appealing for animal toxicology experiments where it may help in pursuing the 3Rs of animal research [67] by *reducing* the number of animals used, since data from individual studies are efficiently combined. An understanding of sources of bias in the individual studies may result from a meta-analysis and lead to improvements in the quality of conducting and reporting human studies and animal experiments [68]. More generally, incorporation of meta-analytical results into comprehensive decision models represents a valuable direction of development of probabilistic risk assessment modeling approaches.

References

- [1] Higgins, J.P.T. & Green, S. (eds) (2007). *Cochrane Handbook for Systematic Reviews of Interventions*, <http://www.cochrane.org/resources/handbook/hbook.htm> (accessed Nov 2007).
- [2] Petticrew, M. (2001). Systematic reviews from astronomy to zoology: myths and misconceptions, *British Medical Journal* **322**, 98–101.
- [3] Peters, J.L., Sutton, A.J., Jones, D.R., Rushton, L. & Abrams, K.R. (2006). A systematic review of systematic reviews and meta-analyses of animal experiments with guidelines for reporting, *Journal of Environmental Science and Health B* **41**, 1245–1258.
- [4] Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A. & Song, F. (2000). *Methods for Meta Analysis in Medical Research*, John Wiley & Sons, Chichester.
- [5] Egger, M., Smith, G.D. & Altman, D.G. (2001). *Systematic Reviews in Health Care: Meta-Analysis in Context*, British Medical Journal, London.
- [6] Weed, D.L. (2005). Weight of evidence: a review of concept and methods, *Risk Analysis* **25**, 1545–1557.
- [7] Blettner, M., Sauerbrei, W., Shlehofer, B., Scheuchengflug, T. & Friedenreich, C. (1999). Traditional reviews, meta-analyses and pooled analyses in epidemiology, *International Journal of Epidemiology* **28**, 1–9.
- [8] Biddle, S.J.H. (2006). Research synthesis in sport and exercise psychology: chaos in the brickyard revisited, *European Journal of Sport Science* **6**, 97–102.
- [9] Centre for Reviews and Dissemination (2001). *Undertaking Systematic Reviews of Research on Effectiveness*, Report CRD4, University of York, at <http://www.york.ac.uk/inst/crd/report4.htm> (accessed Aug 2007).
- [10] Blair, A., Burg, J., Foran, J., Gibb, H., Greenland, S., Morris, R., Blair, A., Burg, J., Foran, J., Gibb, H., Greenland, S., Morris, R., Raabe, G., Savitz, D., Teta, J., Wartenberg, D., Wong, O., Zimmerman, R. & ISLI Risk Science Institute (1995). Guidelines for application of meta-analysis in environmental epidemiology, *Regulatory Toxicology and Pharmacology* **22**, 189–197.
- [11] Stroup, D.F., Berlin, J.A., Morton, S.C., Stroup, D.F., Berlin, J.A., Morton, S.C., Olkin, I., Williamson, G.D., Rennie, D., Moher, D., Becker, B.J., Sipe, T.A. & Thacker, S.B. (2000). Meta-analysis of observational studies in epidemiology. A proposal for reporting, *Journal of the American Medical Association* **283**, 2008–2012.
- [12] Egger, M., Schneider, M. & Davey Smith, G. (1998). Meta-analysis: spurious precision? Meta-analysis of observational studies, *British Medical Journal* **316**, 140–144.
- [13] Mignini, L.E. & Khan, K.S. (2006). Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research, *BMC Medical Research Methodology* **6**, 10.
- [14] Dickersin, K., Scherer, R. & Lefebvre, C. (1994). Systematic reviews – identifying relevant studies

- for systematic reviews, *British Medical Journal* **309**, 1286–1291.
- [15] Cook, D.J., Reeve, B.K., Guyatt, G.H., Heyland, D.K., Griffith, L.E., Buckingham, L., Cook, D.J., Reeve, B.K., Guyatt, G.H., Heyland, D.K., Griffith, L.E., Buckingham, L. & Tryba, M. (1996). Stress-ulcer prophylaxis in critically ill patients – resolving discordant meta-analyses, *Journal of the American Medical Association* **275**, 308–314.
 - [16] Moher, D., Jadad, A.R., Nichol, G., Penman, M., Tugwell, P. & Walsh, S. (1995). Assessing the quality of randomized controlled trials – an annotated bibliography of scales and checklists, *Controlled Clinical Trials* **16**, 62–73.
 - [17] Jüni, P., Witschi, A., Bloch, R. & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis, *Journal of the American Medical Association* **282**, 1054–1060.
 - [18] von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gøtzsche, P.C. & Vandenbroucke, J.P. for the STROBE Initiative. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies, *Annals of Internal Medicine* **147**, 573–577.
 - [19] Myers, J.E. & Thompson, M.L. (1998). Meta-analysis and occupational epidemiology, *Occupational Medicine* **48**, 99–101.
 - [20] Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*, Academic Press, London.
 - [21] Higgins, J.P.T. & Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis, *Statistics in Medicine* **21**, 1539–1558.
 - [22] Emerson, J.D. (1994). Combining estimates of the odds ratio: the state of the art, *Statistical Methods for Medical Research* **3**, 157–178.
 - [23] Greenland, S. & Salvan, A. (1990). Bias in the one-step method for pooling study results, *Statistics in Medicine* **9**, 247–252.
 - [24] DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials, *Controlled Clinical Trials* **7**, 177–188.
 - [25] Peters, J.L., Rushton, L., Sutton, A.J., Jones, D.R., Abrams, K. & Mugglestone, M.A. (2005). Bayesian methods for the cross-design synthesis of epidemiological and toxicological evidence, *Journal of the Royal Statistical Society Series C* **54**, 159–172.
 - [26] Biggerstaff, B.J. & Tweedie, R.L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis, *Statistics in Medicine* **16**, 753–768.
 - [27] Peto, R. (1987). Why do we need systematic overviews of randomised trials? *Statistics in Medicine* **6**, 233–240.
 - [28] Thompson, S.G. (1993). Controversies in meta-analysis: the case of the trials of serum cholesterol reduction, *Statistical Methods in Medical Research* **2**, 173–192.
 - [29] Cooper, H. & Hedges, L.V. (eds) (1994). *The Handbook of Research Synthesis*, Russell Sage Foundation, New York.
 - [30] Thompson, S.G. & Higgins, J.P.T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* **21**, 1559–1573.
 - [31] Lau, J., Ioannidis, J.P.A., Terrin, N., Schmid, C.H. & Olkin, I. (2006). The case of the misleading funnel plot, *British Medical Journal* **333**, 597–600.
 - [32] Rothstein, H., Borenstein, M. & Sutton, A.J. (eds) (2004). *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, John Wiley & Sons, Chichester.
 - [33] Egger, M., Smith, G.D., Schneider, M. & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test, *British Medical Journal* **315**, 629–634.
 - [34] Duval, S. & Tweedie, R. (2000). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis, *Biometrics* **56**, 455–463.
 - [35] Williamson, P.R. & Gamble, C. (2005). Identification and impact of outcome selection bias in meta-analysis, *Statistics in Medicine* **24**, 1547–1561.
 - [36] Tobias, A.M., Saez, M. & Kogevinas, M. (2004). Meta-analysis of results and individual patient data in epidemiological studies, *Journal of Modern Applied Statistical Methods* **3**, 176–185.
 - [37] Stewart, L.A. & Tierney, J.F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data, *Evaluation and the Health Professions* **25**, 76–97.
 - [38] Edler, L., Kopp-Schneider, A. & Heinzl, H. (2005). Dose-response modelling, in *Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment*, L. Edler & C.P. Kitsos, eds, John Wiley & Sons, Chichester.
 - [39] World Health Organisation (1994). *Assessing Human Health Risks of Chemicals: Derivation of Guidance Values for Health-Based Exposure Limits*, Geneva.
 - [40] Watanabe, K.H. (2005). Modeling exposure and target organ concentrations, in *Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment*, L. Edler & C.P. Kitsos, eds, John Wiley & Sons, Chichester, pp. 115–124.
 - [41] Goldbohm, R.A., Tielemans, E.L.J.P., Heederik, D., Rubingh, C.M., Dekkers, S., Willems, M.I., Goldbohm, R.A., Tielemans, E.L., Heederik, D., Rubingh, C.M., Dekkers, S., Willems, M.I. & Dinant Kroese, E. (2006). Risk estimation for carcinogens based on epidemiological data: a structured approach, illustrated by an example on chromium, *Regulatory Toxicology and Pharmacology* **44**, 294–310.
 - [42] Putzrath, R.M. & Ginevan, M.E. (1991). Meta-analysis: methods for combining data to improve quantitative risk assessment, *Regulatory Toxicology and Pharmacology* **14**, 178–188.
 - [43] Reiss, R. & Gaylor, D. (2005). Use of benchmark dose and meta-analysis to determine the most sensitive endpoint for risk assessment for dimethoate, *Regulatory Toxicology and Pharmacology* **43**, 55–65.

- [44] Guth, D.J., Carroll, R.J., Simpson, D.G. & Zhou, H. (1997). Categorical regression analysis of acute exposure to tetrachloroethylene, *Risk Analysis* **17**, 321–332.
- [45] Brown, K.G. & Strickland, J.A. (2003). Utilizing data from multiple studies (meta-analysis) to determine effective dose-duration levels. Example: rats and mice exposed to hydrogen sulfide, *Regulatory Toxicology and Pharmacology* **37**, 305–317.
- [46] DuMouchel, W.H. & Harris, J.E. (1983). Bayes methods for combining the results of cancer studies in humans and other species, *Journal of the American Statistical Association* **78**, 293–308.
- [47] Cox, L.H. & Piegorsch, W.W. (1994). Combining environmental information: environmetric research in ecological monitoring, epidemiology, toxicology, and environmental data reporting, Technical Report 12, National Institute of Statistical Science, North Carolina.
- [48] Simmonds, S.J., Piegorsch, W.W., Nitcheva, D. & Zieger, E. (2003). Combining environmental information via hierarchical modelling: an example using mutagenic potencies, *Environmetrics* **14**, 159–168.
- [49] Parmigiani, G. (2002). *Modeling in Medical Decision Making*, John Wiley & Sons, Chichester.
- [50] Guzelian, P.S., Victoroff, M.S., Halmes, N.C., James, R.C. & Guzelian, C.P. (2005). Evidence-based toxicology: a comprehensive framework for causation, *Human & Experimental Toxicology* **24**, 161–201.
- [51] Juutilainen, J., Kumlin, T. & Naarala, J. (2006). Do extremely low frequency magnetic fields enhance the effects of environmental carcinogens? A meta-analysis of experimental studies, *International Journal of Radiation Biology* **82**, 1–12.
- [52] Bull, S., Fletcher, K., Boobis, A.R. & Battershill, J.M. (2006). Evidence for genotoxicity of pesticides in pesticide applicators: a review, *Mutagenesis* **21**, 93–103.
- [53] Spitzer, W.O. (1991). Meta-meta-analysis: unanswered questions about aggregating data, *Journal of Clinical Epidemiology* **44**, 103–107.
- [54] Dickersin, K. (2002). Systematic reviews in epidemiology: why are we so far behind? *International Journal of Epidemiology* **31**, 6–12.
- [55] Shapiro, S. (1994). Meta-analysis/Shmeta-analysis, *American Journal of Epidemiology* **140**, 771–778.
- [56] Greenland, S. (2005). Multiple-bias modelling for the analysis of observational data, *Journal of the Royal Statistical Society A* **168**, 267–306.
- [57] Federal Focus Inc. (1999). *Epidemiology in Hazard and Risk Assessment*, at <http://www.fedfocus.org/final-report1.html> (accessed Aug 2007).
- [58] Eddy, D.M., Hasselblad, V. & Shachter, R. (1992). *Meta-analysis by the Confidence Profile Method*, Academic Press, San Diego.
- [59] General Accounting Office (1992). *Cross Design Synthesis: A New Strategy for Medical Effectiveness Research*, General Accounting Office, Washington, DC.
- [60] Wolpert, R.L. & Mengersen, K. (2004). Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke, *Statistical Science* **19**, 450–471.
- [61] Prevost, T.C., Abrams, K.R. & Jones, D.R. (2000). Hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening, *Statistics in Medicine* **19**, 3359–3376.
- [62] Spiegelhalter, D.J., Abrams, K.R. & Myles, J.P. (2003). *Bayesian Approaches to Clinical Trials and Health-care Evaluation*, John Wiley & Sons, Chichester.
- [63] Ades, A.E. & Sutton, A.J. (2006). Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches, *Journal of the Royal Statistical Society A* **169**, 5–35.
- [64] Bedford, T.M. & Cooke, R.M. (2001). *Probabilistic Risk Analysis: Foundations and Method*, Cambridge University Press, Cambridge.
- [65] Chalmers, I. (2005). The scandalous failure of science to accumulate evidence scientifically, *Clinical Trials* **2**, 229–231.
- [66] Yokota, F. & Thompson, K.M. (2004). Value of information analysis in environmental health risk management decisions: past, present, and future, *Risk Analysis* **24**, 635–650.
- [67] National Centre for the Replacement, Refinement and Reduction of Animals in Research, at <http://www.nc3rs.org.uk/> (accessed Aug 2007).
- [68] Macleod, M.R., Ebrahim, S. & Roberts, I. (2005). Surveying the literature from animal experiments: systematic reviews and meta-analyses are important contributions, *British Medical Journal* **331**, 110.

Related Articles

Bayesian Statistics in Quantitative Risk Assessment

Decision Analysis

DAVID R. JONES, ALEX SUTTON AND JAIME PETERS

Methylmercury

Biogeochemistry

Mercury (Hg) is a heavy metal that is dispersed into the environment as a result of both natural processes and human activities. The natural sources include volcanic emissions, the weathering of rock containing Hg ore, and undersea vents. The major human-related (anthropogenic) pathway of exposure is the combustion of carbon-based fuels, particularly by coal-fired power plants. The use of Hg as an amalgamator in gold mining operations has resulted in contamination in certain locales. Other anthropogenic pathways include municipal, medical, and hazardous waste incineration, although, in the United States, the importance of these pathways has diminished as a result of more stringent regulations regarding the amount of Hg allowed in the waste stream. Because of the complex biogeochemistry of Hg fate and long-range transport in the atmosphere [1], it is difficult to quantify the relative contributions of these different sources and pathways to global Hg pollution. However, anthropogenic processes are thought to be responsible for at least 50% of the total yearly dispersion of Hg into the environment [2–4].

Hg is present in the environment in metallic, inorganic, and organic forms. Interconversion between forms can occur. Great attention has been paid in recent years to human exposures to one particular organic form, methylmercury (MeHg). The primary route of human exposure to MeHg is the consumption of food substances taken from contaminated water (see **Water Pollution Risk**). MeHg results when elemental and inorganic Hg deposited in waterbodies is biotransformed through the process of methylation by phytoplanktons (marine environments) and sulfate-reducing bacteria in anoxic or suboxic sediments and waters (freshwater environments) [5]. In heterotrophic organisms, MeHg bioaccumulates and biomagnifies as it ascends the aquatic food chain. As a result, the greatest tissue concentrations are generally found in animals at the highest trophic levels, i.e., long-lived predatory species. In the marine environment, these include certain sharks, whales, swordfish, and tuna. In the freshwater environment, they include bass, pike, walleye, and pickerel. MeHg levels can also be high in animals that feed on marine life, such as polar bears, otters, mink, and piscivorous

birds (e.g., loons, eagles, osprey, ducks, herons, and kingfishers).

Toxicokinetics

MeHg is not lipophilic (fat soluble). Therefore, it is present in the largest concentrations in muscle tissues rather than in fat or oil. MeHg concentrations are unaffected by preparation or cooking methods. Because of biomagnification, the concentration of MeHg in fish tissues can be orders of magnitude greater than the concentration of Hg in surrounding water. Approximately 95% of ingested MeHg is absorbed from the gastrointestinal tract. The half-life of MeHg in blood in humans is estimated to be 50 days, and the whole-body half-life to be 70–80 days. The residence time of Hg in the brain, into which MeHg is actively transported on the large neutral amino acid carrier [6], however, appears to be considerably longer [7]. The primary route of excretion of MeHg is by means of demethylation by organisms in the gut and excretion of the resulting inorganic form in feces. Another route of elimination is hair. The Hg concentration in hair is approximately 250 times greater than the concentration in blood, although this ratio can vary widely depending on level of exposure, age, and other factors [8]. Because at least 80% of the Hg in hair is in the form of MeHg, total hair Hg is frequently used in epidemiological studies as an exposure biomarker for MeHg, and is particularly useful in studies of populations of high fish consumers. MeHg is actively transported across the placenta [9] and maternal hair Hg level at delivery correlates reasonably well with the level of Hg in the infant brain ($r = 0.6–0.8$, depending on brain region) [10].

Human Health Effects

The devastating effects of high-dose exposures to MeHg on human health were first recognized following the decades-long poisoning that occurred in the population living near Minamata Bay in southern Japan as a result of industrial discharge of effluent containing methylmercury chloride. Women who consumed MeHg-contaminated fish gave birth to children who suffered from what came to be called *congenital Minamata disease (CMD)*. The signs of CMD included growth disturbances, primitive reflexes,

movement and coordination disorders (cerebellar ataxia, chorea, athetosis, and dysarthria), sensory impairments, cerebral palsy, and mental retardation. Because of the delay in identifying MeHg as the cause, the critical MeHg dose required to cause CMD could not be determined with certainty. Other episodes of mass poisoning were recognized in Niigata, Japan, in the 1960s, and in Iraq in the 1970s. In the latter episode, seed grain treated with a fungicide containing MeHg was consumed rather than planted. The adverse effects in children who were exposed to MeHg prenatally were similar to those observed in Minamata. In both the Japanese and Iraqi episodes, it was noted anecdotally that mothers of many affected children were asymptomatic or suffered only mild difficulties such as transient paresthesias, particularly in the extremities and perioral region. This observation suggested that pregnant women are the critical population subgroup with regard to MeHg exposure. In its 1990 risk assessment, the World Health Organization (WHO) identified the Iraqi episode as the critical study and concluded that, "A prudent interpretation . . . implies that a 5% risk (of developmental delay in offspring) may be associated with a peak Hg level of 10–20 ppm (parts per million) in the maternal hair" [11, p. 103].

Examination of the pathological changes in the brains of individuals poisoned with MeHg toxicity in the Minamata episode provided clues to the explanation for the greater sensitivity of the fetal nervous system, compared with the adult nervous system. A striking age dependence was noted in the distribution of lesions in the brain. The injuries were highly focal in individuals who were exposed only as adults and were located primarily in the cerebellum (particularly granule cells), the calcarine fissure of the occipital cortex, and pre- and post-central gyrus. This pattern accords well with the typical clinical signs of adult MeHg poisoning, which include distal sensory disturbances, constriction of visual fields, dysarthria, tremor, and ataxia. In contrast, lesions were diffusely distributed throughout the brain [12] among individuals exposed prenatally. This is because MeHg disrupts neuronal cell proliferation and migration, resulting in structural abnormalities that include reduced cell densities, islands of heterotopic neurons (i.e., incorrectly located), glial proliferation, incomplete myelination, and disturbances in brain cytoarchitecture. Among the possible mechanisms of MeHg neurotoxicity are disorganization

of the microtubular system, oxidative stress injury, mitochondrial dysfunction, and inhibition in protein and macromolecular synthesis [13].

Recognizing the devastating effects of high-dose exposure to MeHg on the fetus, investigators were led, beginning in the 1980s, to ask whether milder neurological effects were associated with the lower in utero exposures to MeHg that are more typical within the general population of seafood consumers. To evaluate this hypothesis, several longitudinal prospective studies were undertaken in populations of frequent fish consumers. The most important of these studies were conducted in New Zealand, the Faroe Islands (located in the Northern Atlantic Ocean), and the Seychelles Islands (located in the western Indian Ocean). The women enrolled in the Seychelles Islands study, for example, reported eating an average of 12 "fish meals" per week [14]. Approximately half of the Faroese women consumed "fish dinners" three or more times per week, and 80% consumed one or more "pilot whale dinners" per month [15]. The New Zealand and Faroe Islands studies have generally been interpreted as demonstrating inverse associations between subclinical prenatal exposure to MeHg and children's neurodevelopment. In the Faroes Islands study, cord-blood Hg level was inversely associated with children's neuropsychological test scores at ages 7 and 14 years [16, 17]. In addition, higher Hg levels in children's hair at 14 years of age were associated with delayed responses on brain stem auditory evoked potentials [18]. In the New Zealand study, maternal hair Hg level greater than 10 ppm was associated with a doubling of the risk that a child would have an IQ score below 70 [19]. The Seychelles Islands study has generally been interpreted as providing little evidence of adverse effects from higher MeHg exposures [20, 21].

Approaches to Risk Assessment

The apparent discrepancies in the findings of the New Zealand, Seychelles Islands, and Faroe Islands studies have posed challenges to risk assessors attempting to establish intake limits for MeHg, and considerable effort has been invested in the search for possible explanations [7]. Recent integrative analyses suggest that although the findings of the three studies appear to differ if compared solely on the basis of *P* values associated with the dose–effect relationships (*see*

Dose–Response Analysis), the inverse slopes of the relationships are actually quite similar across studies [22]. Nevertheless, different regulatory agencies have made different choices regarding the critical study. The US Environmental Protection Agency (US EPA) relied primarily on the Faroe Islands study in deriving its RfD of $0.1 \mu\text{g Hg/kg body weight/day}$, although the findings of the New Zealand study and, to a lesser extent, the Seychelles Islands study provided supporting evidence [23]. The US Agency for Toxic Substances and Disease Registry used the Seychelles Islands study to derive its minimal risk level of $0.3 \mu\text{g Hg/kg body weight/day}$, expressing concern that some of the apparent adverse effects attributed to MeHg in the Faroe Islands study might be the result of residual confounding by coexposures to persistent organic contaminants present in pilot whale blubber [24]. The Joint Expert Committee on Food Additives and Contaminants (JECFA), a committee of the WHO and the Food and Agriculture Organization, did not consider the New Zealand study in its risk assessment because of concerns about the inordinate influence on the results of one observation, a maternal hair Hg level of 86 ppm, which was more than four times higher than the next highest concentration in this study sample. Therefore, the JECFA derived the provisional tolerable weekly intake of $1.6 \mu\text{g Hg/kg body weight/week}$ (or $0.23 \mu\text{g Hg/kg body weight/day}$) on the basis of the Faroes Islands and Seychelles Islands studies [25]. Another factor contributing to the discrepancies among the intake limits derived by regulatory groups is differences in the underlying assumptions, most notably in the uncertainty factors applied.

To illustrate a MeHg risk assessment, the approach used by the US EPA to derive the RfD is described. The RfD is defined as, "... an estimate of a daily exposure to the human population (including sensitive subgroups) that is likely to be without an appreciable risk of deleterious effects during a lifetime" [26]. To derive the RfD, the US EPA applied benchmark dose (BMD) modeling. In this approach, the dose, termed the *benchmark dose*, is identified at which the prevalence of a defined health abnormality exceeds the background prevalence by a specified amount (i.e., an excess risk). The lower bound of the 95% confidence interval for the BMD, referred to as the BMDL (*benchmark dose lower limit*), is identified. The BMDL is therefore the lowest Hg dose that is statistically consistent with the observed

excess risk. The US EPA considered a neuropsychological test score at the fifth percentile or below to be the critical adverse effect, and a doubling of such scores as the critical excess risk [27]. Based on models of the dose–effect associations relating prenatal MeHg exposure to the target excess risk, the US EPA identified a BMDL of $58 \mu\text{g l}^{-1}$ in cord blood (or 12 ppm of Hg in maternal hair). A total uncertainty factor of 10 was applied to the BMDL to take account of inter-individual toxicokinetic and toxicodynamic variability, resulting in a target cord blood Hg level of $5.8 \mu\text{l}^{-1}$. The next task was to calculate the MeHg intake (i.e., the RfD) that, over a lifetime, would not produce a cord blood Hg level greater than $5.8 \mu\text{l}^{-1}$. To do so, a one-compartment pharmacokinetic model was used, involving the following assumptions: elimination constant, 0.014 days^{-1} ; blood volume, 5 l; MeHg absorption, 0.95; the fraction of absorbed dose in the blood, 0.059; body weight, 67 kg; the ratio of cord blood Hg to maternal blood Hg, 1:1.

The US EPA recognized that the RfD is sensitive to the values assumed for these factors. In the absence of definitive evidence regarding the functional form of the association between Hg body burden and children's neurodevelopment, the EPA's calculations were based on a linear model. Additional analyses of the Faroe Islands study, however, suggested that a log linear model, which would result in a lower value for the RfD, provides a somewhat better fit to the data than does a linear model [28]. Recent analyses indicate that the best point estimate for the ratio of cord blood Hg to maternal blood Hg is 1.7, not 1.1, as the US EPA assumed, and that the ratio at the 95th percentile might be as large as 3.4 [29]. Use of this revised estimate of the ratio would produce a 35% reduction in the RfD. The use of point estimates for the factors considered in the derivation of the RfD ignores the fact that the values on such factors are likely to vary among individuals in the population [30]. The incorporation of information about the distributions of such factors would be important in deriving RfDs that are more protective for potentially susceptible subgroups [27]. Taking account of the inevitable error involved in measuring MeHg exposure, which was not done in the EPA analysis, would also result in a lower RfD [31]. In addition, the findings of some recent studies suggest that the BMDL assumed by the US EPA as the point of departure for calculating the RfD is too high. Hair Hg levels as low as 2 ppm

might be cardiotoxic in adults [32], and adverse neurodevelopmental effects have been observed at maternal hair Hg levels well below the BMDL of 12 ppm [33, 34]. Intake limits for MeHg will therefore need to be reconsidered periodically as new data provide more refined estimates for the assumptions that underlie the risk assessments.

A factor complicating MeHg risk assessment is that seafood, the primary route of human exposure to MeHg, also delivers beneficial nutrients such as omega-3 fatty acids, whose effects on the critical endpoints can be antagonistic to those of MeHg. In other words, the estimate of the association between a health endpoint and MeHg exposure incurred as the result of seafood consumption cannot be interpreted as a “pure” estimate of MeHg toxicity insofar as it reflects the balance between the beneficial effects of some fish constituents and the adverse effects of MeHg on the critical endpoint. The trade-off between the risks and benefits associated with seafood consumption has been demonstrated in studies of children’s neurodevelopment [33] and the progression of cardiovascular disease in adult men [35].

References

- [1] Fitzgerald, W.F. & Clarkson, T.W. (1991). Mercury and monomethylmercury: present and future concerns, *Environmental Health Perspectives* **96**, 159–166.
- [2] Nriagu, J.O. & Pacyna, J.M. (1988). Quantitative assessment of worldwide contamination of air, water and soils by trace metals, *Nature* **333**, 134–139.
- [3] U.S. Environmental Protection Agency (1997). *Mercury Report for Congress*, Office of Air Quality Planning and Standards and Office of Research and Development, Washington, DC, Vol. 1, Executive Summary, EPA-452/R-97-003.
- [4] United Nations Environmental Programme (2002). *Global Mercury Assessment*, Geneva.
- [5] Kraepiel, A.M.L., Keller, K., Chin, H.B., Malcolm, E.G. & Morel, F.M.M. (2003). Sources and variations of mercury in tuna, *Environmental Science and Technology* **37**, 5551–5558.
- [6] Kerper, L.E., Ballatori, N. & Clarkson, T.W. (1992). Methylmercury transport across the blood-brain barrier by an amino acid carrier, *American Journal of Physiology* **267**, R761–R765.
- [7] National Research Council (2000). *Toxicological Effects of Methylmercury*, National Academy Press, Washington, DC.
- [8] Budtz-Jorgensen, E., Grandjean, P., Jorgensen, P.J., Weihe, P. & Keiding, N. (2004). Association between mercury concentrations in blood and hair in methylmercury-exposed subjects at different ages, *Environmental Research* **95**, 385–393.
- [9] Bjornberg, K.A., Vahter, M., Berglund, B., Niklasson, B., Blennow, M. & Sandborgh-Englund, G. (2005). Transport of methylmercury and inorganic mercury to the fetus and breast-fed infant, *Environmental Health Perspectives* **113**, 1381–1385.
- [10] Cernichiari, E., Brewer, R., Myers, G.J., Marsh, D.O., Lapham, L.W., Cox, C., Shamlaye, C.F., Berlin, M., Davidson, P.W. & Clarkson, T.W. (1995). Monitoring methylmercury during pregnancy: maternal hair predicts fetal brain exposure, *Neurotoxicology* **16**, 705–710.
- [11] World Health Organization (1990). *Methylmercury: Environmental Health Criteria No. 101*, Geneva.
- [12] Choi, B.H. (1989). The effects of methylmercury on the developing brain, *Progress in Neurobiology* **32**, 447–470.
- [13] Verity, M.A. (1997). Pathogenesis of methyl mercury neurotoxicity, in *Mineral and Metal Neurotoxicology*, M. Yasui, M.J. Strong, K. Ota & M.A. Verity, eds, CRC Press, Boca Raton, pp. 159–167.
- [14] Shamlaye, C.F., Marsh, D.O., Myers, G.J., Cox, C., Davidson, P.W., Choisy, O., Cernichiari, E., Choi, A., Tanner, M.A. & Clarkson, T.W. (1995). The Seychelles Child Development Study on neurodevelopmental outcomes in children following in utero exposure to methylmercury from a maternal fish diet: background and demographics, *Neurotoxicology* **16**, 597–612.
- [15] Grandjean, P., Weihe, P., Jorgensen, P.J., Clarkson, T., Cernichiari, E. & Videro, T. (1992). Impact of maternal seafood diet on fetal exposure to mercury, selenium, and lead, *Archives of Environmental Health* **47**, 185–195.
- [16] Grandjean, P., Weihe, P., White, R.F., Debes, F., Araki, S., Yokoyama, K., Sorensen, N., Dahl, R. & Jorgensen, P.J. (1997). Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury, *Neurotoxicology and Teratology* **19**, 417–428.
- [17] Debes, F., Budtz-Jorgensen, E., Weihe, P., White, R.F. & Grandjean, P. (2006). Impact of prenatal methylmercury exposure on neurobehavioral function at age 14 years, *Neurotoxicology and Teratology* **28**, 363–375.
- [18] Murata, K., Weihe, P., Budtz-Jorgensen, E., Jorgensen, P.J. & Grandjean, P. (2004). Delayed brainstem auditory evoked potential latencies in 14-year-old children exposed to methylmercury, *Journal of Pediatrics* **144**, 177–183.
- [19] Kjellstrom, T., Kennedy, P., Wallis, S., Stewart, A., Friberg, L., Lind, B., Wutherspoon, N.T. & Mantell, C. (1989). Physical and mental development of children with prenatal exposure to mercury from fish, National Swedish Environmental Protection Board Report No. 3642, The National Swedish Environmental Protection Board, Solna.
- [20] Davidson, P.W., Myers, G.J., Cox, C., Axtell, C., Shamlaye, C., Sloane-Reeves, J., Cernichiari, E., Needham, L., Choi, A., Wang, Y., Berlin, M. & Clarkson, T.W.

- (1998). Effects of prenatal and postnatal methylmercury exposure from fish consumption on neurodevelopment, *Journal of the American Medical Association* **280**, 701–707.
- [21] Myers, G.J., Davidson, P.W., Cox, C., Shamlaye, C.F., Palumbo, D., Cernichiari, E., Sloane-Reeves, J., Wilding, G.E., Kost, J., Huang, L.S. & Clarkson, T.W. (2003). Prenatal methylmercury exposure from ocean fish consumption in the Seychelles child development study, *Lancet* **361**(9370), 1686–1692.
- [22] Axelrad, D.A., Bellinger, D.C., Ryan, L.M. & Woodruff, T.J. (2007). Dose-response relationship of prenatal mercury exposure and IQ: an integrative analysis of epidemiologic data, *Environmental Health Perspectives* **115**, 609–615.
- [23] U.S. Environmental Protection Agency (2001). *Risk assessment for methylmercury, Water Quality Criterion for the Protection of Human Health: Methylmercury*, EPA Office of Sciences and Technology, Office of Water, Washington, DC, Chapter 4.
- [24] U.S. Agency for Toxic Substances and Disease Registry (1999). *Toxicological Profile for Mercury*, U.S. Department of Health and Human Services, Public Health Service, Atlanta.
- [25] FAO/WHO Joint Expert Committee on Food Additives (2004). Evaluation of certain food additives and contaminants (Sixty-first report of the Joint FAO/WHO Expert Committee on Food Additives), WHO's Technical Report Series, No. 922, The International Programme On Chemical Safety, Geneva.
- [26] Rice, D.C., Schoeny, R. & Mahaffey, K. (2003). Methods and rationale for derivation of a reference dose for methylmercury by the U.S. EPA, *Risk Analysis* **23**, 107–115.
- [27] Rice, D.C. (2004). The US EPA reference dose for methylmercury: sources of uncertainty, *Environmental Research* **95**, 406–413.
- [28] Budtz-Jorgensen, E., Grandjean, P., Keiding, N., White, R.F. & Weihe, P. (2000). Benchmark dose calculations of methylmercury-associated neurobehavioral deficits, *Toxicology Letters* **112–113**, 193–199.
- [29] Stern, A.H. & Smith, A.E. (2003). An assessment of the cord blood:maternal blood methylmercury ratio: implications for risk assessment, *Environmental Health Perspectives* **111**, 1465–1470.
- [30] Canuel, R., de Grosbois, S.B., Atikesse, L., Lucotte, M., Arp, P., Ritchie, C., Mergler, D., Chan, H.M., Amyot, M. & Anderson, R. (2006). New evidence on variations of human body burdens of methylmercury from fish consumption, *Environmental Health Perspectives* **114**, 302–306.
- [31] Budtz-Jorgensen, E., Keiding, N. & Grandjean, P. (2004). Effects of exposure imprecision on estimation of the benchmark dose, *Risk Analysis* **24**, 1689–1696.
- [32] Stern, A.H. (2005). A review of the studies of the cardiovascular health effects of methylmercury with consideration of their suitability for risk assessment, *Environmental Research* **98**, 133–142.
- [33] Oken, E., Wright, R.O., Kleinman, K.P., Bellinger, D., Hu, H., Rich-Edwards, J.W., Gillman, M.W. (2005). Maternal fish consumption, hair mercury, and infant cognition in a US cohort, *Environmental Health Perspectives* **113**, 1376–1380.
- [34] Jedrychowski, W., Jankowski, J., Flak, E., Skarupa, A., Mroz, E., Sochacka-Tatara, E., Lisowska-Miszczuk, I., Szpanowska-Wohn, A., Rauh, V., Skolicki, Z., Kaim, I. & Perera, F. (2006). Effects of prenatal exposure to mercury on cognitive and psychomotor function in one-year-old infants: epidemiologic cohort study in Poland, *Annals of Epidemiology* **16**, 439–447.
- [35] Rissanen, T., Voutilainen, S., Nyyssonen, K., Lakka, T.A. & Salonen, J.T. (2000). Fish oil-derived fatty acids, docosahexenoic acid and docosapentaenoic acid, and the risk of acute coronary events-The Kuopio Ischaemic Heart Disease Risk Factor Study, *Circulation* **102**, 2677–2679.

Related Articles

Mercury/Methylmercury Risk

DAVID C. BELLINGER

Microarray Analysis

Microarray Analysis and Genomics

The fields of genetics, combinatorial chemistry, and statistics combine to produce a technology, known as *microarray analysis*, that can be used to quantitatively assess the risk of developing a disease for which the genetic expression has been identified. The technology is one of the cornerstones of the science of functional genomics and produces measurements of the expression levels of genes in a particular organism, generally in response to some condition(s) believed to stimulate the organism's genes. By comparing the expression levels of the full complement of genes from the organism with those from a "normal" individual, one can identify genes that are "differentially expressed" (expressed at different levels in the two genomes). To the extent that a gene has been identified as either causative or highly associated with a particular disease (e.g., *BRCA1* gene on chromosome 17 for suppression of early onset breast cancer tumor, or the *Rb* gene on chromosome 13 for suppression for retinoblastoma tumor), risk of disease can be assessed and mechanisms of causation can be elucidated. In some cases, the analysis may permit early intervention to prevent further disease (e.g., the identification of the absence of a gene may permit early diagnosis and treatment).

Microarray analysis can assist in the quantification of risk in developing disease by virtue of technology that permits measurements of gene expression. If increased gene expression can be associated with increased risk of disease, then accurate measurements of gene expression, and quantification of the uncertainty in these measurements, becomes essential. Thus, a complete understanding of the extent to which microarray analyses quantify disease risk requires clarification of the measurement process, which exposes possible sources of uncertainties in gene expression levels. Also, it is important to note that these measurements reflect levels of messenger ribonucleic acid (mRNA), not the levels of proteins, which the organism manufactures in response to measured elevated levels of mRNA.

Presently, two technologies, one using *complementary deoxyribonucleic acid (cDNA) slides* and the other using *oligonucleotide arrays*, are available for measuring gene expression levels. Both are based

on the biological concept of hybridization between matching nucleotides and both technologies contain multiple copies of single-stranded genes or gene fragments, called *probes*, linked to a substrate or surface for binding with expressed transcripts from target tissues.

Biological Basis and Measurements from Microarrays

The genetic code for an organism is contained in organized strings of four nucleotides (A: adenine, C: cytosine, G: guanine, T: thymine), arranged in triplets such that each triplet codes for one of 20 amino acids. (Multiple triplets may code for the same amino acid.) Strings of amino acids are called *peptides*. Peptides can act independently in a cell, or they can combine with other peptides to form complex proteins used by the organism for cell function. Genetic material known as *deoxyribonucleic acid (DNA)* is arranged in a double-stranded helical structure, with complementary base pairs on either side of the helix. In response to a stimulus to produce a protein, the coding genes in the DNA are transcribed into mRNA for translation into peptides. To test for gene expression, cells from tissues are harvested, and the mRNA is reverse transcribed into its more stable form, cDNA, split into smaller strands, and labeled with a chemical that can be detected by an instrument. For quantitative measurement, the mixture of cDNA strands is placed onto the slide or chip containing the gene probes, and the strands of nucleotides in the target sample are allowed to bind (hybridize) to their matching partners. Spots on the slide or chip where hybridization has occurred indicate genes that are present in larger quantities and may have been expressed in response to the stimulus. With either technology, the reported intensity level at a particular location on the slide or chip is a summary of fluorescence measurements detected in a series of pixels that comprise the spot on the slide.

The two technologies for measuring gene expression differ in the process that is used to manufacture the probes on the slide or chip. In cDNA slides, the probes are obtained from actual cellular DNA that has been spliced using restriction enzymes to cleave the DNA at specific locations corresponding to specific genes. Once isolated, the DNA fragments undergo a series of complex processes that amplify and multiply their numbers, resulting in a gene "library" for

a given laboratory, which is then used to bond to the substrate on slides. In oligonucleotide arrays, the gene fragments consist of “known” strings of manufactured nucleotides that are placed on the chip using photolithography. For both the chip and the slide, as many as 50 000 or more mRNA transcript probes are placed in an array of rows and columns, hence the term *microarray*.

The technologies also differ in the experimental protocol that yields the data on gene expression levels. In cDNA experiments, a target sample contains a mixture of two types of cells, control and experimental, whose mRNA is reverse transcribed into the more stable cDNA and then labeled with two different fluorophores: the control cells are often labeled with cyanine 3, or Cy3 (green dye), and the experimental cells (e.g., cells that have been subjected to some sort of treatment, such as stress, heat, radiation, or chemicals), or those known to originate from a disease tissue, are often labeled with cyanine 5, or Cy5 (red dye). When mRNA concentration is high in these cells, their cDNA will bind to their corresponding probes on the spotted cDNA slide; an optical detector in a laser scanner will measure the fluorescence at wavelengths corresponding to the green and red dyes (532 and 635 nm, respectively). Good experimental design will interchange the dyes in a separate experiment to account for imbalances in the signal intensities from the two types of fluorophores and the expected degradation in the cDNA samples between the first scan at 532 nm (green) and second scan at 635 nm (red). The ratio of the relative abundance of red and green dyes at these two wavelengths on a certain spot indicates relative mRNA concentration between the experimental and control genes. Thus, the gene expression levels in the target cells can be directly compared with those from the control cells.

Oligonucleotide arrays circumvent the possible inaccuracies that can arise in the preparation of a gene “library” and the control and experimental samples for spotted array slides, by using 11–20 predefined and prefabricated sequences of 25–60 nucleotides for each gene. Rather than circular-shaped spots, these probes are deposited onto the chip in square-shaped cells. The probe cells measure 24×24 or $50 \times 50 \mu\text{m}^2$ and are divided in 8×8 pixels; cells from the target sample, labeled again with a fluorophore, will hybridize to those squares on the chip that correspond to the complementary strands of the target sample’s single-stranded cDNA. For these experiments, the

target sample contains only one type of cell (e.g., treatment or control); the assessment of expression is in comparison to the expression level on an adjacent probe, which is exactly the same as the gene probe except for the middle nucleotide (e.g., 13th out of 25 nucleotides). This “mismatch” (MM) for the “perfect match” (PM) sequence is only rough, since a target sample with elevated mRNA concentration for a certain gene may sufficiently hybridize to both the PM and MM probes. However, the results are believed to be less variable, since the probes on the chips are manufactured in more carefully controlled concentrations. Gene expression levels are measured again by a laser scanner that detects the optical energy in the pixels at the various probes (PM and MM) on the chip.

Statistical Analysis of Microarray Measurements

The analysis of the data (fluorescence levels at the various locations on the slide or chip) depends upon the technology. For cDNA experiments, the analysis usually involves the logarithm of the ratio of the expression levels between the target and control samples. For oligonucleotide experiments, the analysis involves a weighted linear combination of the logarithm of the PM expression level and the logarithm of the MM expression level (with some authors choosing zero for the weights of the MM values). Microarray analysis involves several considerations, including the separation of “spot” pixels from “background” pixels and the determination of the expression level from the intensities recorded from the data “spot” pixels; the adjustment of the calculated spot intensity for background (“background correction”); the normalization of the range of fluorescence values from one experiment to another, particularly with oligonucleotide chips; experimental design of multiple slides or chips [1, 2]; data transformations [3, 4], statistical methods of inference and combining information from multiple cDNA experiments [5, 6] and from multiple oligonucleotide arrays [7, 8]; and adjustments for multiple comparisons [9, 10]. The “low-level” analysis consists of the necessary “preprocessing” steps, including data transformation (usually the logarithm) to partially address the non-normality of the expression levels, and normalization and background correction methods to adjust for different signal intensities across different microarray

experiments and sources of variation arising from the chip manufacturing process and background intensity levels. The “high-level” analysis usually involves clustering the gene expression levels into groups of genes that are believed to respond similarly, but no consensus has been achieved on the best methods for normalizing, clustering, and reducing the number of genes to consider as “significantly differentially expressed” when searching for associations between disease and gene locations on chromosomes.

Microarray Analysis in Risk Identification

Microarray analyses have become a standard screening tool in the exploration and elucidation of mechanisms of disease and for studying the interface between the environment and the genome. They also have been used to better understand the effect of certain exposures, such as anthrax, on the cells from human and animal populations, and hence to better characterize the risk of such agents to these populations [11]. As the technology matures and becomes more widely accessible, genome- and proteome-scale microarrays will gain greater acceptance in quantitative risk assessment for many more diseases, vaccines, and other public health initiatives. Examples of useful gene microarray-based investigations that can potentially improve public health practice include discoveries of candidates for biomarkers of disease; measurements of perturbations in the cell cycle under differing conditions; uncovering genetic underpinnings of numerous human, animal, and plant diseases; and explorations of the impacts of environmental change on ecosystems and populations of humans, animals, plants, and disease-causing organisms. An example of the potential utility of microarrays is in the tracking of antigenic drifts and shifts of influenza viruses and the changes in the virus – host relationship. As scientists increase their expertise and knowledge with these types of analyses, and as more results from laboratories are validated in nature, the fields of genomics, proteomics, and further studies of the full complement of cellular components will contribute substantially to how we perceive, measure, and address disease and environmental change.

References

- [1] Kerr, M. & Churchill, G. (2001). Statistical design and the analysis of gene expression microarray data, *Genetics Research* **77**, 123–128.
- [2] Yang, Y. & Speed, T. (2002). Design issues for cDNA microarray experiments, *Reviews* **3**, 579–588.
- [3] Yang, Y., Dudoit, S., Luu, P. & Speed, T. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research* **30**, E15.
- [4] Kafadar, K. & Phang, T. (2003). Transformations, background estimation, and process effects in the statistical analysis of microarrays, *Computational Statistics and Data Analysis* **44**, 313–338.
- [5] Dudoit, S., Yang, Y., Callow, M. & Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica* **12**, 111–139.
- [6] Amaratunga, D. & Cabrera, J. (2001). Analysis of data from viral DNA microchips, *Journal of the American Statistical Association* **96**, 1161–1170.
- [7] Efron, B., Tibshirani, R., Storey, J. & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association* **96**, 1151–1160.
- [8] Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U. & Speed, T. (2002). Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics* **19**, 185–193.
- [9] Reiner, A., Yekutieli, D. & Benjamini, Y. (2002). Identifying differentially expressed genes using false discovery rate controlling procedures, *Bioinformatics* **19**, 368–375.
- [10] Storey, J. (2003). The positive false discover rate: a Bayesian interpretation and the q-value, *Annals of Statistics* **31**, 2013–2035.
- [11] U.S. Department of Energy (2002). Microbial genome program, *U.S. Department of Energy Human Genome News* **12**, 20, http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v12n1/HGN121_2.pdf.

Related Articles

Association Analysis

Linkage Analysis

KAREN KAFADAR AND KATHE E. BJORK

Mobile Phones and Health Risks

Possible health risks of nonionizing electromagnetic fields (EMF) have been controversially discussed amongst researchers, in politics, and in the public for some time (*see* **Extremely Low Frequency Electric and Magnetic Fields**). The main issue is whether EMF levels below permitted levels, directly or indirectly, yield adverse health effects.

This paper summarizes current epidemiological data on health effects associated with EMF emitted by mobile phones. Up to now little evidence linking short-term mobile phone use to health risks has been found. The health effects of long-term (>10 years) mobile phone use, however, need to be clarified. This is important, as the use of mobile phones is increasing worldwide. Hence, eventual minimum effects regarding health risks could be of public health importance [1, 2].

Current Information on Health Risks of Mobile Phone Use

The majority of epidemiological studies carried out in this area to date have focused on mobile phone use and the possible increased risk of brain tumors. Several case-control (*see* **Case-Control Studies**) and cohort studies (*see* **Cohort Studies**) investigating this association have been published. Initially, these were mainly conducted in the United States, Finland, and Sweden. The studies were difficult to compare, as they investigated different brain tumor types and used different exposure assessment methods. Their results were also not always uniform. These first studies were basically all disadvantaged by the relatively short mobile phone use periods of their participants which did not allow the investigation of long-term mobile phone use.

The World Health Organisation (WHO) is coordinating an international case-control study called *INTERPHONE*, being carried out in 13 countries and incorporating more than 6000 brain tumor cases. Currently, eleven reports of this study have been published, three of which report results for acoustic neuroma, two for glioma only, three for glioma and meningioma, two for glioma, meningioma and

acoustic neuroma and one for parotid gland tumor. None of these studies showed an association between brain tumor risk and mobile phone usage of a duration less than 10 years. A combined study of five “INTERPHONE” countries based on 678 patients with acoustic neuroma and 3533 control persons, however, did show a slight increased risk for long-term mobile phone users (10 years or more) [3]. One of the most recent studies also reported an increased risk for glioma amongst long-term mobile phone users [4]. It should, however, be noted that these preliminary results are not confirmed in the other publications and are based on small numbers. Their results thus need to be affirmed. They will have to be verified when a pooled analysis of all the data from the 13 “INTERPHONE” countries is completed. Combined results from the international study are expected in the near future.

Conclusion

Results of studies that have been published up to date do not show any irrefutable association between mobile phone use and an increased risk of glioma or meningioma. Many of the first studies conducted in this area did not have long-term mobile phone users amongst their participants. The study populations were also small, thus limiting their results. More recent studies have been able to incorporate long-term mobile phone users in their analyses. Although final results of the “INTERPHONE” study, especially regarding long-term mobile phone use are still outstanding, preliminary results lead to the assumption that no definite association will be found in this group either.

References

- [1] Blettner, M. & Berg, G. (2000). Are mobile phones harmful? *Acta Oncologica* **39**(8), 927–930.
- [2] Kundi, M. (2004). Mobile phone use and cancer, *Occupational Environment Medicine* **61**, 560–570.
- [3] Schoemaker, M.J., Swerdlow, A.J., Ahlbom, A., Auvinnen, A., Blaasaas, K.G., Cardis, E., Christensen, H.C., Feytching, M., Hepworth, S.J., Johansen, C., Klaeboe, L., Lönn, S., McKinney, P.A., Muir, K., Raitanen, J., Salminen, T., Thomsen, J. & Tynes, T. (2005). Mobile phone use and risk of acoustic neuroma: results of the Interphone case-control study in five North European countries, *British Journal of Cancer* **93**, 842–848.

- [4] Schüz, J., Böhler, E., Berg, G., Schlehofer, B., Hettinger, I., Schläefer, K., Wahrendorf, J., Kunna-Grass, K. & Blettner, M. (2006). Cellular phones, cordless phones, and the risks of glioma, meningioma (Interphone study group, Germany), *American Journal of Epidemiology* **163**, 512–520.

Environmental Health Risk

Environmental Risk Regulation

MARIA BLETTNER AND FLORENCE
SAMKANGE-ZEEB

Related Articles

Environmental Hazard

Model Risk

Overview

Reliance on models to price, trade, and manage risks carries its own risks. Models are susceptible to errors: from incorrect assumptions about price dynamics and market interactions, from inaccurate estimation of volatilities and correlations and other inputs that are not directly observable and must be forecasted.

Model risk relates to the risks involved in using models to value and hedge securities. Model risk becomes a compelling issue for institutions that trade over-the-counter (OTC) derivative products and execute arbitrage strategies.

In liquid and more or less efficient securities markets, the market price is, on average, the best indicator of the asset's value. In the absence of liquid markets and price discovery mechanisms, there is no other alternative than theoretical valuation models to mark-to-model the position, to assess the risk exposure along the various risk factors, and to derive the appropriate hedging strategy.

Since 1973, with the publication of the Black and Scholes [1] and Merton [2] option-pricing model (*see Risk-Neutral Pricing; Importance and Relevance; Statistical Arbitrage*), we have witnessed relentless increased complexity in the theoretical approach to valuation. The area of fixed income instruments and derivatives is striking. With increased sophistication in the design of securities, like the addition of imbedded options, valuation is now often based on complex multifactor stochastic interest rate models, which are only understood by "rocket scientists".

The pace of model development has accelerated to support the rapid growth of financial innovations such as caps, floors, swaptions (*see Nonparametric Calibration of Derivatives Models*), spread options (*see Credit Migration Matrices*), and other exotic derivatives (*see Equity-Linked Life Insurance; Insurance Pricing/Nonlife*). These innovations have also been made possible because of new developments in financial theory, which allow a better capture of the many facets of financial risks. At the same time, these models would have never been implemented in practice, and not so well accepted on the trading floor, had the growth in computing power not accelerated so dramatically. It seems that financial innovations, model development, and computing

power are engaged in a sort of leapfrog game; financial innovations call for more model complexity, which in turn requires more computing power. Using complex models and huge computing power increases the traders' level of comfort and gives them the incentive to innovate more.

The main causes of model risk are as follows:

- *Model error*

The model might contain mathematical errors or, more likely, be based on simplifying assumptions that are misleading or inappropriate.

- *Implementing a model wrongly*

The model might be implemented wrongly, either by accident or as part of a deliberate fraud.

Model Error

Derivatives trading depends heavily on mathematical models that make use of complex equations and advanced mathematics. In the simplest sense, a model is incorrect if there are mistakes in the analytical solution (in the set of equations, or in the solution of a system of equations).

A model is also said to be incorrect if it is based on wrong assumptions about the underlying asset price process – and this is perhaps both a more common and a more dangerous risk. The history of the financial industry is littered with examples of trading strategies based on shaky assumptions and some model risks are really just a formalization of this kind of mistake. For example, a model of bond pricing might be based on a flat and fixed term structure, at a time when the actual term structure of interest rates is steep and unstable.

The most frequent error in model building is to assume that the distribution of the underlying asset is stationary (i.e., unchanging), when, in fact, it changes over time. The case of volatility is particularly striking. Derivatives practitioners know very well that volatility is not constant. The ideal solution would be to acknowledge that volatility is stochastic and to develop an option-pricing model that is consistent with this, but option-valuation models become computationally difficult when any sort of stochastic volatility is included. (Moreover, introducing new unobservable parameters associated with the volatility process into the valuation model makes the estimation problem even more severe.)

Instead, derivative practitioners find themselves engaged in a continual struggle to find the best compromise between complexity (to better represent reality) and simplicity (to improve the tractability of their modeling).

While traders know that they are making simplifying assumptions about price behavior, it is not easy for them (or for risk managers) to assess the impact of this kind of simplifying assumption on any given position or trading strategy. For example, practitioners often assume that rates of return are normally distributed, i.e., that they have a classic “bell-shaped” distribution. However, empirical evidence points to the existence of “fat tails” in distributions; in these distributions, unlikely events are, in fact, much more common than would be the case if the distributions were normally distributed. Therefore, empirical distributions rather than theoretical distributions should be used to help alleviate the danger of an unrecognized fat tail, where possible. However, such fat tails are not accounted for in the theoretical distributions that lie behind many classic models, such as the Black–Scholes option-pricing model.

Another way to oversimplify a model is to underestimate the number of risk factors that it must take into account. For simple “vanilla” investment products, such as a callable bond, a one-factor term structure model (*see* **Dynamic Financial Analysis; Insurance Applications of Life Tables**), where the factor represents the spot short-term rate, may be enough to produce accurate prices and hedge ratios. For more complex products, such as spread options or exotic structures, a two- or three-factor model may be required, where the factors are, for example, the spot short-term and long-term rates for a two-factor model.

Another problem is that models are almost always derived under the assumption that perfect capital markets exist. In reality, many markets, especially those in less-developed countries, are far from perfect. Meanwhile, even in developed markets, OTC derivative products (*see* **Alternative Risk Transfer; Weather Derivatives**) are not traded publicly and usually cannot be perfectly hedged.

Liquidity, or rather the absence of liquidity, may also be a major source of model risk. Models assume that the underlying asset can be traded long or short at current market prices, and that prices will not change dramatically when the trade is executed.

In the same way, models that are safe to use for certain kinds of product might not perform well

at all on subtly different instruments. Many OTC products have options embedded within them that are ignored in the standard option-pricing model. For example, using a model to value warrants may yield biased results if the warrant is also extendable. Other common errors include using the Black–Scholes option-valuation model to price equity options, while adjusting for dividends by subtracting their present value from the stock price. This ignores the fact that the options can be exercised early. Applying the wrong model is also easy, if the researcher is not clear about whether the underlying instrument is a primary asset or is itself a contingent claim on another asset (or basket of assets).

Implementing a Model Wrongly

Even if a model is correct and is being used to tackle an appropriate problem, there remains the danger that it will be wrongly implemented. With complicated models that require extensive programming, there is always a chance that a programming “bug” may affect the output of the model. Some implementations rely on numerical techniques that exhibit inherent approximation errors and limited ranges of validity. Many programs that seem error-free have been tested only under normal conditions, and so may be error-prone in extreme cases and conditions.

In models that require a Monte Carlo simulation (*see* **Risk Measures and Economic Capital for (Re)insurers; Structured Products and Hybrid Securities; Reliability Optimization**), large inaccuracies in prices and hedge ratios can creep in if insufficient simulation runs or time steps are implemented. In this case, the model might be right, and the data might be accurate, but the results might still be wrong if the computation process is not given the time it needs.

For models evaluating complex derivatives, data are collected from many different sources. The implicit assumption is that for each time period, the data for all relevant assets and rates pertain to exactly the same time instant, and thus reflect simultaneous prices. Using nonsimultaneous price inputs may be necessary for practical reasons, but again, it can lead to wrong pricing.

When implementing a pricing model, statistical tools are used to estimate model parameters such as volatilities and correlations. An important question, then, is how frequently should input parameters be

refreshed? Should the adjustment be made on a periodic basis, or should it be triggered by an important economic event? Similarly, should parameters be adjusted according to qualitative judgments, or should these adjustments be purely based on statistics? The statistical approach is bound to be in some sense “backward looking”, while a human adjustment can be forward looking, i.e., it can take into account a personal assessment of likely future developments in the relevant markets.

All statistical estimators are subject to estimation errors involving the inputs to the pricing model. A major problem in the estimation procedure is the treatment of “outliers”, or extreme observations. Are the outliers really outliers, in the sense that they do not reflect the true distribution? Or are they important observations that should not be dismissed? The results of the estimation procedure will be vastly different depending on how such observations are treated. Each bank, or even each trading desk within a bank, may use a different estimation procedure to estimate the model parameters. Some may use daily closing prices, while others may use transaction data. Whether the researcher uses calendar time (i.e., the actual number of days elapsed), trading time (i.e., the number of days on which the underlying instrument is traded), or economic time (i.e., the number of days during which significant economic events take place) affects the calculation.

Finally, the quality of a model depends heavily on the accuracy of the inputs and parameter values that feed the model. This is particularly true in the case of relatively new markets, where best-practice procedures and controls are still evolving. The old adage “garbage in, garbage out” should never be forgotten when implementing models that require the estimation of several parameters. Volatilities and correlations are the hardest input parameters to judge accurately. For example, an option’s strike price and maturity are fixed, and asset prices and interest rates can be easily observed directly in the market – but volatilities and correlations must be forecast. This gives rise to many opportunities for both genuine mistakes and deliberate tampering that can be countered only through robust control procedures and independent vetting.

The most frequent problems in estimating values, on the one hand, and assessing the potential errors in valuation, on the other hand, are as follows:

- *Inaccurate data*

Most financial institutions use internal data sources as well as external databases. The responsibility for data accuracy is often not clearly assigned. It is therefore very common to find data errors that can significantly affect the estimated parameters.

- *Inappropriate length of sampling period*

Adding more observations improves the power of statistical tests and tends to reduce the estimation errors; but, the longer the sampling period, the more weight is given to potentially stale and obsolete information. Especially in dynamically changing financial markets, “old” data can become irrelevant and may introduce noise into the estimation process.

- *Problems with liquidity and the bid/ask spread*

In some markets, a robust market price does not exist. The gap between the bid and ask prices may be large enough to complicate the process of finding a single value. Choices made about the price data at the time of data selection can have a major impact on the output of the model.

How Can We Mitigate Model Risk?

One important way to mitigate model risk is to invest in research to improve models and to develop better statistical tools. It is critical for a financial institution to keep up with the latest developments and to spread new knowledge quickly throughout the business.

An even more vital way of reducing model risk is to establish a process for independent vetting of how models are both selected and constructed. This should be complemented by independent oversight of the profit and loss (P&L) calculation.

The role of vetting is to offer assurance to the firm’s management that any model for the valuation of a given security proposed by, say, a trading desk is reasonable. In other words, it provides assurance that the model offers a reasonable representation of how the market itself values the instrument, and that the model has been correctly implemented. Vetting should consist of the following phases:

1. Documentation

The vetting team should ask for full documentation of the model, including both the assumptions underlying the model and its mathematical expression. This should be independent of any particular implementation, such as a spreadsheet or a C++ computer code, and should include

(a) the term sheet or, equivalently, a complete description of the transaction;

(b) a mathematical statement of the model, which should include

(i) an explicit statement of all the components of the model: stochastic variables and their processes, parameters, equations, and so on;

(ii) the payoff function and/or any pricing algorithm for complex structured deals;

(iii) the calibration procedure for the model parameters;

(iv) the hedge ratios/sensitivities;

(v) implementation features, i.e., inputs, outputs, and numerical methods employed;

(vi) a working version of the implementation.

2. Soundness of model

The model vetter needs to verify that the mathematical model is a reasonable representation of the instrument that is being valued. For example, the manager might reasonably accept the use of a particular model (e.g., the Black model) for valuing a short-term option on a long-maturity bond, but reject (without even looking at the computer code) the use of the same model to value a 2-year option on a 3-year bond. At this stage, the risk manager should concentrate on the finance aspects and not become overly focused on the mathematics.

3. Independent access to financial rates

The model vetter should check that the bank's middle office has independent access to an independent market-risk-management financial rates database (to facilitate independent parameter estimation).

4. Benchmark modeling

The model vetter should develop a benchmark model based on the assumptions that are being made and on the specifications of the deal. Here, the model vetter may use an implementation that is different from the implementation that is being proposed. A proposed analytical model can be tested against a numerical approximation technique or against a simulation approach. (For example, if the model to be vetted is based on a "tree" implementation, one may rely on the partial differential equation approach instead and use the finite-element technique to derive the numerical results.) Compare the results of the benchmark test with those of the proposed model.

5. Health check and stress test the model

Also, make sure that the model possesses the basic properties that all derivatives models should possess, such as put-call parity and other nonarbitrage conditions. Finally, the vetter should stress test the model. The model can be stress tested by looking at some limit scenario in order to identify the range of parameter values for which the model provides accurate pricing. This is especially important for implementations that rely on numerical techniques.

6. Build a formal treatment of model risk into the overall risk-management procedures, and periodically reevaluate models

Also, reestimate parameters using best-practice statistical procedures. Experience shows that simple, but robust, models tend to work better than more ambitious, but fragile, models. It is essential to monitor and control model performance over time.

Conclusion

Models are an inevitable feature of modern finance, and model risk is inherent in the use of models. The most important aspect is to be aware of the dangers. Firms must avoid placing undue faith in model values, and must hunt down all the possible sources of inaccuracy in a model. In particular, they must learn to think through situations in which the failure of a model might have a significant impact.

We have stressed the technical elements of model risk, but we should also be wary of the human factor in model risk losses. Large trading profits tend to lead to large bonuses for senior managers, and this creates an incentive for these managers to believe the traders who are reporting the profits (rather than the risk managers or other critics who might be questioning the reported profits). Often, traders use their expertise in formal pricing models to confound any internal critics, or they may claim to have some sort of informal but profound insight into how markets behave. The psychology of this behavior is such that we are tempted to call it the "Tinkerbell" phenomenon, after the scene in *Peter Pan* in which the children in the audience shout, "I believe, I believe" in order to revive the poisoned fairy Tinkerbell. The antidote is for senior managers to approach any model that seems to record or deliver above-market returns with a healthy skepticism, to insist that models are made

transparent, and to make sure that all models are independently vetted. The topic of model risk is discussed in detail in [3]. Also, many articles on this subject appear in [4].

The comments on model risk in this chapter are made in the context of a trading environment. But, they can be easily transposed in other contexts such as risk management.

References

- [1] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.
- [2] Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.
- [3] Crouhy, M., Galai, D. & Mark, R. (2005). *The Essentials of Risk Management*, McGraw-Hill.
- [4] Gibson, R. (2005). *Model Risk, Concepts, Calibration and Pricing*, Risk Books.

MICHEL CROUHY, DAN GALAI
AND ROBERT MARK

Modifiable Areal Unit Problem (MAUP)

Geographical scale and spatial resolution are important considerations in the design, implementation, and interpretation of research studies aimed at understanding the link between environmental hazards and related adverse effects on people and ecosystems [1] (see **Geographic Disease Risk; Spatiotemporal Risk Analysis**). Sources of pollution, ambient pollutant concentrations, population exposure levels, and occurrence of environmentally related adverse effects all vary according to spatial scale and resolution – where “scale” refers to the spatial (or areal) extent of the analysis and “resolution” identifies the level of aggregation for statistical data gathering. The range of geographical scales typically applied to environmental exposure studies ranges from personal exposure investigations where emphasis is on pollutant concentrations within a few centimeters of a person’s body (10^{-1} km) to studies of global exposure to UV radiation that encompass continents, if not the entire world (10^4 km). The scope of statistical aggregation for health outcome data is illustrated by information collected by the US census, which reports data at the block, block group, tract, minor civil division or MCD, county, state, and regional levels.

Spatial (geographical) data on environmental health are often aggregated to make communication of research results convenient and informative. The modifiable areal unit problem or *MAUP* is a term coined by geographers to describe a potential source of error that can affect spatial studies using aggregate sources of data [2, 3]. The MAUP is a formal statement of the fact that statistical results can and do vary by (a) scale – the number of spatial units (or zones) used in the analysis and (b) aggregation – the grouping of smaller geographical areas into larger spatial units. Thus a variable that is significant in a regression analysis at one geographic scale may be insignificant at another, or a spatial pattern of data points that appears random at one level of aggregation may seem clustered at another.

Among other problems, the MAUP can contribute to a misinterpretation of aggregate data by which incorrect inferences about the nature of individuals

are based exclusively on combined statistics from the group to which those individuals belong. This mistake is called the *ecological fallacy*, where the fallacy lies in assuming that all group members have specific characteristics of the larger group (e.g., assuming that all basketball players are tall or that everyone from Ireland has red hair). The fact that statistical findings from analysis of aggregate spatial data depend on subjective and often arbitrary decisions about geographic scale and resolution is a sobering reality, especially given that results often serve as the basis for risk management decisions, public policy choices, and communication strategies.

Public health professionals have known about the MAUP for years and routinely recognize the importance of addressing spatial issues [1]. Exposure analysts, for example, must deal with the uneven geographical distribution of people and pollution and they direct significant resources toward estimating exposures in ways that will minimize misclassification errors. Biostatisticians know that the MAUP can affect statistical analysis of exposure–response relationships and they therefore take steps to deal with issues of misaligned data, data incompatibility, and validity of inferences. Epidemiologists are trained to take account of spatial scale and resolution in designing studies, and they are cognizant of the need to avoid the ecological fallacy when interpreting findings. Ultimately, the development of suitable statistical frameworks and models along with practical guidance on choosing appropriate spatial units and aggregation levels will be necessary to solve MAUP-related quandaries.

Scientists today continue to examine the MAUP and its impact on research design and interpretation of scientific findings. For example, although the study of geographical patterns in cancer prevalence and/or incidence has shed much light on disease control and prevention, there is still no scientific consensus on how best to designate geospatial locations of health events (i.e., geocoding) so that results are accurate (within acceptable error), precise (to a desired areal unit of analysis), and “fit for use” (applicable to other available data) [4]. Thus researchers must choose, based on pragmatic rather than scientific considerations, between using smaller areal units that contain few at-risk subjects and yield less reliable rates or larger areal units that may blur meaningful variation occurring within locales. A recent study by Gregorio and his colleagues [4] suggests that if

surveillance of small area cancer clusters is not the goal, then the costs of obtaining and preparing data at spatial units smaller than the census tract may not be justified.

As Sheppard *et al.* [5] point out, results of spatial analysis of *environmental justice* (see **Statistics for Environmental Justice**) issues (i.e., questions of whether poor minority populations bear a disproportionate burden of pollution) are dependent on choices of geographic scale of the unit of analysis (e.g., municipalities, counties, metropolitan areas, and states) and spatial resolution of data within each unit (e.g., zip codes, census tracts, and block groups). Thus individual environmental justice studies may be independently accurate, but give apparently contradictory results, (which are in fact not conflicting), if proper account is not taken of the effect of diverse spatial scales and levels of resolution on analytical outcomes.

To illustrate the types of difficulties created by the MAUP, consider two examples. First, imagine a hypothetical town of four people, all of whom are exposed to a particular environmental pollutant. As shown in Figure 1, two residents live north of a river dividing the town and two live south. Both (a) the disease outcomes actually experienced by the people in the town during the study period and (b) the potential disease outcomes that would have occurred if no one was exposed to the environmental pollutant are

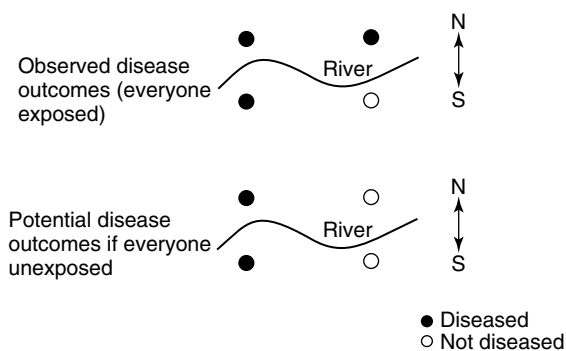


Figure 1 The implications of spatial scale for causal inference – comparison of observed and potential disease outcomes for a hypothetical town during a specified period of observation under two different exposure scenarios [Reproduced from [1]. © Taylor & Francis Group, 2002.]

depicted. The causal incidence proportion ratio (IPR) is a parameter used by epidemiologists to compare the incidence proportion (risk) of disease that a target population would experience during a specified period under two different exposure conditions. It is interpreted as the proportionate increase in risk for the target population over the study period that is caused by the differences in the two exposure conditions. The important point is that the IPR changes depending on the spatial scale. For people north of

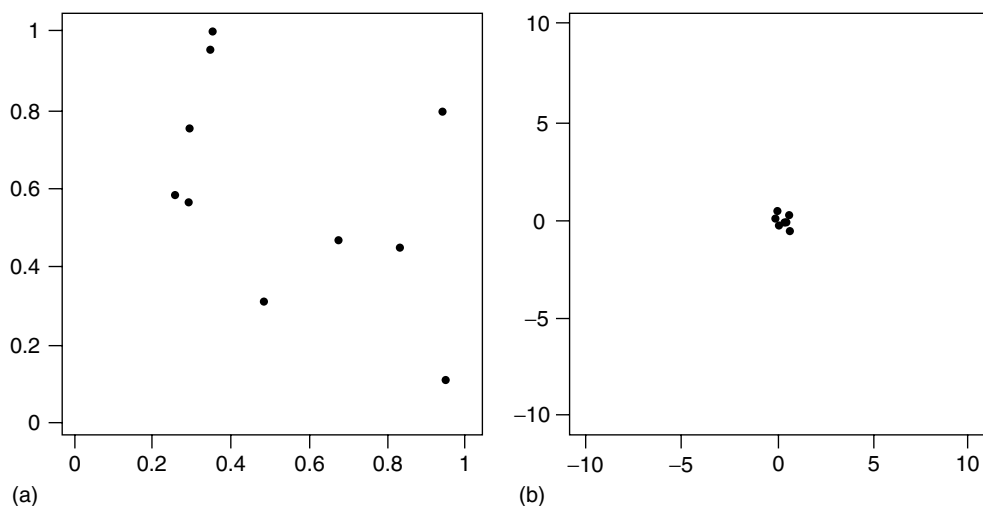


Figure 2 The implications of spatial scale for disease clustering – the same number and relative positioning of disease cases can appear random at one scale and clustered at another [Reproduced from [1]. © Taylor & Francis Group, 2002.]

the river, the risk of disease is 1 (two people/two people with the disease) if they are exposed and 0.5 (1/2) if they are not, which means the IPR is equal to 2 (1/0.5). On the other hand, the risk for the two people south of the river is 0.5 (1/2) if exposed and also 0.5 (1/2) if not exposed, which makes the IPR equal to 1. When the entire town is considered, the IPR becomes 1.5 (0.75/0.5) based on a 0.75 (3/4) risk of disease if exposed, and a 0.5 (2/4) risk if not exposed [1].

Second, consider an investigation of the pattern of disease incidence in an area where the population is suspected of being exposed to a toxic environmental agent, while others in outlying areas are thought not to be exposed. An analyst typically will assign a location (e.g., residential address) to each person who has the disease thought to be caused by the environmental agent. Statistical techniques are then used to look for a spatial pattern of disease cases that might suggest a higher incidence in the area of putative exposure. Detecting a “disease cluster” in the area where people are likely to be exposed provides suggestive evidence that the “cluster” is an environmentally induced effect. However, as shown in Figure 2, even though the number and pattern of disease cases remains the same, they can appear random at one scale and clustered at another [1].

In summary, issues of spatial scale and resolution are an intrinsic part of protecting and enhancing environmental health. The MAUP has important ramifications for the design and interpretation of studies

aimed at increasing our knowledge and furthering our understanding of environmental exposures and environmentally induced effects. Ultimately, sound policies to safeguard public health and protect environmental quality depend on our ability to make the right choices about spatial units of analysis and levels of data aggregation within those units.

References

- [1] Sexton, K., Waller, L.A., McMaster, R.B., Maldonado, G. & Adgate, J.L. (2002). The importance of spatial effects for environmental health policy and research, *Human and Ecological Risk Assessment* **8**(1), 109–125.
- [2] Openshaw, S. & Taylor, P.J. (1979). A million or so correlated coefficients: three experiments on the modifiable areal unit problem, in *Statistical Applications in the Spatial Sciences*, N. Wrigley & R.J. Bennet, eds, Pion, London.
- [3] Unwin, D.J. (1996). GIS, spatial analysis and spatial statistics, *Progress in Human Geography* **20**(4), 540–541.
- [4] Gregorio, D.I., DeChello, L.M., Samociuk, H. & Kulldorf, M. (2005). Lumping or splitting: seeking the preferred areal unit for health geography studies, *International Journal of Health Geographics* **4**, 6.
- [5] Sheppard, E., Lettner, H., McMaster, R.B. & Tian, H. (1999). GIS-based measures of environmental equity: exploring their sensitivity, *Journal of Exposure Analysis and Environmental Epidemiology* **9**(1), 18–28.

KEN SEXTON

Molecular Markers and Models

Our understanding of the mechanisms underlying cellular processes, the interactions of these various processes, and the ways in which abrogation of one or more of these processes can lead to adverse health outcomes has increased enormously over the past decade. In addition, a considerable effort is being expended in trying to determine how environmental agents (particularly chemicals, nanomaterials, radiation, and microbes) can play a role in this abrogation of cellular processes, and in turn enhance the disease process (*see Environmental Health Risk; Environmental Hazard; Gene–Environment Interaction*). The challenge is how to bring together these two components – mechanistic understanding of disease outcomes and the impact of environmental agents – in the science of risk assessment. This linkage has been addressed in a formal manner through the US Environmental Protection Agency’s *Guidelines for Carcinogen Risk Assessment* [1] as well as in a preliminary manner by the European Commission in its *First Report on the Harmonization of Risk Assessment Procedures* (ec.europa.eu/food/fs/sc/ssc/out83_en.pdf). A concise review of the history and practices of quantitative-risk assessment has been presented by Hrudey [2]. The main component of this linkage is that the use of mechanistic data in cancer-risk assessment is encouraged to reduce uncertainty, in particular the uncertainty that is associated with any default approach. A similar approach can be applied to risk assessments for other adverse outcomes, including reproductive, neurological, and developmental disorders. This chapter addresses the general types of approaches that can be used for incorporating mechanistic data into quantitative-risk assessments, and the types of data that would help enhance these approaches. There will be an emphasis on cancer risk in order to focus the discussion.

Definition of Terms

A number of terms are used in the context of risk assessment that need to be defined for the purposes of comprehension of their subsequent use.

Mode of action (MOA): is described by a sequence of key events and processes, starting with interaction of an agent with a cell, proceeding through functional and anatomical changes, and resulting in cancer formation [1] (e.g., DNA reactivity, mitogenesis, cytotoxicity with regenerative cell proliferation, immune suppression, and receptor-mediated events).

Mechanism of action: implies a more detailed understating and description of events, often at the molecular level, than is meant by MOA.

Key event: an empirically observable precursor step that is itself a necessary element of the MOA or is a biologically based marker for such an element [1]. For cancer formation, a scheme for the key events whereby a DNA-reactive carcinogen induces tumors has been presented recently by Preston and Williams [3]. It is presented here as an example of the types of key events that could prove to be necessary for molecular-modeling approaches (Table 1). Key events along the pathway to tumor development for DNA-reactive carcinogens, for example, can be assessed both qualitatively and quantitatively by experimental studies and, where possible, by human studies (*see Environmental Carcinogenesis Risk*). The general approach is to establish the relevance to humans of an MoA developed for an animal model, for which data are likely to be more readily available, using the human relevance framework [4].

Key Events and Molecular Markers

In the previous section, a set of key events is presented for tumor production, in general, for exposures to DNA-reactive carcinogens. This same approach can be used for developing a set of key events for non-DNA-reactive (but mutagenic) chemicals. In this case, DNA damage is produced in the absence of covalent binding or other direct DNA–chemical interaction. Once DNA damage is produced, the remaining key events (3–10) are the same as for DNA-reactive carcinogens. The same general principles can be applied for defining MOA in terms of key events for noncancer endpoints. In this situation, however, the key events tend to be chemical specific rather than adverse-outcome specific because of the difficulty of identifying a set of key events for noncancer outcomes. The approach is described in a set of articles in a recent volume of *Critical Review in Toxicology* [5]. As more is learned about

Table 1 Key events for tumor development: DNA-reactive MOAs^(a)

1	Exposure of target cells (e.g., stem cells) to ultimate DNA-reactive and mutagenic species – in some cases this requires metabolism
2	Reaction with DNA in target cells to produce DNA damage
3	Misreplication on damaged DNA template or misrepair of DNA damage
4	Mutations in critical genes in replicating target cell
5	These mutations result in initiation of new DNA/cell replication
6	New cell replication leads to clonal expansion of mutant cells
7	DNA replication can lead to further mutations in critical genes
8	Imbalanced and uncontrolled clonal growth of mutant cells may lead to preneoplastic lesions
9	Progression of preneoplastic cells results in emergence of overt neoplasms, solid tumors (which require neoangiogenesis), or leukemia
10	Additional mutations in critical genes as a result of uncontrolled cell division results in malignant behavior

^(a) Reproduced from [3]. © John Wiley & Sons, Inc., 2005

the specific mechanisms underlying reproductive, developmental, and neurological disorders and the potential role of exposure to environmental agents, it is likely that a set of key events can be developed for any particular outcome. This will certainly enhance the risk-assessment process for these non-cancer diseases.

The great value of using key events in a risk-assessment process is that information gathered on the response of a key event to an environmental stressor is, by definition, informative of the impact of that same stressor on disease induction. Thus, key events, or more likely, a combination of several key events, can be used for the various extrapolations that are a requirement for risk assessment. Such extrapolations include high to low dose, species to species, and acute to chronic exposures, and are frequently the “Achilles heel” of the risk-assessment process [6]. In this regard, it is important to note that considerable caution has to be taken when extrapolating from data obtained with animal models of human diseases to predict disease outcomes in humans. There are frequently significant differences in the specifics of disease formation between humans and animals, even for the same apparent disease.

The recent, very significant enhancement of the ability to unravel the processes underlying a range of diseases at the molecular level has been a consequence of the advances in molecular biology techniques, computational approaches to data analysis, and the development of systems-based models. These together have allowed for a definition of adverse health outcomes, particularly cancer, in terms of key molecular markers.

A very clear example of the application of such approaches is the development of a “universal” tumor model described by Hanahan and Weinberg [7] in their *Hallmarks of Cancer*. This model consists of a set of six acquired characteristics that are essential for the formation of all tumors, irrespective of tumor type and species. These characteristics are broadly described as follows: self-sufficiency in growth signals, insensitivity to antigrowth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis. It seems probable that there is no specific order for obtaining these characteristics, although it might be predicted that changes in growth signaling will occur early on in the process. This model has been elegantly extended by Hahn and Weinberg [8]. They described an approach for modeling the molecular circuitry of cancer, whereby a comparison is made of the shared and unique elements (based on the acquired characteristics) of human and mouse models of cancer to aid in the identification of the molecular circuits that function differently in humans and mice. This knowledge can then be used to improve existing models of cancer. An interactive website (<http://www.nature.com/nrc/poster/subpathways/index.html>) utilized a so-called subway map of cancer pathways, and allows the user to investigate specific pathways along the route to cancer. These various models for describing pathways to tumorigenesis based on the identification of key events is synthesized in a comprehensive review by Hahn and Weinberg, *Rules for Making Human Tumor Cells* [9].

It is likely that the most informative key events for tumor development following exposures to environmental chemicals are to be found within the set of

rules for making human (or rodent) tumor cells. The more proximate a key event is to the development of a tumor itself (i.e., it is itself on the pathway to tumorigenesis), the more likely it is that it can be used to predict tumor outcomes both qualitatively and quantitatively. This application is critical for predicting tumor responses at doses below those at which tumors themselves are observed, and for predicting human tumor frequencies from rodent tumor responses. These are essential steps in cancer-risk assessment because the essential epidemiological data are almost always lacking for exposures to environmental chemicals (*see Environmental Risks*).

Molecular Models and Systems Approaches

The development of diseases is a complex process that involves whole organ or tissue responses, even if much of the pertinent data might be collected at the single cell/single pathway level or from averaged responses at the cell population level. What is really needed for risk assessment is a systems approach, where a disease outcome is described by modifications to and abrogations of a system (cellular, tissue, or even whole organism) in response to environmental chemicals. At this juncture, we are some distance away from meeting this need, but there are certainly encouraging steps along this path. A very informative review on the approaches for incorporating quantitative modeling into cell biology is provided by Mogilner *et al.* [10]. They discuss the utility of models of cell signaling, cytoskeletal self-organization, nuclear transport, and the cell cycle. These are all basic cell functions whose abrogation is involved in disease pathways. Thus, a model of normal biology is essential before embarking on the assessment of effects of dysfunction in any of these fundamental processes.

In a similar vein, Mogilner *et al.* [11] have developed a molecular model of mitosis and the mechanisms by which sister chromatids are segregated. In particular, they combine mathematical models and computer simulations together with experimental studies to describe the molecular mechanisms by which the spindle machinery performs its critical functions. Chromosome gains and losses are a significant component of the formation of all tumors, and a high proportion of birth defects and reproductive loss, and so knowledge of the normal process of

mitosis and how this might be disrupted can form an informative component of risk assessments.

On an even broader basis, Kleppe *et al.* [12] have addressed the issue of how signal transduction is orchestrated at the molecular level, which is essential to the understanding of most, if not all, cellular functions. The investigators' approach is to take advantage of high-throughput methods in proteomics and genomics that provide quantitative and temporal information. Such data are integrated to generate predictive models of cellular processes. The authors argue (appropriately) that the generation of such models provides powerful tools for understanding how systems operate in healthy and pathologic states.

Finally, sophisticated computational models can be used to identify key events on the basis of whole genome analysis. These types of models can be used to compare essential functions and their controls across species. Lehner *et al.* [13] have developed such an approach for the systematic mapping of genetic interactions in *C. elegans* to identify common modifiers of diverse signaling pathways. They note that most heritable traits, including disease susceptibility, are impacted by interactions between multiple genes. To test such a hypothesis, Lehner *et al.* used RNA interference to systematically test about 65 000 pairs of genes for their interaction potential. This approach has led to the identification of about 350 genetic interactions among genes involved in signaling pathways that are also mutated in human diseases. Of particular note, in the context of potential key events, the authors identified a class of "highly connected 'hub' genes". If one of these "hub" genes is inactivated it can enhance the phenotypic effects of mutations in a range of different genes. These hub genes *all* encode chromatin modulators, and this activity appears to be conserved across animal species. It is possible that these chromatin modulator hub genes could act as modifier genes in multiple, mechanistically unrelated genetic diseases in humans. This view is supported by studies that link chromatin modulation with human diseases (reviewed in [14]). This type of molecular modeling at the systems level is clearly one of the ways forward for better defining the cellular "master genes" and, when their functions are modified, their role in disease etiology. The subsequent application of such systems-based models to a risk-assessment process that incorporates the use of mechanistic data is both clear to see and inevitable.

Acknowledgments

I wish to thank Dr Stephen Edwards and Dr Rory Conolly for their very insightful review of this manuscript. I also thank Carolyn Fowler for her expert assistance in the preparation of the manuscript.

This manuscript has been reviewed in accordance with the U.S. Environmental Protection Agency policy and approved for publication although it does not necessarily reflect Agency policy.

References

- [1] U.S. Environmental Protection Agency (U.S. EPA) (2005). *Guidelines for Carcinogen Risk Assessment, Risk Assessment Forum*, Washington, DC, <http://www.epa.gov/cancerguidelines>.
- [2] Hrudey, S.E. (1998). Quantitative cancer risk assessment – pitfalls and progress, in *Risk Assessment and Risk Management*, R.E. Hester & R.M. Harrison, eds, Royal Society of Chemistry, London, pp. 57–90.
- [3] Preston, R.J. & Williams, G.M. (2005). DNA-reactive carcinogens: mode of action and human cancer hazard, *Critical Reviews in Toxicology* **35**, 673–683.
- [4] Meek, M.E., Bucher, J.R., Cohen, S.M., Dellarco, V., Hill, R.N., Lehman-McKee, L.D., Longfellow, D.G., Pastoor, T., Seed, J. & Patton, D.E. (2003). A framework for human relevance analysis of information on carcinogenic modes of action, *Critical Reviews in Toxicology* **33**, 591–654.
- [5] McClellan, R.O. (ed) (2005). *Critical Reviews in Toxicology*, Taylor and Francis, Philadelphia, Vol. 35, no. 8–9, pp. 663–781.
- [6] Preston, R.J. (2005). Extrapolations are the Achilles heel of risk assessment, *Mutation Research* **589**, 153–157.
- [7] Hanahan, D. & Weinberg, R.A. (2000). The hallmarks of cancer, *Cell* **100**, 57–70.
- [8] Hahn, W.C. & Weinberg, R.A. (2002). Modelling the molecular circuitry of cancer, *Nature Reviews Cancer* **2**, 331–341.
- [9] Hahn, W.C. & Weinberg, R.A. (2002). Rules for making human tumor cells, *New England Journal of Medicine* **347**, 1593–1603.
- [10] Mogilner, A., Wollman, R. & Marshall, W.F. (2006). Quantitative modeling in cell biology: what is it good for? *Developmental Cell* **11**, 279–287.
- [11] Mogilner, A., Wollman, R., Civelekoglu-Scholey, G. & Scholey, J. (2006). Modeling mitosis, *Trends in Cell Biology* **16**, 88–96.
- [12] Kleppe, R., Kjarland, E. & Selheim, F. (2006). Proteomic and computational methods in systems modeling of cellular signaling, *Current Pharmaceutical Biotechnology* **7**, 135–145.
- [13] Lehner, B., Crombie, C., Tischler, J., Fortunato, A. & Fraser, A.G. (2006). Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways, *Nature Genetics* **38**, 896–903.
- [14] Cho, K.S., Elizondo, L.I. & Boerkoel, C.F. (2004). Advances in chromatin remodeling and human disease, *Current Opinion in Genetics Development* **14**, 308–315.

Related Articles

Cumulative Risk Assessment for Environmental Hazards
Gene–Environment Interaction
Statistics for Environmental Mutagenesis

R. JULIAN PRESTON

Monte Carlo Methods

Monte Carlo techniques are used in situations in which the analytic solution of the problem is either intractable or time consuming. Instead of calculating exact quantities, simulation is used to produce stochastic approximations to the solution.

In practice, Monte Carlo methods are discussed interchangeably with simulation. Useful discussions can be found in a variety of monographs, such as [1–3], to which the reader is referred for a thorough review.

Monte Carlo techniques are used in a wide range of problems such as the construction of Monte Carlo hypothesis tests, bootstrap distributions, and for numerical integration in Bayesian calculations (*see Bayesian Statistics in Quantitative Risk Assessment*). More specialized uses are the analysis of complex stochastic systems.

Monte Carlo techniques have a long history in mathematics. An early example is Buffon's Needle dating from 1733, described in [1]. Realistically, the widespread application of Monte Carlo techniques has only become feasible with the availability of cheap and efficient computing since the 1970s. In fact, the exponential increase in available computing power has led to Monte Carlo techniques facilitating major statistical advances.

As a trivial example of a Monte Carlo method, we consider the calculation of the mean μ of a density $g(x)$. The analytic solution is

$$\mu = \int x g(x) dx \quad (1)$$

which may be difficult to evaluate. In the Monte Carlo approach, we sample k observations, $X_1, \dots, X_k$, from $g(x)$ and form the Monte Carlo estimate

$$\hat{\mu} = \frac{\sum_{i=1}^k X_i}{k} \quad (2)$$

A law of large numbers can be used to show that this estimate will converge strongly to the true value, provided the integral exists. Thus, we can produce an estimate of a required accuracy by manipulating k .

An important point to note is that the term *Monte Carlo* refers not to a particular stochastic algorithm

but only to the fact that a stochastic algorithm has been used. For instance, in the previous example there are infinite number of different stochastic algorithms that we could use to sample the observations and estimate μ .

The choice of algorithm depends on a variety of factors. In many problems, there is a “natural” formulation. For example, we estimate the mean of a distribution by taking the mean of a sample from the distribution. Another criterion that should be considered when choosing an algorithm is the efficiency of the method. This leads to the contemplation of variance reduction techniques, which are used to produce more accurate estimates for the same computational effort.

In practice, there is a trade-off between the use of an inefficient algorithm that is easy to design and implement and an efficient algorithm that could take detailed analysis to design. Provided that the inefficient algorithm is not pathological, it is sufficient merely to run the algorithm for an appropriately longer period of time to obtain results comparable to those with the efficient algorithm. An important caveat is that great care must be taken to ensure that satisfactory convergence has occurred. In any application, the relevant literature should be consulted.

We now briefly describe several important examples of the use of Monte Carlo techniques.

Monte Carlo Hypothesis Testing

The first technique considered is that of Monte Carlo hypothesis testing. Assume we have a statistic T , and a simple null hypothesis H_0 and composite alternative H_a . Consider constructing a test of size $1 - \alpha$. If the sampling distribution of $T|H_0$ is known and analytically tractable, then we can use standard calculus methods to construct a critical region for rejecting H_0 .

Alternatively, we consider constructing a Monte Carlo test of H_0 . Specification of the rejection region involves the calculation of a quantile of the null distribution of T . In the Monte Carlo approach, we estimate the required quantile stochastically. To do this we sample from the distribution $h(t)$ of $T|H_0$, to produce $T_1, \dots, T_k$. Depending on the alternative, we reject H_0 if the observed t is more “extreme” than $100(1 - \alpha)\%$ of the combined simulated values and the observed t .

As a simple example, we consider the one sample t test that the mean of a normal distribution is μ against the alternative that it is less than μ . If we observe a sample $X_1, \dots, X_n$, then our statistic is

$$T = \frac{\bar{X} - \mu}{\text{se}} \quad (3)$$

with

$$\text{se} = \frac{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}{n(n-1)} \quad (4)$$

the standard error of the mean. Using the classical approach, we calculate that the rejection region under the null distribution is $t < t_{\alpha, n-1}$, where $t_{\alpha, n-1}$ is the α quantile of Student's t distribution, with $n-1$ degrees of freedom.

Under a Monte Carlo approach we could sample $T_1, \dots, T_k$ from a t distribution with $n-1$ degrees of freedom. We would then reject H_0 if T is less than $T_{[\alpha]}$, where $T_{[\alpha]}$ is the α quantile of $(T_1, \dots, T_k, T)$. Thus, we use the sample from the t distribution to estimate the α quantile.

Alternatively, we can consider the problem of estimating the P value of the test. In the example given, this is equivalent to finding the value of the integral

$$\int_{-\infty}^t h(x) dx \quad (5)$$

which is consistently estimated by

$$\hat{P} = \frac{\sum_{i=1}^k I(T_i \leq t)}{k+1} \quad (6)$$

where $I(T_i \leq t)$ is the indicator function for the event $T_i \leq t$ and $\hat{P}$ is the Monte Carlo estimate.

In practice, the Monte Carlo approach to hypothesis testing is advantageous if the distribution of the chosen test statistic is intractable; however, it is convenient to sample from this distribution. Some practical examples can be found in [4].

Jöckel [5] has investigated the properties of Monte Carlo hypothesis tests under quite general conditions and produced the following conclusions. First, if we choose large enough samples (k) to produce our Monte Carlo quantities, then the results will

be equivalent to the analytic results, with high probability. Secondly, the power of the Monte Carlo test is an increasing function of k .

Bootstrap

The second example we consider is the bootstrap [6]. The bootstrap is a simple but powerful idea that is applicable to a wide range of problems. The central idea of the bootstrap is the following [7]. Suppose we wish to estimate some functional θ of a distribution $F(\beta; x)$,

$$\theta = \int g(x) dF(\beta; x) \quad (7)$$

The bootstrap estimate is found by replacing the unknown population distribution in the integral with the sample's empirical distribution, as follows:

$$\hat{\theta} = \int g(x) d\hat{F}(x) \quad (8)$$

where $\hat{F}(x)$ is the empirical distribution of the observed sample.

Now consider estimating the variability of this statistic. To estimate the variance of the sampling distribution of our estimate, we again replace the population distribution with the sample distribution to give

$$\hat{\sigma}^2 = E_{\hat{F}}[\hat{\theta} - E_{\hat{F}}(\hat{\theta})]^2 \quad (9)$$

where $\hat{\sigma}^2$ is called the *bootstrap estimate* of the sampling variation of $\hat{\theta}$ and $E_{\hat{F}}$ denotes expectation over the empirical distribution of the sample. This calculation could be performed analytically, and requires tabulating a complicated set of permutations. Instead, we can form a Monte Carlo estimate by generating draws of the same size as the original sample from the empirical distribution of the sample. For each such draw, we calculate this statistic. We iterate this procedure k times to obtain $\theta_1^*, \dots, \theta_k^*$, and then calculate

$$\hat{\sigma}^2 = \frac{1}{k} \sum_{i=1}^k (\theta_i^* - \bar{\theta}^*)^2 \quad (10)$$

where $\bar{\theta}^*$ is the mean of the bootstrap samples.

This is equivalent to drawing with replacement from the original sample, which is how the bootstrap is sometimes described.

Complex Systems

Another example of the use of Monte Carlo methods is in the analysis of complicated systems. Simulation methods are often used to explore the properties of complex systems, likelihoods, or estimators. Deterministic simulation systems are frequently used in areas such as finance where the impact of varying input parameters is assessed by changing the values in a spreadsheet [8].

Simulations involving stochastic components are often called *Monte Carlo simulations*. In many situations, uncertainty or randomness plays a key role in the dynamics of the system. Examples include queuing for operations and the demand for inventory items. Mathematical models can be proposed and analytical solutions can be sought, but in any realistic formulation the mathematics quickly becomes intractable.

Epidemic theory has been a fertile area for Monte Carlo simulations. A recent epidemic that has received considerable attention is the HIV/AIDS epidemic. Monte Carlo methods have been used to build realistic behavior patterns into a model for this epidemic and to explore its evolution based on these assumptions. Monte Carlo has also been used as an estimation technique. Often, good information exists for some aspects of the model. The other parameters of the model are varied until the results of the simulation agree best with the observed data.

An example of this methodology is given in [9] and [10]. Data from homosexual men in San Francisco provided input parameters for several aspects of the model. The model was a compartmental one with different levels of sexual activity, immigration, emigration, and death. Data on the probability of transmission related to sexual activity collected from several other studies were used to complement the primary data set.

The simulation study generated incidence projections for HIV/AIDS. A sensitivity analysis was conducted to see the range of behavior variables that were consistent with the observed incidence patterns. The conclusion was that there was a range of parameter values that could generate the observed data. The most important outcome of the analysis was the flagging of the most influential parameters for further collection of data.

Markov Chain Monte Carlo

The final example we consider is the use of Markov chain Monte Carlo (MCMC) techniques such as Gibbs sampling. These techniques use stochastic simulation to explore and summarize a multivariate density that may not be analytically tractable. This problem arises often in the area of Bayesian statistics.

With MCMC techniques, standard results [3] are used to construct Markov chains with the required stationary distributions. Popular choices are variants on the Metropolis–Hastings algorithm. From an initial state vector, the chain is simulated until the sequence reaches approximate stationarity. This is the Monte Carlo part of the algorithm. The resulting sequence is used to make stochastic approximations to any required functional of the multivariate distribution.

References

- [1] Morgan, B.J.T. (1984). *Elements of Simulation*, Chapman & Hall, London.
- [2] Ripley, B.D. (1987). *Stochastic Simulation*, John Wiley & Sons, New York.
- [3] Tanner, M. (1993). *Tools for Statistical Inference*, Springer-Verlag, New York.
- [4] Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns*, Academic Press, New York.
- [5] Jöckel, K.-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests, *Annals of Statistics* **14**, 336–347.
- [6] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics* **7**, 1–26.
- [7] Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, London.
- [8] Winston, W.L. (1991). *Operations Research Applications and Algorithms*, 2nd Edition, PWS-Kent, Boston.
- [9] Hethcote, H.W., Van Ark, J.W. & Karon, J.M. (1991). A simulation model for AIDS in San Francisco: II. Simulations, therapy and sensitivity analysis, *Mathematical Biosciences* **106**, 223–247.
- [10] Hethcote, H.W., Van Ark, J.W. & Longini, I.M. (1991). A simulation model for AIDS in San Francisco: I. Model formulation and parameter estimation, *Mathematical Biosciences* **106**, 203–222.

TERENCE J. O'NEILL, SIMON C. BARRY AND
BOREK PUZA

Article originally published in *Encyclopedia of Biostatistics*, 2nd Edition (2005, John Wiley & Sons, Ltd).

Moral Hazard

Policyholder Behavior

Providing insurance protection to an individual may lead that person to behave more carelessly than he or she did before having coverage. If the insurer cannot predict this behavior and relies on past loss data from uninsured individuals to estimate rates, the resulting premium is likely to be too low to cover losses.

Moral hazard refers to an increase in the probability of loss caused by the behavior of the policyholder. Obviously, it is extremely difficult to monitor and control behavior once a person is insured. How does one monitor carelessness? Is it possible to determine whether a person might decide to collect more on a policy than he or she deserves by making false claims?

Probability of Loss Increases after a Person Becomes Insured

The numerical example used to illustrate adverse selection can also demonstrate moral hazard. With adverse selection, the insurer cannot distinguish between good and bad risks, but the probability of a loss for each group is assumed not to change after a policy is sold. With moral hazard, the actual probability of a loss becomes higher after a person becomes insured. For example, suppose the probability of a loss increases from $p = 0.1$ before insurance to $p = 0.3$ after coverage has been purchased. If the insurance company does not know that moral hazard exists, it would sell policies at a price of \$10 to reflect the estimated actuarial loss ($0.1 \times \$100$). The actual loss would be \$30 since p has increased to 0.3. Therefore, the firm would lose \$20 ($\$30 - \10) on each policy it sells.

Mechanisms to Avoid Losses from Moral Hazards

One way to avoid the problem of moral hazard is to raise the premium to \$30 to reflect the known increase in the probability p , that occurs once a policy has

been purchased. In this case, there is *no* decrease in coverage as there was in the adverse selection example. Those individuals willing to buy coverage at a price of \$10 would still want to buy a policy at \$30 since they know that their probability of a loss with insurance would be 0.3.

Another way to avoid moral hazard is to introduce *deductibles* and *coinsurance* as part of the insurance contract. A sufficiently large deductible can act as an incentive for the insureds to continue to behave carefully after purchasing coverage because they would be forced to cover a significant portion of their loss themselves. With coinsurance, the insurer and the insured share the loss together. An 80% coinsurance clause in an insurance policy means that the insurer pays 80% of the loss (above a deductible), and the insured pays the other 20%. As with a deductible, this type of risk-sharing arrangement encourages safer behavior because those insured want to avoid having to pay for some of the losses.^a Another way to encourage safer behavior is to place upper limits on the amount of coverage an individual or enterprise can purchase. If the insurer provides only \$500,000 worth of coverage on a structure and contents worth \$1 million, then the insured knows that he or she would have to incur any residual costs of losses above \$500,000. This assumes that the insured is not able to purchase a second insurance policy for \$500,000 to supplement the first one and hence be fully protected against a loss of \$1 million, except for deductibles and insurance clauses.

Even with these clauses in an insurance contract, the insureds may still behave more carelessly with coverage than without it simply because they are protected against a large portion of the loss. For example, they may decide not to take precautionary measures that they would have adopted had they been uninsured. The cost of adopting mitigation may now be viewed as too high relative to the dollar benefits that the insured would receive from this investment. If the insurer knows in advance that an individual would be less interested in loss-reduction activity after purchasing a policy, then it can charge a higher insurance premium to reflect this increased risk or it can require specific mitigation measures as a condition of insurance. In either case, this aspect of the moral hazard problem is overcome.

End Notes

^a. For more details on deductibles and coinsurance in relation to moral hazard, see Pauly [1].

Reference

- [1] Pauly, M. (1968). The economics of moral hazard: comment, *American Economic Review* **58**, 531–536.

HOWARD KUNREUTHER

Multiattribute Modeling

Attributes, Criteria, and Objectives

In the context of decision aid or support, Roy [1] defines a *criterion* as “a tool constructed for evaluating and comparing potential actions according to a point of view that must be (as far as it is possible) well defined”. This concept is exploited in the approach to complex decision making and planning problems that has been termed *Multiple Criteria Decision Analysis* or *Aid (MCDA)*, or *Multiple Criteria Decision Making (MCDM)*. See Keeney and Raiffa [2] for early developments in this field, and Belton and Stewart [3] for a more recent and general review.

The fundamental paradigmatic basis underlying MCDA/MCDM is first to identify the appropriate criteria against which alternative courses of action or policies may be evaluated and compared in more-or-less unambiguous terms, before seeking to construct an aggregate measure of preference that integrates performance across all criteria. It is within this process that we introduce the concept of multiple attributes, or *multiattribute modeling*. Although there may not be absolute uniformity concerning terminology in the literature, it is a useful and quite common practice to adopt the following definitions:

- *Criterion*

A perspective according to which, decision alternatives may be evaluated, representing some specific interest, concern or point of view, perhaps associated with a particular stakeholder group.

- *Attribute*

An evaluation or measure of performance of a decision alternative in terms of a particular criterion, that may be used to distinguish between decision alternatives in terms of this criterion.

- *Objective*

A statement of preference in terms of an attribute, typically expressed as maximization or minimization of the numerical value of the attribute.

- *Goal*

A specific numerical value for an attribute that is deemed to represent a satisfactory level of performance in terms of the associated criterion.

For example, transportation safety may be identified as an important criterion in comparing hazardous waste disposal alternatives. An appropriate attribute might be road distance within urban areas over which material is to be transported. The associated objective would be minimization of this distance, while an associated goal may be to restrict this distance to not more than 50 km.

In the next section, we consider the process of eliciting the criteria relevant to a particular decision context by interaction with decision makers and/or other stakeholders (*see Stakeholder Participation in Risk Management Decision Making*), after which we turn to the practicalities of constructing an operationally meaningful *value tree* of attributes for purposes of decision aid and support. In the final section, we discuss the manner in which conflicting views of different stakeholder groups need to be taken into consideration.

Identification of Criteria

By definition, a criterion is a representation of human judgment regarding an issue that is important in a decision making context. Elicitation of criteria is therefore not a process of identifying a preexisting reality as is research in the natural sciences; it is a constructive process of problem structuring. A comprehensive review of problem-structuring methods is provided by Rosenhead and Mingers [4], while practicalities of facilitating the associated decision processes are discussed by Papamichail *et al.* [5].

The problem-structuring approaches discussed in the cited references have broader aims, but many have direct relevance to the identification of criteria. We have, in particular, found mind mapping concepts (e.g., the cognitive mapping of Eden and Ackermann [6], or the causal mapping described in Montibeller and Belton [7]) to be useful in eliciting criteria. This approach is illustrated in Figure 1, which displays such a map that was developed as a summary of concerns expressed by one fisher community during an investigation into the evaluation of fishing rights allocations in the Western Cape province of the Republic of South Africa (discussed by Stewart *et al.* [8]). The map connects a number of concepts by directed arcs, indicating the effects or influences of one concept on another. Concepts having incoming but no outgoing arcs are indicative of more-or-less

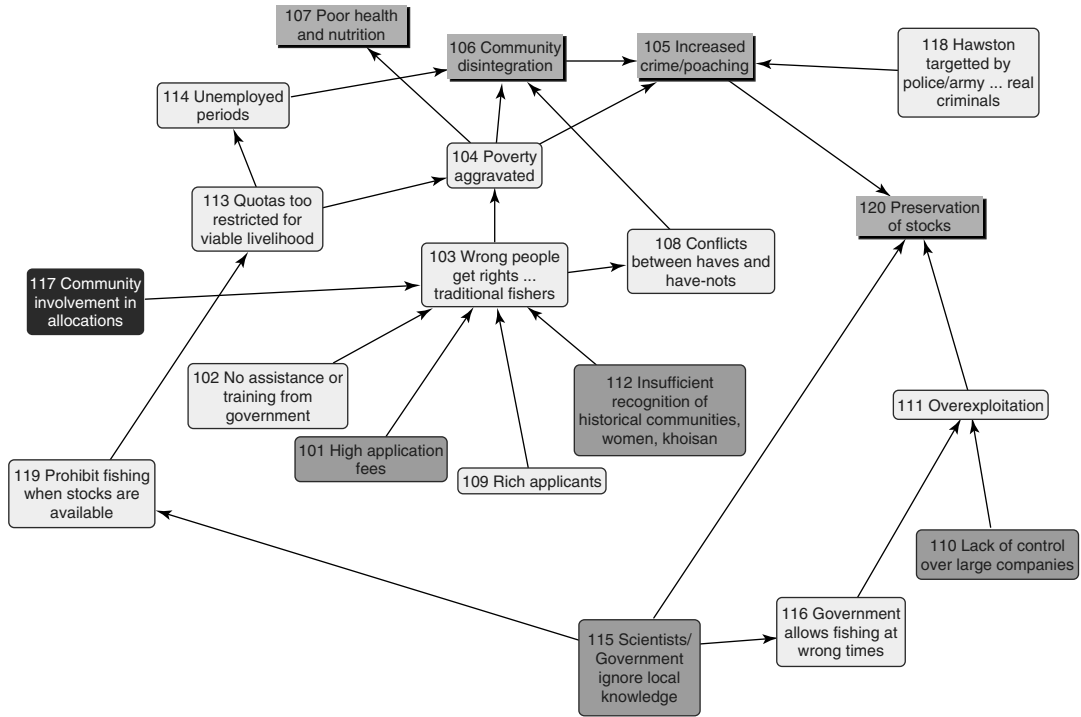


Figure 1 Illustrative causal map for identification of fundamental criteria

fundamental criteria of importance to the group. In the case illustrated in Figure 1, these criteria can be identified as health and nutrition status of the community (concept 107), social integration of the community (concept 106), crime (concept 105), and preservation of fishing stocks (concept 120). These criteria represent issues of concern to one stakeholder group, and the same process had to be repeated with other groups, and with the management of relevant government department, before a complete set of criteria could be identified.

In guiding the process of identifying criteria, the facilitator may find it useful to distinguish between what Keeney [9] terms fundamental and means-ends objectives (where in our terminology, his “objectives” correspond to our criteria and attributes). In this sense, “fundamental” refers to ultimate aims and purposes, while “means-ends” refers to more immediate objectives or goals (in the sense we have defined above) that can lead to attaining these ultimate aims or purposes. Such considerations lead to two distinct approaches to the elicitation of criteria as follows:

- *Top-down*

Start by identifying the fundamental criteria, often linked to broad concerns such as economic, social, and environmental issues. These need then to be broken down into more operational means of achieving the desired ends, until ultimately a clear set of measures of performance or achievement (the “attributes”) emerge.

- *Bottom-up*

Identify those characteristics (“attributes”) which differentiate between the available courses of action, and then consider their value relevance to fundamental aims or ends.

The facilitator may well alternate between the two approaches in working with decision makers or other stakeholders to construct a relevant set of criteria. Both approaches lead naturally to a hierarchical structure of criteria, sometimes termed the *value tree*, with an overall statement of ultimate intent at the root, progressively broken down into attribute measures against which decision alternatives can be evaluated

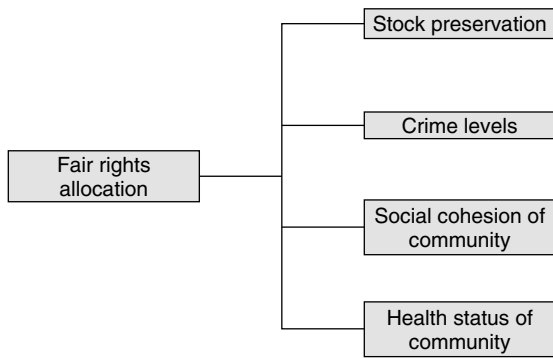


Figure 2 Value tree extracted from the causal map of Figure 1

and compared. A simple value tree extracted from the causal map of Figure 1 is illustrated in Figure 2; other examples may be found *inter alia* in Keeney [9]. The construction of such value trees is considered further in the next section.

Constructing the Value Tree

The problem-structuring process of the previous section typically requires the exercise of group value judgments by a number of decision makers and stakeholders, and thus often results in the tabling of a large number of issues, not all equally well defined operationally. For purposes of practical aid to decision making, further structure is then necessary to ensure that the resultant decision making process is both coherent and transparent. The use of a hierarchical value tree as mentioned above provides a format for clear and transparent structure that may easily be communicated between all role players. However, the criteria included in the tree, and especially the attributes at the detailed level, need to be subject to critical and careful scrutiny. The following (taken from Belton and Stewart [3], which is an expansion of an earlier list suggested by Keeney and Raiffa [2]), constitutes a useful checklist of aspects of the chosen set of criteria that need to be considered in constructing the final value tree of attributes on which decision making is to be based.

1. Value relevance

Are stakeholders able to express objectives or goals directly in terms of each attribute? Can they meaningfully state which of two decision alternatives

are preferred in terms of the associated criterion by consideration of this attribute alone (all other things assumed equal at this stage)?

2. Understandability

Do stakeholders share a common understanding of what is being measured by each attribute?

3. Measurability

Can the measure of performance of alternatives defined by the attribute be applied in a consistent manner? For example, it is well known that different people will apply a 1–5 scale in widely differing ways, unless the meaning of each point on the scale is very carefully defined, preferably with reference to external objective standards.

4. Operationally feasible

Can performance measures of alternatives in terms of each attribute be obtained or estimated with reasonable expenditure of effort, i.e., not placing excessive demands on decision makers and/or their staff.

5. Judgmental independence

Is it possible to think about trade-offs between performance on any two attributes, all other things being equal, without having to know the levels of achievement on the other attributes?

6. Nonredundancy

Do more than one attribute address essentially the same concerns, perhaps in different words or as subcriteria of different branches of the value tree (for example, employment as a component of both social and economic criteria)?

7. Balancing completeness and conciseness

The value tree should be complete, i.e., all important aspects of the problem should be captured, but should also be concise, keeping the level of detail to the minimum required.

8. Balancing simplicity and complexity

One should strive for the simplest value tree that adequately captures the essence of the decision making problem, i.e., is “requisite” in the sense defined by Phillips [10].

In some situations, there is a natural link from the criterion to a quantitative attribute. For example, in selecting between competing industrial processes,

one criterion might relate to levels of air pollution generated, which would naturally be measured in terms of concentrations of a specific set of contaminants or pollutants. Such measures would then represent the *natural attributes* at the lowest (most detailed) level of the value tree. Typically, a natural attribute creates little problem in terms of the first four aspects listed above. The attributes are, by definition, directly measurable and linked to the values of stakeholders, and are usually easily understood. In many cases, values for natural attributes are also readily available without excessive effort.

The first four aspects listed above demand greater care in the precise operational definition of the attributes when the associated criteria or objectives are of a more qualitative nature. Two approaches may then commonly be found:

- *Proxy attribute(s)*

Although there may not be a direct quantitative measure available, it may be possible to identify one of more related measures of performance that may serve as a proxy for the criterion. Belton and Stewart [3, Section 5.2], in an example of choosing the city in which to locate a new office, suggest a criterion of “availability of staff”. Such a criterion encompasses a broad range of concerns about both the range of skills available in the region, and the relationships between supply and demand in certain skills categories. It was then suggested that one surrogate or proxy measure may be the average number of applicants per advertisement placed for a particular category of post. Such an attribute cannot capture the full richness of concerns implied by the criterion, but may plausibly be correlated with many of these concerns.

Use of a proxy attribute (or a few such proxies at the lowest level of the value tree) may certainly be more operationally feasible than a complete exploration of issues, but care is necessary to ensure that the selected attribute has clear value relevance to stakeholders and decision makers.

- *Subjectively constructed attributes*

An alternative to seeking proxy attributes may be to construct a scale, by setting up verbal descriptions or reference scenarios of what would constitute different levels of performance in terms of the underlying criterion. Typically, some three to five levels are defined. Eventually, alternative policies or courses of action will then be evaluated by determining which

of the constructed scenarios most closely describes the alternative under consideration.

Belton and Stewart [3] also provide an example of a constructed scale for a “quality of life” criterion. A number of factors are identified, such as climate, pollution, safety, and cultural life, such that for each factor, any alternative location can be described as favorable, acceptable, or unfavorable. Three reference scenarios may then be defined:

- all factors are favorable.
- all factors are at least acceptable, with at most one unfavorable factor, if this is compensated by at least one favorable factor.
- there are no favorable factors, while three or more factors are unfavorable.

Two intermediate descriptions might also be provided to create a 5-point scale.

Constructed scales are more demanding in time and effort to ensure a broad level of understandability, but can be made highly value relevant.

The issue of judgmental independence (aspect 5, above) is not equally critical of all methods of multicriteria decision analysis. Nevertheless, for ease of understanding of the process, it is generally advantageous to be able to think about the importance of achieving specific levels of performance on one attribute without needing to modify such value judgments in the light of performance on other attributes. Keeney [11] has demonstrated how a redefinition of attributes can lead to greater levels preferential independence, so that judgmental independence may often be a matter of how attributes are defined (rather than a fundamental property of the problem context).

Consideration of the final three aspects listed above is a matter of judgment as to what detail of modeling is necessary for the problem at hand. Is the decision problem of sufficient importance to warrant additional effort? Are there outstanding issues that could substantially change conclusions reached?

Figure 3 is adapted from Belton and Stewart [3, page 81], and provides a more comprehensive illustration of a value tree which may finally emerge in a particular problem setting. The example was based on an earlier study by Stewart and Joubert [12] of a problem involving regional land use planning. The overall fundamental criterion was that of improving quality of life in the region; this criterion was subdivided into more fundamental concerns with social, economic, and environmental benefits. The third level represents

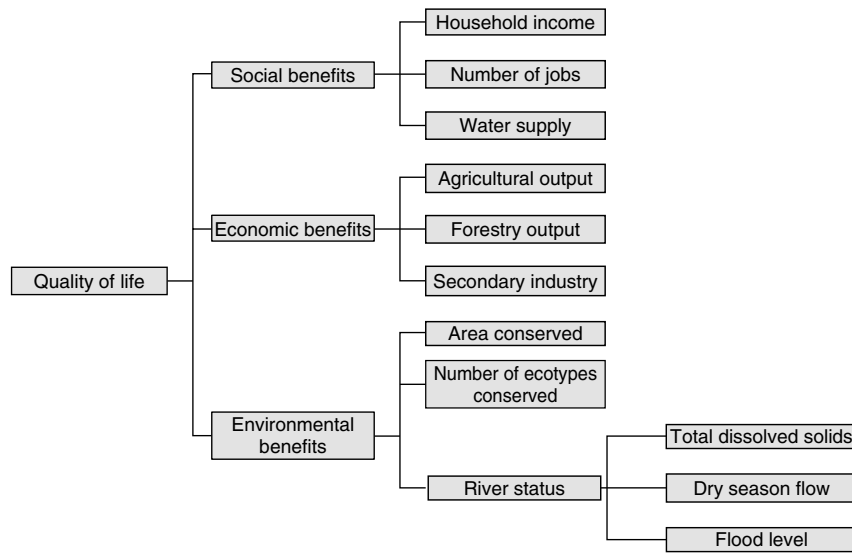


Figure 3 Illustrative value tree for regional water resources planning

more operational means to achieve the stated ends, most of which can be viewed as attributes except for “river status” which still needed to be further unpacked into the more specific attributes shown. All attributes are at this stage quantitative, although some may be viewed as proxies, for example, the number of ecotypes conserved as a proxy for biodiversity, and the two river flow measures as proxies for broader concerns with maintaining hydrological patterns.

Multiple Stakeholders

Multicriteria decision problems often arise because of the existence of a number of different interest groups. In much of the preceding discussion we have made reference to *decision makers* and *stakeholders*. In doing so, we have recognized the following characteristics of complex decision making problems:

- *Group decision making*

Corporate decisions are seldom made by an individual decision maker. The decision making process tends to be one of seeking consensus between different groups, and the techniques of MCDA are designed especially for this purpose.

- *External stakeholder or interest groups*

In addition to the previous point, there often exist a number of lobbies or pressure groups who do

not have direct responsibility for decision making, but who will attempt to influence decisions and/or take legal action to prevent unfavorable (to them) decisions from being implemented. Decision makers may well wish to consult with such groups in order, either to preempt later action or (at very least) to demonstrate that divergent interests have been taken into account.

The existence of interest and stakeholder groups does complicate the process of multiattribute modeling. Ideally, one might like to capture the concerns of all such groups in the value tree used for decision analysis. Unfortunately, this becomes operationally infeasible in some instances, as the value tree of attributes may become too large and detailed to allow for meaningful analysis. In such cases, it may be advisable to construct different value trees for different groups, for each to explore their own values and preferences, and to act as a medium of communication between groups.

References

- [1] Roy, B. (2005). Paradigms and challenges, in *Multiple Criteria Decision Analysis – State of the Art Annotated Surveys, International Series in Operations Research and Management Science*, J. Figueira, S. Greco & M. Ehrgott, eds, Springer, New York, Vol. 76, Chapter 3, pp. 3–24.

- [2] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives*, John Wiley & Sons, New York.
- [3] Belton, V. & Stewart, T.J. (2002). *Multiple Criteria Decision Analysis: An Integrated Approach*, Kluwer Academic Publishers, Boston.
- [4] Rosenhead, J. & Mingers, J. (eds) (2001). *Rational Analysis for a Problematic World Revisited*, 2nd Edition, John Wiley & Sons, Chichester.
- [5] Papamichail, K.N., Alves, G., French, S., Yang, J.B. & Snowdon, R. (2007). Facilitation practices in decision workshops, *Journal of the Operational Research Society* **58**, 614–632.
- [6] Eden, C. & Ackermann, F. (2001). SODA – the principles, in *Rational Analysis for a Problematic World Revisited*, 2nd Edition, J. Rosenhead & J. Mingers, eds, John Wiley & Sons, Chichester, pp. 21–41.
- [7] Montibeller, G. & Belton, V. (2006). Causal maps and the evaluation of decision options – a review, *Journal of the Operational Research Society* **57**, 779–791.
- [8] Stewart, T.J., Joubert, A. & Janssen, R. (2006). MCDA framework for fishing rights allocation in South Africa. Submitted for publication.
- [9] Keeney, R.L. (1992). *Value-Focused Thinking: A Path to Creative Decision Making*, Harvard University Press, Cambridge.
- [10] Phillips, L.D. (1984). A theory of requisite decision models, *Acta Psychologica* **56**, 29–48.
- [11] Keeney, R.L. (1981). Analysis of preference dependencies among objectives, *Operations Research* **29**, 1105–1120.
- [12] Stewart, T.J. & Joubert, A. (1998). Conflicts between conservation goals and land use for exotic forest plantations in South Africa, in *Multicriteria Analysis for Land Use Management*, E. Beinat & P. Nijkamp, eds, Kluwer Academic Publishers, Dordrecht, pp. 17–31.

Related Articles

Considerations in Planning for Successful Risk Communication

Group Decision

Risk and the Media

Role of Risk Communication in a Comprehensive Risk Management Approach

THEODOR J. STEWART

Multiatribute Utility Functions

General Principles and Assumptions

A *multiatribute utility function* (MAUF) is a simple mechanism for evaluating decision alternatives in the context of a *multicriteria decision making* (MCDM) problem when outcomes are subject to risk or uncertainty, in other words, the consequences of any action are *random variables*. It is assumed that the MCDM problem itself has been formulated and structured in multiatribute terms (*see Multiatribute Modeling*), and that the resulting attribute measures are fully quantified in a cardinal sense (i.e., defined on a scale such that averaging is a meaningful operation). The MAUF can then be viewed as an integration of two concepts:

- the von Neumann–Morgenstern utility function for representation of risk preferences (*see Risk Attitude*); and
- a *multiatribute value function* (MAVF) (*see Multiatribute Value Functions*), extended to ensure that the output represents a strength of preference, rather than simply an ordinal rank.

We assume now that the outcomes consequent upon the choice of a specific decision alternative (course of action) may be represented in terms of a multivariate probability distribution of the relevant attribute values. We shall adopt the following notational conventions:

- Z_i^a : The performance of an alternative a in terms of attribute i (for $i = 1, 2, \dots, m$), viewed as a random variable whose distribution depends on the chosen alternative a .
- $\mathbf{Z}^a$: The vector of random variables $Z_1^a, Z_2^a, \dots, Z_m^a$.
- $z_i^a, \mathbf{z}^a$: Specific realizations of the random variables Z_i^a and $\mathbf{Z}^a$.

Without loss of generality, we may suppose that the attributes are defined such that increasing values are always preferred. The superscripts on $Z_i^a, \mathbf{Z}^a, z_i^a$, and $\mathbf{z}^a$ may be omitted when an undefined alternative is under consideration.

The function $U(\mathbf{z})$ is an *MAUF* representing decision maker preferences if it satisfies the property that alternative a is preferred to alternative b if and only if $E[U(\mathbf{Z}^a)] \geq E[U(\mathbf{Z}^b)]$, where expectation is taken with respect to the respective multivariate distributions on $\mathbf{Z}$ implied by choice of alternatives a or b .

The questions to be addressed in this article are the appropriate choice of functional form for $U(\mathbf{z})$, and the process of eliciting parameter values within this function to represent decision maker preferences. The development of the associated multiatribute utility theory is well documented in the classic texts of Keeney and Raiffa [1] and von Winterfeldt and Edwards [2], while a practitioner-oriented summary is provided in [3].

In the next section, we examine the implications and limitations of assuming a simple additive function for $U(\mathbf{z})$. Thereafter, we consider some relaxed assumptions concerning multivariate risk preferences, and identify the forms of function that arise then. In the final two sections, we give attention to elicitation of parameter values and comment on the ranges of practical applications of the MAUF model.

Assumptions and Implications of Additive Utility Functions

In the article *Multiatribute Value Functions*, it was noted that an additive functional form could be justified on the basis of relatively weak assumptions concerning preference structures. With this result in mind, it is tempting to attempt to represent $U(\mathbf{z})$ also in additive form, i.e., to assume that $U(\mathbf{z})$ can be expressed as

$$U(\mathbf{z}) = \sum_{i=1}^m k_i u_i(z_i) \quad (1)$$

where the k_i are some form of weights, and the $u_i(z_i)$ are marginal (single attribute) utility functions.

Clearly, the $u_i(z_i)$ must be one-dimensional utility functions in the von Neumann–Morgenstern sense (see [3], for example). This assertion follows directly by considering alternatives that have identical fixed (deterministic) values on all attributes except attribute, say, i . With the assumption of the additive utility, an alternative a is preferred to an alternative b if and only if $E[u_i(Z_i^a)] \geq E[u_i(Z_i^b)]$.

In terms of the additive model, the evaluation of $E[U(\mathbf{Z})]$ depends only on the marginal probability distributions of the components of $\mathbf{Z}$, and not on any properties of their joint distributions. The additive model is thus only justifiable if decision maker preferences depend only on the marginal distributions, a property that is termed *additive independence*. In order to understand the restrictions implicit in the assumption of additive independence, consider the two-dimensional case, i.e., $m = 2$, with only two possible outcomes on each criterion, which we shall denote by $z_i^0 < z_i^1$ for $i = 1, 2$. Without loss of generality, we may then standardize the marginal utility functions such that $u_1(z_1^0) = u_2(z_2^0) = 0$ and $u_1(z_1^1) = u_2(z_2^1) = 1$. Now suppose that the two alternatives a and b generate outcomes as follows:

- *alternative a* results in one or other of the two outcomes $(z_1^0; z_2^0)$ and $(z_1^1; z_2^1)$, with equal probabilities on each; while
- *alternative b* results in one or other of the two outcomes $(z_1^0; z_2^1)$ and $(z_1^1; z_2^0)$, with equal probabilities on each.

The additive model results in an expected utility of $(k_1 + k_2)/2$ for both alternatives, implying that the decision maker should be indifferent between a and b irrespective of the specific details. Such indifference, however, may well be violated because of the following reasons:

- In some cases, the values on the two attributes may be compensatory, so that alternative b , which guarantees some gain over the worst case may well be preferred by a risk averse decision maker to alternative a .
- In other contexts, the attributes may refer to gains to different stakeholders. Under such circumstances, decision makers may prefer alternative a , which avoids the inevitable conflict that would arise when there is always one winner and one loser.

The additive model (1) may thus represent too strong an assumption in many cases. Although it has been demonstrated (e.g., Stewart [4]) that the effect of the additive assumption may be small in comparison to other errors and biases, such as in the assessment of the marginal utilities $u_i(z_i)$, models arising from less restrictive assumptions may be more defensible. These are discussed in the next two sections.

Utility Independence and Multiplicative Aggregation

In order to weaken the assumption of additive independence, we need to introduce the concept of *utility independence*. Suppose that we partition the set of attributes into two disjoint subsets, say $\mathbf{Z} = (\mathbf{Y}, \bar{\mathbf{Y}})$, where $\mathbf{Y}$ is a vector of $p < m$ attributes, and $\bar{\mathbf{Y}}$ is the complement of $\mathbf{Y}$, i.e., the elements of $\mathbf{Z}$ not included in $\mathbf{Y}$. We say that the subset of attributes $\mathbf{Y}$ is *utility independent* of its complement if the decision maker's preferences between probability distributions on $\mathbf{Y}$, given that outcomes on $\bar{\mathbf{Y}}$ are deterministically fixed, do not depend on these fixed values.

The attributes in $\mathbf{Z}$ are said to be *mutually utility independent* if every subset of the attributes in $\mathbf{Z}$ is utility independent of its complement. Note that this condition implies *inter alia* that preferences for gambles on any one attribute, say Z_i , given that all other attributes are fixed, do not depend on these fixed levels for the other attributes. It is then still operationally meaningful to establish a marginal utility function $u(z_i)$ for each attribute in the standard manner.

If the attributes in a given decision problem are mutually utility independent, then it can be shown (see for example Keeney and Raiffa [1], Section 6.3) that the aggregate utility function $U(\mathbf{z})$ is functionally related to the marginal utility functions $u_i(z_i)$ in multiplicative form defined by

$$1 + kU(\mathbf{z}) = \prod_{i=1}^m [1 + kk_i u_i(z_i)] \quad (2)$$

where the parameter k is related to the relative importance parameters k_i through

$$1 + k = \prod_{i=1}^m [1 + kk_i] \quad (3)$$

It may be confirmed that $-1 < k < \infty$ when $k_1 = k_2 = \dots = k_m$. Furthermore, the solution with $k = 0$ reduces to the additive model (1) with $\sum_{i=1}^m k_i = 1$, so equation (1) is just a special case of equation (2).

In the two-attribute example of the previous section, the expected utility based on equation (2) can be calculated to be $(k_1 + k_2 + kk_1 k_2)/2$ for alternative a , but $(k_1 + k_2)/2$ for alternative b . The multiplicative model thus no longer forces an equivalence between alternatives a and b , since a is preferred to

b for $k > 0$, but *vice versa* for $k < 0$. In this sense, the parameter k is seen to be a measure of preference for more equitable outcomes (as opposed to compensation between criteria).

Weaker Forms of Independence Assumptions

The multiplicative model arising from the assumption of mutual utility independence allows for modeling a wide range of preference structures over multiattribute outcomes under uncertainty in a relatively simple manner, requiring only the need to elicit the marginal utility functions and the m scaling parameters k_i (see section titled “Practical Application”). Nevertheless, the assumption of mutual utility independence may be quite difficult to validate fully in practice, so that it may be desirable to seek less restrictive assumptions.

If we expand both sides of equation (2) then we obtain an equivalent expression for $U(\mathbf{z})$ as follows:

$$\begin{aligned} U(\mathbf{z}) = & \sum_{i=1}^m k_i u_i(z_i) + k \sum_{i=1}^{m-1} \sum_{j=i+1}^m k_i k_j u_i(z_i) u_j(z_j) \\ & + k^2 \sum_{i=1}^{m-2} \sum_{j=i+1}^{m-1} \sum_{\ell=j+1}^m k_i k_j k_\ell u_i(z_i) u_j(z_j) u_\ell(z_\ell) \\ & + \cdots + k^{m-1} \prod_{i=1}^m k_i u_i(z_i) \end{aligned} \quad (4)$$

Clearly equation (4) is a special case of the following multilinear aggregation function:

$$\begin{aligned} U(\mathbf{z}) = & \sum_{i=1}^m k_i u_i(z_i) + \sum_{i=1}^{m-1} \sum_{j=i+1}^m k_{ij} u_i(z_i) u_j(z_j) \\ & + \sum_{i=1}^{m-2} \sum_{j=i+1}^{m-1} \sum_{\ell=j+1}^m k_{ij\ell} u_i(z_i) u_j(z_j) u_\ell(z_\ell) \\ & + \cdots + k_{12\dots m} \prod_{i=1}^m u_i(z_i) \end{aligned} \quad (5)$$

in which independent weighting factors apply to every cross product term.

In fact, for the case of $m = 2$ there is no real difference between the multiplicative and multilinear

models (the only difference being a replacement of the parameter k by $k_{12} = k k_1 k_2$). However, for $m > 2$, the models diverge, and equation (5) becomes a nontrivial extension of equation (2).

It has been shown (Keeney and Raiffa [1], Section 6.3) that a set of assumptions that are weaker than mutual utility independence are sufficient to demonstrate the validity of the more general aggregation rule given by equation (5). The required assumption is only that each individual attribute Z_i be utility independent of its complement (i.e., of all attributes except the i th). Since the utility independence of each Z_i from its complement is in any case necessary in order to construct the marginal utility functions, this assumption is about the weakest that can be made in order to form an aggregate multiattribute utility, so that equation (5) is in effect the most general model that we can find for our purposes.

Unfortunately, however, the number of parameters that need to be elicited from decision makers in order to construct the multilinear model grows rather rapidly. For the multiplicative model, we require the elicitation of precisely m parameters. For the multilinear model, however, a total of $2^m - 1$ scaling parameters are needed (giving, for example, 31 parameters when $m = 5$). These numbers rapidly become practically unmanageable (see next section).

Practical Application

The first step in the elicitation of any MAUF (whether additive, multiplicative, or multilinear) is the assessment of the marginal (single attribute) utility functions $u_i(z_i)$. This requires the same procedures as described in the article **Risk Attitude**, applied independently to each attribute.

In the case of the additive model, the weights k_i can be determined by direct consideration of deterministic trade-offs, essentially as described in the article **Multiattribute Value Functions**.

For the multiplicative model, the *relative* magnitudes of the k_i can also be assessed as for deterministic MAVFs. However, in order to determine the appropriate value for k (or, equivalently, the scaling to be applied to the k_i), it is necessary to elicit preferences between multivariate lotteries. In order to illustrate this last elicitation step, suppose that the relative weights have been established (as in the deterministic case), and standardized to give weights

$w_1, \dots, w_m$ such that $\sum_{i=1}^m w_i = 1$. It then follows that $k_i = \alpha w_i$ for some as yet unknown scaling factor α , giving $\sum_{i=1}^m k_i = \alpha$.

One approach might then be to select two attributes (which with no loss of generality can be assumed to be the first two), and to present the decision maker with a choice between two hypothetical alternatives that have identical and deterministic outcomes on attributes $3, \dots, m$, but which have uncertain outcomes on attributes 1 and 2 specified as follows (using the same definitions of z_i^0 and z_i^1 as in the earlier example, i.e., such that $u_i(z_i^0) = 0$ and $u_i(z_i^1) = 1$ for $i = 1, 2$):

- *alternative a* results in the outcome $(z_1^1; z_2^1)$ with probability p , and in the outcome $(z_1^0; z_2^0)$ with probability $1 - p$; while
- *alternative b* results in one or other of the two outcomes $(z_1^0; z_2^1)$ and $(z_1^1; z_2^0)$, with equal probabilities on each.

In this hypothetical case, the decision maker would be asked to specify the probability p that would make him/her indifferent between the two alternatives. Simple algebraic manipulation confirms that the parameters of the multiplicative model must then satisfy the following:

$$\begin{aligned} & \frac{1}{2} (1 + kk_1) + \frac{1}{2} (1 + kk_2) \\ &= p (1 + kk_1) (1 + kk_2) + 1 - p \end{aligned} \quad (6)$$

Now define $c = \alpha k$. The above equality simplifies to

$$\frac{c}{2} (w_1 + w_2) = pc (w_1 + w_2) + pc^2 w_1 w_2 \quad (7)$$

for which the only solutions are $c = 0$ or

$$c = \frac{(0.5 - p)(w_1 + w_2)}{pw_1 w_2} \quad (8)$$

Since $c = 0$ implies $k = 0$, the zero solution is only valid if the additive model applies, in which case (as we have previously seen) we must have $p = 1/2$. Thus for $p \neq 0.5$ the only solution is given by equation 8. Finally, the value k will be obtained from

$$1 + k = \prod_{i=1}^m (1 + cw_i) \quad (9)$$

Table 1 Multiattribute risk parameters as function of assessed probabilities

p	c	k
0.3	2.667	4.444
0.4	1.000	1.250
0.5	0.000	0.000
0.6	-0.667	-0.556
0.7	-1.143	-0.816

Table 2 Marginal utilities for different values of two attributes

Level	Cost		Time	
	\$ m	Utility	Mins	Utility
z_i^0	25	0	30	0
z_i^1	20	0.7	22.5	0.4
z_i^2	15	1	15	1

Illustrative Numerical Values

For the case of $m = 2$ and $w_1 = w_2$, the Table 1 indicates the values of k corresponding to a few sample values of p .

Numerical Example

We return to the example of choosing between alternative routes for a highway introduced in the article **Multiattribute Value Functions**. This example involved only two attributes, namely total investment cost (\$ million) and mean transit time (min). Suppose now that for any alternative there are only three possible outcomes, namely \$15, \$20, and \$25 million for investment cost and 15, 22.5, and 30 min for transit time, but that alternatives differ in terms of the probabilities on each of the nine possible outcomes.

Without loss of generality, utilities of 0 and 1 can be associated with the worst and best outcomes on each criterion, respectively. The marginal utility functions will thus be fully defined by assessing the relative utilities for the intermediate values on each criterion. For purposes of illustration, suppose that the marginal utilities for each attribute have been assessed as shown in Table 2 (by consideration of gambles, as discussed in the article **Risk Attitude**).

Finally, suppose that equal weights are associated with each criterion (i.e., $k_1 = k_2$).

For this simple case with $m = 2$ and $k_1 = k_2$ (i.e., $w_1 = w_2$), it is easily confirmed that equations (8) and (9) lead to $c = 2(1 - 2p)/p$, $k = c(1 + c/4)$, and $k_1 = k_2 = p$.

Consider now three alternative routes with (stochastic) outcomes described as follows:

- A. This option results in (cost = \$25 million, time = 30 min) with probability $1/8$; (cost = \$25 million, time = 22.5 min) with probability $1/8$; and (cost = \$15 million, time = 22.5 min) with probability $3/4$. The expected utility then becomes $1.1p + 0.3p^2$.
- B. This option results in two possible outcomes occurring with equal probabilities, namely (cost = \$25 million, time = 30 min) and (cost = \$15 million, time = 15 min). The expected utility is then 0.5 for any value of p (or k).
- C. This option results in the certain outcome (cost = \$25 million, cost = 15 min), which has (expected) utility of $k_2 = p$.

Table 3 gives the expected utilities for the three options for three different values of p and the associated k .

This example thus demonstrates that each of the three options can in principle be optimal, depending on the level of the decision maker's multivariate risk aversion measured by k .

Conclusions

Although the simple example of the previous section illustrates the potential for the parameter k to

substantially influence recommendations from the model, it should be evident that questions regarding preferences between multivariate lotteries is cognitively much more difficult than for univariate lotteries. The analyst would thus be advised to repeat the exercise for a number of different combinations of attributions and levels, as a check on consistency. Such an approach, however, places additional cognitive burdens on the decision maker.

The situation with the multilinear linear model is even more difficult. A minimum of $2^m - m$ comparisons between multivariate lotteries would be needed in order to assess all the parameters in equation (5), but in fact many more would be needed for purposes of consistency checks.

The complexity of the questions that need to be answered in order to elicit the parameters of multiplicative or (even worse) multilinear models means that many reported applications of MAUFs have been restricted to the additive model. Fortunately, simulation studies (e.g., Stewart [4]) have revealed that in many instances the sensitivity of results to departures from additivity are substantially smaller than sensitivity to other errors, such as in the assessment of the marginal utilities $u_i(z_i)$. Nevertheless, the analyst should be aware of the potential for biases arising from use of the additive model when undertaking sensitivity analyses.

References

- [1] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives*, John Wiley & Sons, New York.
- [2] von Winterfeldt, D. & Edwards, W. (1986). *Decision Analysis and Behavioral Research*, Cambridge University Press, Cambridge.
- [3] Goodwin, P. & Wright, G. (1997). *Decision Analysis for Management Judgement*, 2nd Edition, John Wiley & Sons, Chichester.
- [4] Stewart, T.J. (1995). Simplified approaches for multi-criteria decision making under uncertainty, *Journal of Multi-Criteria Decision Analysis* 4, 246–258.

THEODOR J. STEWART

Table 3 Expected utilities of three options at different levels of multivariate risk aversion

p	k	Options		
		A	B	C
0.3	4.444	0.45	0.5	0.3
0.5	0	0.55	0.5	0.5
0.7	−0.816	0.65	0.5	0.7

Multiatribute Value Functions

General Principles and Assumptions

A *multiatribute value function (MAVF)* is a simple mechanism for evaluating decision alternatives in the context of a *multicriteria decision making (MCDM)* problem. The MAVF is a special case of a value function (see **Value Function**), arising when the MCDM problem is formulated and structured in multiatribute terms (see **Multiatribute Modeling**).

The process of multiatribute modeling leads to a representation of decision alternatives in terms of performance measured on each of a number of identified and quantified *attributes* (although some attributes may be measured on a subjectively constructed scale). The assumption is that preferences between alternatives can, for purposes of practical decision making, be expressed entirely in terms of the associated attribute values. We adopt the following notational conventions:

$a, b, \dots$:	decision alternatives which are to be compared and/or evaluated;
$z_i(a)$:	performance of alternative a as measured on attribute i , for $i = 1, 2, \dots, m$;
z_i :	an arbitrary measure of performance in terms of attribute i ;
$\mathbf{z}, \mathbf{z}(a)$:	the m -dimensional vectors with elements z_i and $z_i(a)$ respectively.

Without loss of generality, we may suppose that the attributes are defined such that increasing values are always preferred, i.e., the alternative a is preferred to b in terms of attribute i if and only if $z_i(a) > z_i(b)$.

For most of this article, we assume that there are a finite number of alternatives, but some generalizations are discussed in the section titled “Mathematical Programming Structures”.

In principle, a *value function* $V(a)$ defined on the alternatives is such that the alternative a is preferred to alternative b ($a \succ b$) if and only if $V(a) > V(b)$. In some cases, stronger properties than simple preservation of preference order (e.g., representation of strength of preference) may be required of the value function. For purposes of the present article, it is sufficient (except in the section titled “Generalized Functional Forms”) to restrict

discussion to the simplest assumption of preservation of preference order only.

The basis of multiatribute modeling is that an alternative a may be represented fully by the associated vector of attribute values $\mathbf{z}(a)$, so that $V(a)$ can be expressed as $V[\mathbf{z}(a)]$. For this reason, the MAVF approach is to construct a function of the attribute values directly, say $V(\mathbf{z})$, which may then subsequently be applied to the specific alternatives $a, b, \dots$. The process of constructing $V(\mathbf{z})$ involves the following two distinct stages:

- construction of, what we term, *partial value functions* $v_i(z_i)$ for each attribute: by the assumption made about the direction of preference on attributes, $v_i(z_i)$ must be a monotone strictly increasing function of its argument, but we shall see that there are additional properties that need to be satisfied.
- aggregation of the partial value functions into the complete value function $V(\mathbf{z})$.

Detailed discussions of the principles underlying MAVFs may, for example, be found in Keeney and Raiffa [1] or in von Winterfeldt and Edwards [2].

In the next section, we investigate conditions under which the aggregation step is of a simple additive form. It turns out that these conditions are relatively mild, and lead to very simple and transparent processes for decision analysis. In the following section, we also describe the methods by which the components of the value function can be elicited from decision makers or other stakeholders. We then discuss a number of extensions to the basic value function model, taking into consideration mathematical programming decision structures, stronger value function properties leading to nonadditive forms, and the use of the attribute structure to deal with risk and/or time preferences.

Additive Value Functions

We consider now a simple additive aggregation, i.e., construction of an overall value function in the form

$$V(\mathbf{z}) = \sum_{i=1}^m w_i v_i(z_i) \quad (1)$$

where the parameters w_i indicate a form of importance weight on the criterion represented by attribute

i. The appeal of the model given by equation (1) is in its simplicity, which makes the processing transparent to decision makers and other stakeholders. The question may, nonetheless, arise as to whether the model is not too simplistic and reductionistic, perhaps invoking unrealistic assumptions. Surprisingly, however, the additive form is a legitimate order-preserving value function, provided that some care and discipline is exercised in selecting the attributes, defining the functions $v_i(z_i)$, and specifying the weights. Two critical issues relate to *preferential independence* of the attributes and to a required *interval scale property* of the partial value functions. Both issues can be understood in terms of simple algebraic properties of equation (1), as discussed in the next two subsections.

Preferential Independence

Suppose that two alternatives a and b differ on r attributes only, where $1 < r < m$. Let $\mathcal{D}$ be the set of attributes on which the two alternatives differ. Since, by definition, $z_i(a) = z_i(b)$ for $i \notin \mathcal{D}$, it follows that a is preferred to b if and only if

$$\sum_{i \in \mathcal{D}} w_i v_i(z_i(a)) > \sum_{i \in \mathcal{D}} w_i v_i(z_i(b)) \quad (2)$$

irrespective of the performances of these two alternatives in terms of attributes not in $\mathcal{D}$ (i.e., on which their performances are identical). It thus follows that for an additive aggregation to apply, the decision maker must be able to express meaningful trade-offs between levels of achievement on the attributes in $\mathcal{D}$, assuming that levels of achievement on the other attributes are fixed, *without* needing to be concerned with what these fixed levels of achievement are. This is the property that we term *preferential independence*, and which is a stronger form of judgmental independence alluded to in the article on multiattribute modeling.

Example 1: Suppose that three attributes relevant to choice of a water development scheme in an arid region are identified as investment cost (from taxpayers' money), person-days of access to recreational facilities, and number of invertebrate species conserved. From a political viewpoint, it may be much more difficult to justify restrictions on recreational access to achieve higher numbers of invertebrate species conserved if investment costs have been high

(i.e., high demands on taxpayers) than if these costs are low. If this is true, then person-days of access to recreational facilities and number of invertebrate species are not preferentially independent of investment cost. Some reformulation will be necessary to justify use of an additive value function model (perhaps as discussed in Keeney [3]).

The preferential independence condition amongst the attributes is thus a *necessary* condition for an additive value function (1) to be valid; in fact this condition, together with some technical assumptions regarding continuity, transitivity, and completeness of preferences, are in fact *sufficient* to prove the existence of an additive value function, i.e., there will always exist an additive model on a set of attributes satisfying preferential independence (see Keeney and Raiffa [1, Chapter 3], for a proof). Before proceeding with the use of an additive value function modeling approach in multicriteria decision analysis (MCDA), therefore, it is necessary to confirm that the criteria and associated attributes are so defined that the decision maker's values are preferentially independent amongst the modeled attributes. In the event of serious doubt about the validity of these assumptions, one should return to the structuring process.

It is perhaps important, at this stage, to differentiate between the concept of preferential dependence, and that of structural or statistical dependence between performance levels for different attributes. It may well happen that the structure of the decision space is such that improved performance in terms of one attribute is strongly associated with poor performance in terms of another. This does not, in any way, preclude the possibility that the decision maker's preferences may still satisfy the preferential independence properties.

Interval Scale Property

We have noted that the partial value functions $v_i(z_i)$ are monotone increasing in the attribute values z_i , but we have not yet otherwise specified the functional form. The preferential independence property is sufficient to show that *some* additive value function exists, but this does not imply that the additive aggregation applies to *any* arbitrary set of increasing functions $v_i(z_i)$.

It is clear that the addition or subtraction of a constant term to each of the $v_i(z_i)$ does not affect the

preference ordering implied by the resultant $V(\mathbf{z}(a))$. The addition of such a constant term simply redefines the (generally arbitrary) level of performance that is allocated a “zero value” so as to provide a reference point against which other performances are measured. Ratios of partial values will accordingly not be meaningful, in general (since addition of constants to each $v_i(z_i)$ will change their ratios), except in special circumstances under which a natural and unambiguous zero value point exists, which would preclude such addition or subtraction of constant terms. Consequently, only ratios of *differences* between the $v_i(a)$ (i.e., the relative magnitudes of differences) have any absolute meaning independent of the generally arbitrary choices of zero point and scaling. This is the defining property of an *interval scale* of preferences.

The implication of the interval scale property is best seen by looking again at trade-offs between performance levels for different attributes. For attribute i , consider two pairs of performance levels, say $z_i^1 > z_i^0$ and $z_i^3 > z_i^2$. Suppose that the corresponding value differences are equal, i.e., that $v_i(z_i^1) - v_i(z_i^0) = v_i(z_i^3) - v_i(z_i^2)$. Clearly, if an increase from z_i^0 to z_i^1 just compensates for specified losses in other attributes, then the increase from z_i^2 to z_i^3 will equally well just compensate for those same losses. In constructing the partial value functions, therefore, it is important to check that equal increments in $v_i(z_i)$ do indeed represent equal trade-offs against other attributes, i.e., that the property of *constant relative trade-offs* applies to comparisons between performance as measured on the value scales $v_i(z_i)$. These constant trade-offs are illustrated in the next example.

Example 2: Suppose that a choice needs to be made between alternative routes for a highway, and that just two attributes (criteria) have been identified: total investment cost and mean transit time. In this case, for both attributes, smaller values are preferred to larger values, so that we might formally define the algebraic attributes, for example, by

$$z_1 = 25 - \text{Investment cost in millions of dollars} \quad (3)$$

$$z_2 = 30 - \text{Transit time in minutes} \quad (4)$$

where the zero points are defined by the worst outcomes under consideration (\$25 M and 30 min). In

the next section, we look at the process of eliciting value functions, but for illustrative purposes suppose, for now, that we have established a partial value function for mean transit time as $v_2(z_2) = z_2^2$. It is easily confirmed that $v_2(9) - v_2(5) = 56$ (corresponding to transit times of 21 and 25 min respectively) and $v_2(15) - v_2(13) = 56$ (corresponding to transit times of 15 and 17 min respectively).

The implication is that whatever the value function for investment cost, if decision makers are prepared to accept an increase in investment cost from 20 to 21 million dollars (say) to reduce mean transit time from 25 to 21 min, then for any additive model to hold, they should equally well be prepared to accept an increase in investment cost from 20 to 21 million dollars to reduce mean transit time from 17 to 15 min. As the value functions are constructed, these implied trade-offs need regularly to be fed back to decision makers, to ensure that the functions do indeed (at least approximately) satisfy the interval scale property.

Elicitation of Additive Value Functions

In the previous section, we have seen that the additive value function model provided by equation (1) is simple, transparent, and yet broadly justifiable provided that certain assumptions are met. As the analyst constructs the model by eliciting the partial value functions and weights from clients (decision makers or other stakeholders), care needs to be exercised to ensure that the assumptions do hold, at least to a good level of approximation.

The process of elicitation involves two distinct steps, namely, elicitation of the individual partial value functions $v_i(z_i)$, followed by elicitation of the weights w_i . The first assumption to be verified is that of preferential independence, which can, in fact, be checked as part of both steps in the elicitation:

- In the process of constructing $v_i(z_i)$, can decision makers express judgments on the worth of different increments on the z_i scale without needing to think about performance levels on other attributes?
- In the process of estimating weights, can decision makers express trade-offs between subsets of attributes without needing to think about performance levels on other attributes?

If no serious difficulties are encountered, then we may safely assume that a satisfactory level of preferential independence exists. We now assume that this independence is satisfied as we move on the elicitation process itself.

Partial Value Functions

We require that $v_i(z_i)$ be an increasing function of z_i (recalling the assumed definition of the attribute in increasing preference terms). A further assumption which is frequently encountered in reported application of MAVFs is that a linear function can be used, i.e., $v_i(z_i) \approx az_i + b$. This is, in fact, the assumption that *constant relative trade-offs* apply to the immediately defined attributes z_i .

In constructing the partial value functions, the interval scale property needs to be verified, which may be done by systematic checks on whether constant relative trade-offs appear to hold. It has been our experience, however, that constant relative trade-offs between directly or naturally defined attributes (z_i) are perhaps the exception rather than the rule, so that the linear partial value function may typically not be valid. For example, in fisheries risk management, an important attribute is that of the population size of an endangered species. As an illustration, suppose that population sizes for one such species ranges from 0 to 5000. It might well be argued in a particular case that there is very little advantage gained (in terms of long-term survival of the species) in going from a population of 0 to say 1000. However, from that point onward, survivability is greatly enhanced by further population increases up to say 3500, at which point the species is quite safe. Clearly, therefore, increments in the population size from (say) 1500 to 2500 is much more important than either from 0 to 1000, or from 3500 to 4500. The natural attribute measurement does not then satisfy the interval scale property, and it is more likely that the sigmoidally shaped partial value function illustrated in Figure 1 will capture true value preferences (where the squares indicate points at which values may have been elicited, as discussed later).

It has been demonstrated by Stewart [4, 5] that the inappropriate assumption of linearity of the partial value function can be the major source of bias in applying MAVFs (tending to favor more extreme alternatives which are very good on some criteria and

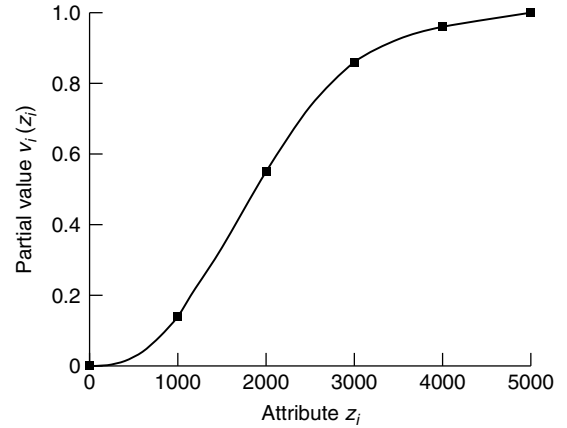


Figure 1 Illustration of a nonlinear partial value function

very poor on others, over more moderate compromises). The same work, however, has demonstrated that it is usually sufficient to assess relative values at about five different performance levels, and to interpolate between these to define the complete functions. It is also common practice when constructing categorical scales, to keep to about five categories. Typically, therefore, the construction of the partial value function requires only that relative values be assessed for a relatively small number of performance levels, which for convenience we denote by $z_i^0, z_i^1, \dots, z_i^p$, say. We describe the process on the basis of a small number of performance levels defined in this way.

To facilitate the later interpretation of the weights w_i (next subsection), it is generally preferable to scale the partial values $v_i(z_i)$ in a consistent manner. For this purpose, it is important to identify the extremes of $z_i(a)$ which are under consideration, which may be specified in either of the following ways:

- “*locally*” as the best and worst outcomes observed in the range of alternatives under consideration or
- “*globally*” as the best and worst outcomes that are conceivably ever likely to present themselves, now or in the future.

Let z_i^* and z_i^0 be the best and worst performance levels identified in this way for criterion i . It is convenient to define the partial value functions such that $v_i(z_i^0) = 0$ for each i , while $v_i(z_i^*)$ is set at some fixed value, often 1, 10, or 100.

There are two approaches that may then be adopted in constructing the value functions:

- Select a set of performance levels $z_i^0, z_i^1, \dots, z_i^p$ (often evenly spaced), followed by an elicitation from the decision maker of the relative values or worth of the increments $z_i^r - z_i^{r-1}$ for $r = 1, 2, \dots, p$ (e.g., Watson and Buede [6]).
- Assuming that the worst-case performance (z_i^0) has been specified, select an arbitrary higher level of performance z_i^1 (at about 10–20% of the range up to z_i^*); then, for $r = 2, 3, \dots$ elicit from the decision maker, the performance level z_i^r such that the increments from z_i^{r-2} to z_i^{r-1} and from z_i^{r-1} to z_i^r are of equal value, and continue until z_i^* is exceeded (e.g., von Winterfeldt and Edwards [2]).

Example 3: We illustrate the two approaches by reference to the previous example of a fish population size.

First method:

We have $z_i^0 = 0$ and $z_i^* = 5000$. Suppose that four equally spaced intermediate levels were selected to give: $z_i^1 = 1000$, $z_i^2 = 2000$, $z_i^3 = 3000$, $z_i^4 = 4000$, and $z_i^5 = 5000$. The decision maker might first be asked which of the five increments represents the greatest importance or worth toward achievement of the underlying criterion. Suppose that the second interval (from 1000 to 2000) is identified. Thereafter, the relative worths of each of the other increments (relative to the second interval) in performance level might be elicited as shown in the second column of Table:

The incremental relative values are cumulated in the third column of Table 1, and provides values of a partial value function at the end points of each interval. As indicated above, it is useful to apply a consistent scaling. For example, to obtain a 0–100 value scale, the values in the last column may be divided by 2.45 to give $v_i(1000) = 14$,

Table 1 Value assessment by comparing equal increments

Range of performance levels	Relative value of increment	Cumulative relative value
0–1000	35	35
1000–2000	100	135
2000–3000	75	210
3000–4000	25	235
4000–5000	10	245

$v_i(2000) = 55$, $v_i(3000) = 86$, $v_i(4000) = 96$, and (by definition) $v_i(5000) = 100$. These are in fact the points plotted for the value function illustrated in Figure 1.

It is often useful when interacting with decision makers using this method, to display the values on a “thermometer scale” as illustrated in Figure 2. It is our experience that decision makers relate easily to shifting the end points along the scale to get a visual impression of the value gaps being proposed.

Second method:

Start with (say) a first interval between 0 and 1000 (20% of the range). The decision maker would be asked to suggest a value x , such that the value of an increment from 1000 to x is equivalent to that of an increment from 0 to 1000. Suppose that the answer is given as $x = 1400$. Further questions of the same form would elicit ranges such as those displayed in Table 2, all being of equal worth:

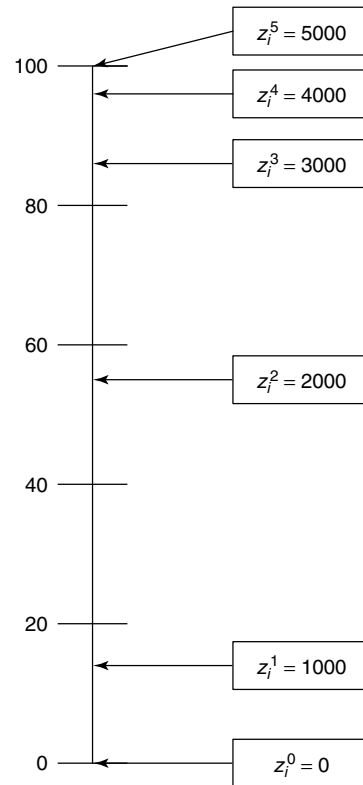


Figure 2 Illustration of a thermometer scale for value function elicitation

Table 2 Value assessment by identifying increments of equal worth

Increment		Unstandardized Cumulative value
From	To	
0	1000	1
1000	1400	2
1400	1750	3
1750	2100	4
2100	2500	5
2500	3200	6
3200	6500	7

The last column provides a relative “value” at the end points of each interval, which can be standardized as desired. There is no reason why the scale should not now be chosen to be 0–6500 even though no actual alternative reaches that far. Alternatively, an interpolation between 3200 and 6500 might be applied to give an approximate value on the unstandardized scale of about 6.5 at $z_i = 5000$. Rescaling to make $v_i(5000) = 100$ would then *pro rata* give $v_i(1000) = 15$, $v_i(1400) = 31$, $v_i(1750) = 46$, $v_i(2100) = 62$, $v_i(2500) = 77$, $v_i(3200) = 92$, and (again by definition) $v_i(5000) = 100$.

Weights

To complete the construction of the additive value function, we also need to elicit the appropriate weights w_i . It is important to recognize the algebraic meaning of the weight parameters, however. Let us suppose that we have scaled the partial value functions between the defined best and worst performance levels z_i^* and z_i^0 , so that $v_i(z_i^0) = 0$ and $v_i(z_i^*) = 100$ (say) for each attribute i . Consider now two hypothetical alternatives, a and b , that differ in terms of two attributes only, say i and j , and such that for alternative a $z_i(a) = z_i^*$, $z_j(a) = z_j^0$, while for alternative b $z_i(b) = z_i^0$, $z_j(b) = z_j^*$. It is clear then that $V(\mathbf{z}(a)) - V(\mathbf{z}(b)) = w_i - w_j$, so that a is preferred to b (implying that the “swing” from z_i^0 to z_i^* is worth more to the decision maker than the “swing” from z_j^0 to z_j^*) if and only if $w_i > w_j$. In this sense, the ratio w_i/w_j indicates the relative importance of these two “swings”.

The relative weights can also be understood in terms of trade-offs. In the example of the previous paragraph, suppose that $w_i > w_j$. Since the swing

from z_i^0 to z_i^* is worth more than the “swing” from z_j^0 to z_j^* , it should be possible to identify a smaller swing of values on attribute i that would compensate exactly for the full swing on attribute j . For example, decision makers may be asked to state a value $\hat{z}_i$ such that they are indifferent between the outcome pairs $\{\hat{z}_i(a); z_j^0\}$ and $\{z_i^0; z_j^*\}$ (again assuming equality on all other attributes). This defines a *trade-off* as the amount the decision maker would wish to gain on attribute i (i.e., $\hat{z}_i - z_i^0$), to compensate for the loss of going from z_j^* down to z_j^0 on attribute j . The indifference between the two pairs implies by the additive value function model that $w_i v_i(\hat{z}_i) + w_j \cdot 0 = w_i \cdot 0 + w_j \cdot 100$ or $w_j/w_i = v_i(\hat{z}_i)/100$.

In either interpretation, we note that it is only the weight ratios that are meaningful. The individual weights may be scaled arbitrarily, a common choice (which tends to facilitate interpretation) being to scale the weights so that $\sum_{i=1}^m w_i = 1$.

The critical observation to be made is that the algebraic meaning of weight ratios, interpreted either as swings or trade-offs, is defined by the range of attribute values under consideration. If this range is modified for one attribute, then the weight parameter must change accordingly. The implication is that weight ratios cannot meaningfully be assessed in absolute terms, independently of context. In eliciting weights, the analyst or modeler must ensure that the client (decision maker or stakeholder) takes the context into account when answering questions.

Practical means of eliciting weights in a manner consistent with the above principles are discussed in the cited references (e.g., [1, 2, 7]). It is generally convenient to have the decision maker consider not more than two to five attributes at a time. Different sets of attributes for evaluation are selected in such a way that any pair of attributes will be linked directly or indirectly through a sequence of comparisons.

Evaluation of relative weights within any one set of attributes might start with a simple ranking of the relative importance of the swings on each attribute. (For algebraic convenience, we have defined swings above by the full range from z_i^0 to z_i^* , but it is possible with appropriate algebraic manipulation to consider only portions of each range.) The following two possible approaches may then be available to the decision maker:

- to assess the relative worths of each swing or
- to specify pairwise trade-offs as defined above.

The second approach is in some senses the most precise, but decision makers are often resistant to specifying trade-offs, especially in politically sensitive areas, in which case the first approach may be found to be more acceptable.

Extensions

We conclude with brief mention of a number of extensions to the simple additive value function model as described above for discrete choice multicriteria decision problems with deterministic outcomes (i.e., no uncertainties).

Generalized Functional Forms

It is important to note that when we speak of preferential independence (together with the other technical assumptions) being sufficient to prove the *existence* of an additive value function, this result only applies to ordinal value functions, i.e., with the property that $a \succ b$ if and only if $V(a) > V(b)$. It is not necessarily true that stronger cardinal properties hold. For example, if for three alternatives $a \succ b \succ c$ we have $V(a) - V(b) > V(b) - V(c)$, it is not legitimate to claim that the strength of preference for a over b is greater than the strength of preference for b over c .

In many applications, the ordinal property is sufficient for decision making purposes. The main context in which relative strengths of preference become an important consideration is that of decision making under uncertainty or risk. Under such circumstances, each alternative is associated with a number of different possible outcomes, and it becomes necessary to consider the combined effects of many different possible trade-offs, for which some measure of strengths of preference for different possible trade-offs is required.

The extension of MAVF theory into *multiattribute utility theory* for purposes of dealing with MCDM under risk or uncertainty is discussed in a separate article. This theory applies to a large extent to the more general problem of modeling strength of preference, and will not be repeated here. We simply record the two more general aggregation models that arise as an alternative to equation (1), which emerges as a special case of either model.

Multiplicative model:

The aggregate value function $V(\mathbf{z})$ is defined by the relationship:

$$1 + kV(\mathbf{z}) = \prod_{i=1}^m [1 + kk_i v_i(z_i)] \quad (5)$$

where the $k_i > 0$ are measures of the relative importance of the associated attribute, and the parameter k is the solution to

$$1 + k = \prod_{i=1}^m [1 + kk_i] \quad (6)$$

which of necessity satisfies $k > -1$.

In the limit as $k \rightarrow 0$, we find that equation (5) tends to the additive model of equation (1) with $w_i = k_i$. In general, although the k_i can still in a sense be viewed as “importance weights”, they can no longer be arbitrarily scaled, in contrast to the weights w_i the additive multiattribute value model. Note, however, that equation (6) can be expressed as follows:

$$1 + k = 1 + k \sum_{i=1}^m k_i + k^2 \sum_{i=1}^{m-1} \sum_{j=i+1}^m k_i k_j + \cdots + k^m k_1 k_2 \dots k_m \quad (7)$$

so that if $\sum_{i=1}^m k_i = 1$, k must satisfy the equation

$$0 = k^2 \sum_{i=1}^{m-1} \sum_{j=i+1}^m k_i k_j + \cdots + k^m k_1 k_2 \dots k_m \quad (8)$$

for which the only solution is $k = 0$, giving the additive model with standardized weights.

Multilinear model:

For $k \neq 0$, the multiplicative model of equation (5) can alternatively be expressed in the following form:

$$V(\mathbf{z}) = \sum_{i=1}^m k_i v_i(z_i) + k \sum_{i=1}^{m-1} \sum_{j=i+1}^m k_i k_j v_i(z_i) v_j(z_j) + \cdots + k^{m-1} k_1 k_2 \dots k_m \times v_1(z_1) v_2(z_2) \dots v_m(z_m) \quad (9)$$

A generalization of the above, which can be shown to invoke weaker assumptions concerning preference structures, is thus the following *multilinear* aggregation:

$$V(\mathbf{z}) = \sum_{i=1}^m k_i v_i(z_i) + \sum_{i=1}^{m-1} \sum_{j=i+1}^m k_{ij} v_i(z_i) v_j(z_j) + \cdots + k_{12\dots m} v_1(z_1) v_2(z_2) \dots v_m(z_m) \quad (10)$$

Note that, whereas the multiplicative model only requires the elicitation of relative values for the k_i and the appropriate scaling (by consideration of multidimensional trade-offs), the multilinear model requires the elicitation of a much larger number of parameters, which is seldom feasible in a practical sense.

Mathematical Programming Structures

We have been assuming that the basic problem is that of choosing between a finite number of decision alternatives, which is perhaps the context within which multiattribute value modeling is most often used. There is, nevertheless, no fundamental reason for restricting application to the finite case. The construction of the MAVF is essentially independent of the decision space.

Suppose then that we have a multiobjective (vector) optimization problem of the form: maximize $z_1(\mathbf{x}), z_2(\mathbf{x}), \dots, z_m(\mathbf{x})$, subject to $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n$. An MAVF $V(\mathbf{z})$ can still be constructed as previously described, so that the vector optimization can be replaced by the conventional scalar optimization problem: maximize $V(\mathbf{z}(\mathbf{x}))$ subject to $\mathbf{x} \in \mathbb{X}$, which can, in principle, be solved by standard constrained maximization problems.

An interesting special case is that of *multiobjective linear programming (MOLP)*, in which each of the $z_i(\mathbf{x})$ functions are linear in $\mathbf{x}$, and the feasible space $\mathbb{X}$ is defined by a set of linear constraints on the $\mathbf{x}$. linear programming (LP) is a powerful and widely used tool of management science, for which very efficient numerical algorithms exist, and it is extremely important to be able to extend these algorithms to provide efficient means of solving the MOLP problem. Multiattribute value function modeling converts the MOLP into maximization of a generally nonlinear scalar function subject to linear constraints. However, if the additive form is used, and

the partial value functions $v_i(z_i)$ are approximated by a piecewise linear form, then the problem is still soluble using standard LP solvers, although possibly requiring the introduction of binary integer variables to keep track of which segments are active.

As we have indicated earlier, the practical elicitation of partial value functions is typically carried out by consideration of a small number of discrete values for the attribute, with subsequent interpolation of the value function between these points. It can be argued that piecewise linear interpolation is no less justifiable than arbitrarily smoothed functions, and it has been shown that for purposes of decision analysis there is no practical distinction between use of a smooth curve and a piecewise linear form with as few as four segments (e.g., Stewart [5]). There is, therefore, very little loss in generality in applying piecewise linear value function modeling to exploit the power of standard LP software in solving MOLP problems.

Treatment of Risk

The previous discussion has assumed that consequences of decisions are known deterministically. Many, if not most, real-world problems include some level of risk or uncertainty of outcome. If such uncertainties can be described in probabilistic terms, then it is tempting to base decisions on the expectation of $V(a)$ for each alternative. This takes us into the realm of multiattribute utility theory, however, which is discussed in a separate article. It is demonstrated there that the expectation of an additive value function may often not be a justifiable approach, leading to consideration of the more complicated functions given by equation (5) or (10) which create practical problems of elicitation.

As an alternative to the use of expectations, it is possible to define *risk attributes*, i.e., measures of performance related to risks. Such an approach is familiar even for fundamentally single objective problems. For example, the standard Markowitz portfolio theory (cf. Jia and Dyer [8] or Goodwin and Wright [9, pp. 196–199]) represents a risky single objective (monetary reward) in terms of what are effectively two nonstochastic measures, namely, expectation and standard deviation of returns. In this sense, a single attribute decision problem under uncertainty is structured as a deterministic two-attribute decision problem. The extension to risk components for each of number of fundamental criteria

is obvious (see, for example, Millet and Wedley [10, p. 104], although this is in the context of analytic hierarchy process (AHP) rather than conventional multiattribute value theory).

The important question relates to that of identifying suitable attributes to model risk preferences in a value function model. The use of standard deviation as a measure of risk is based on special assumptions concerning either the underlying probability distribution or on the structure of risk preferences. More general measures of risk are discussed by Sarin and Weber [11] and by Jia and Dyer [8]. Other risk measures that might be included as attributes in a multiattribute value model include probabilities of occurrence of certain undesirable outcomes or quantiles of the underlying distribution.

Consider, for example, the problem of managing fisheries. An important consideration is often the risk that one or more fish populations may collapse, with serious ecological and economic consequences. A complete analysis of all possible futures resulting from a given policy may not be a practical approach. It may, nevertheless, be possible to assess probabilities that each stock falls to a critical level within the next 20 years. Such probabilities may then serve as useful attributes for MCDA, supplemented perhaps by the expectation and standard deviation of financial returns from the fishery. (See, for example, Stewart [12] for a discussion of related issues in this context.)

Time Values

Similar considerations to that of risk apply also to the problem of costs and benefits being realized at different points in time. Conventional economic analysis tends to apply geometric discount factors to future values. While there is justification for this approach when such costs or benefits are clearly financial, questions need to be raised for some of the nonfinancial measures (e.g., environmental and social impacts) that tend to arise frequently in multiattribute modeling. In such cases, there is no plausible banking system that can invest or withdraw benefits at arbitrary points in time, which is the fundamental basis for geometric discounting. Prelec and Loewenstein [13], for example, discuss a number of different functional forms for the weighting to be applied to outcomes at different points in time, derived from different

axiomatic assumptions regarding preferences over time.

A complete analysis of the appropriate discounting factors to be applied to each attribute identified is thus unlikely to be realistically a practical option, while the simplistic collapsing of future streams of costs or benefits by geometric weighting may not be justifiable for nonfinancial attributes. Much as with the risk considerations in the previous section, it is possible to define outcomes at different points in time as individual attributes in their own right, and to apply the MAVF to all of the resultant attributes.

For example, in consideration of environmental impacts from water resources planning, the base attributes may be identified as minimum flow levels during the year and pollution measured by total dissolved solids. Different policies may well generate different profiles for these measures over time. Each of the two attributes may then be subdivided into the levels of these measures after 5, 20, and 50 years, say. With little additional effort, partial value functions for levels at each point in time and relative weights can be elicited as previously described.

References

- [1] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives*, John Wiley & Sons, New York.
- [2] von Winterfeldt, D. & Edwards, W. (1986). *Decision Analysis and Behavioral Research*, Cambridge University Press, Cambridge.
- [3] Keeney, R.L. (1981). Analysis of preference dependencies among objectives, *Operations Research* **29**, 1105–1120.
- [4] Stewart, T.J. (1993). Use of piecewise linear value functions in interactive multicriteria decision support: a Monte Carlo study, *Management Science* **39**, 1369–1381.
- [5] Stewart, T.J. (1996). Robustness of additive value function methods in MCDM, *Journal of Multi-Criteria Decision Analysis* **5**, 301–309.
- [6] Watson, S.R. & Buede, D.M. (1987). *Decision Synthesis. The Principles and Practice of Decision Analysis*, Cambridge University Press.
- [7] Belton, V. & Stewart, T.J. (2002). *Multiple Criteria Decision Analysis: An Integrated Approach*, Kluwer Academic Publishers, Boston.
- [8] Jia, J. & Dyer, J.S. (1996). A standard measure of risk and risk-value models, *Management Science* **42**, 1691–1705.
- [9] Goodwin, P. & Wright, G. (1997). *Decision Analysis for Management Judgement*, 2nd Edition, John Wiley & Sons, Chichester.

- [10] Millet, I. & Wedley, W.C. (2002). Modelling risk and uncertainty with the analytic hierarchy process, *Journal of Multi-Criteria Decision Analysis* **11**, 97–107.
- [11] Sarin, R.K. & Weber, M. (1993). Risk-value models, *European Journal of Operational Research* **70**, 135–149.
- [12] Stewart, T.J. (1998). Measurements of risk in fisheries management, *ORION* **14**, 1–15.
- [13] Prelec, D. & Loewenstein, G. (1991). Decision making over time and uncertainty: a common approach, *Management Science* **37**, 770–786.

Related Articles

Decision Analysis

Group Decision

Players in a Decision

THEODOR J. STEWART

Multistate Models for Life Insurance Mathematics

The Scope of This Survey

Fetching its tools from a number of disciplines, actuarial science is defined by the direction of its applications rather than by the models and methods it makes use of. However, there exists at least one stream of actuarial science that presents problems and methods sufficiently distinct to merit status as an independent area of scientific inquiry – Life insurance mathematics. This is the area of applied mathematics that studies risk associated with life and pension insurance, and methods for management of such risk. Under this strict definition, its models and methods are described as they present themselves today, sustained by contemporary mathematics, statistics, and computation technology. No attempt is made to survey its two millenniums of past history, its vast collection of techniques that were needed before the advent of scientific computation, or the countless details that arise from its practical implementation to a wide range of products in a strictly regulated industry. For a brief account of statistics for multistate life history models, the reader is referred to [1].

Basic Notions of Payments and Interest

In the general context of financial services, consider a financial contract between a company and a customer. The terms and conditions of the contract generate a stream of payments, benefits from the company to the customer less contributions from the customer to the company. They are represented by a payment function

$$B(t) = \text{total amount paid in the time interval } [0, t] \quad (1)$$

$t \geq 0$, time being reckoned from the time of inception of the policy. This function is taken to be right-continuous, $B(t) = \lim_{\tau \searrow t} B(\tau)$, with existing left-limits, $B(t-) = \lim_{\tau \nearrow t} B(\tau)$. When different from 0, the jump $\Delta B(t) = B(t) - B(t-)$ represents a lump sum payment at time t . There may also be payments due continuously at rate $b(t)$ per time unit at any

time t . Thus, the payments made in any small time interval $[t, t + dt)$ total

$$dB(t) = b(t) dt + \Delta B(t) \quad (2)$$

Payments are currently invested (contributions deposited and benefits withdrawn) into an account that bears interest at rate $r(t)$ at time t . The balance of the company's account at time t , called the *retrospective reserve*, is the sum of all past and present payments accumulated with interest,

$$\mathcal{U}(t) = \int_{0-}^t e^{\int_{\tau}^t r(s) ds} d(-B)(\tau) \quad (3)$$

where $\int_{0-}^t = \int_{[0, t]}$. Upon differentiating this relationship, one obtains the dynamics

$$d\mathcal{U}(t) = \mathcal{U}(t) r(t) dt - dB(t) \quad (4)$$

showing how the balance increases with interest earned on the current reserve and net payments into the account.

The company's discounted (strictly) future net liability at time t is

$$\mathcal{V}(t) = \int_t^n e^{-\int_t^{\tau} r(s) ds} dB(\tau) \quad (5)$$

Its dynamics,

$$d\mathcal{V}(t) = \mathcal{V}(t) r(t) dt - dB(t) \quad (6)$$

shows how the debt increases with interest and decreases with net redemption.

Valuation of Financial Contracts

At any time t , the company has to provide a *prospective reserve* $V(t)$ that adequately meets its future obligations under the contract. If the future payments and interest rates are known at time t (e.g., a fixed plan savings account in an economy with fixed interest), then so would the present value in equation (5), and an adequate reserve would be $V(t) = \mathcal{V}(t)$. For the contract to be economically feasible, no party should profit at the expense of the other, so the value of the contributions must be equal to the value of the benefits at the outset:

$$\Delta B(0) + V(0) = 0 \quad (7)$$

If future payments and interest rates would be uncertain so that $\mathcal{V}(t)$ is unknown at time t , then reserving must involve some principles beyond those of mere accounting. A clear-cut case is when the

payments are *derivatives* (see **Enterprise Risk Management (ERM); Actuary; Premium Calculation and Insurance Pricing; Default Correlation**) (functions) of certain assets (securities), one of which is a money account with interest rate r . Principles for valuation of such contracts are delivered by modern financial mathematics, see [2]. We describe them briefly in lay terms. Suppose there is a finite number of basic assets that can be traded freely, in unlimited positive or negative amounts (taking long or short positions) and without transaction costs. An investment *portfolio* is given by the number of shares held in each asset at any time. The size and the composition of the portfolio may be dynamically changed through sales and purchases of shares. The portfolio is said to be *self-financing* if, after its initiation, every purchase of shares in some assets is fully financed by selling shares in some other assets (the portfolio is currently rebalanced, with no further infusion or withdrawal of capital). A self-financing portfolio is an *arbitrage* if the initial investment is negative (the investor borrows the money) and the value of the portfolio at some later time is nonnegative with probability one. A fundamental requirement for a market to be well functioning is that it does not admit arbitrage opportunities. A financial claim that depends entirely on the prices of the basic securities is called a financial *derivative*. Such a claim is said to be *attainable* if it can be perfectly reproduced by a self-financing portfolio. The no arbitrage regime dictates that the price of a financial derivative must at any time be the current value of the self-financing portfolio that reproduces it. It turns out that the price at time t of a stream of attainable derivative payments B is of the form

$$V(t) = \tilde{\mathbb{E}}[\mathcal{V}(t)|\mathcal{G}_t] \quad (8)$$

where $\tilde{\mathbb{E}}$ is the expected value under a so-called equivalent martingale measure derived from the price processes of the basic assets, and $\mathcal{G}_t$ represent all market events observed up to and including time t . Again, the contract must satisfy equation (7), which determines the initial investment $-\Delta B(0)$ needed to buy the self-financing portfolio. The financial market is *complete* (see **Equity-Linked Life Insurance; Risk Measures and Economic Capital for (Re)insurers; Weather Derivatives; Mathematical Models of Credit Risk**) if every financial derivative is attainable. If the market is

not complete, there is no unique price for every derivative, but any pricing principle obeying the no arbitrage requirement must be of the form (8).

Valuation of Life Insurance Contracts by the Principle of Equivalence

Characteristic features of life insurance contracts are, firstly, that the payments are contingent on uncertain individual life-history events (largely unrelated to market events) and, secondly, that the contracts are long term and binding to the insurer. Therefore, there exists no liquid market for such contracts and they cannot be valued by the mere principles of accounting and finance described above. Pricing of life insurance contracts and management of the risk associated with them are the paramount issues of life insurance mathematics.

The paradigm of traditional life insurance is the so-called principle of equivalence (see **Insurance Applications of Life Tables**), which is based on the notion of risk diversification in a large portfolio under the assumption that the future development of interest rates and other relevant economic and demographic conditions are known. Consider a portfolio of m contracts currently in force. Switching to calendar time, denote the payment function of contract No. i by B^i and denote its discounted future net liabilities at time t by $\mathcal{V}^i(t) = \int_t^\tau e^{-\int_t^\tau r(s)ds} dB^i(\tau)$. Let $\mathcal{H}_t$ represent the life history by time t (all individual life-history events observed up to and including time t) and introduce $\xi_i = \mathbb{E}[\mathcal{V}^i(t)|\mathcal{H}_t]$, $\sigma_i^2 = \text{Var}[\mathcal{V}^i(t)|\mathcal{H}_t]$, and $s_m^2 = \sum_{i=1}^m \sigma_i^2$. Assume that the contracts are independent and that the portfolio grows in a balanced manner such that s_m^2 goes to infinity without being dominated by any single term σ_i^2 . Then, by the central limit theorem, the standardized sum of total discounted liabilities converges in law to the standard normal distribution as the size of the portfolio increases:

$$\frac{\sum_{i=1}^m (\mathcal{V}^i(t) - \xi_i)}{s_m} \xrightarrow{L} N(0, 1) \quad (9)$$

Suppose the company provides individual reserves given by

$$V^i(t) = \xi_i + \epsilon \sigma_i^2 \quad (10)$$

for some $\epsilon > 0$. Then

$$\begin{aligned} & \mathbb{P} \left[\sum_{i=1}^m V^i(t) - \sum_{i=1}^m \mathcal{V}^i(t) > 0 \middle| \mathcal{H}_t \right] \\ &= \mathbb{P} \left[\frac{\sum_{i=1}^m (\mathcal{V}^i(t) - \xi_i)}{s_m} < \epsilon s_m \middle| \mathcal{H}_t \right] \end{aligned} \quad (11)$$

and it follows from equation (9) that the total reserve covers the discounted liabilities with a (conditional) probability that tends to 1. Similarly, taking $\epsilon < 0$ in equation (10), the total reserve covers discounted liabilities with a probability that tends to 0. The benchmark value $\epsilon = 0$ defines the *actuarial principle of equivalence*, which for an individual contract reads (dropping the topscript i):

$$V(t) = \mathbb{E} \left[\int_t^n e^{-\int_t^\tau r(s) ds} dB(\tau) \middle| \mathcal{H}_t \right] \quad (12)$$

In particular, for given benefits the premiums should be designed so as to satisfy equation (7).

Life and Pension Insurance Products

Consider a life insurance policy issued at time 0 for a finite term of n years. There is a finite set of possible states of the policy, $\mathcal{Z} = \{0, 1, \dots, J\}$, 0 being the initial state. Denote the state of the policy at time t by $Z(t)$. The uncertain course of the policy is modeled by taking Z to be a stochastic process. Regarded as a function from $[0, n]$ to $\mathcal{Z}$, Z is assumed to be right-continuous, with a finite number of jumps, and commencing from $Z(0) = 0$. We associate with the process Z the *indicator processes* I_g and *counting processes* N_{gh} defined, respectively, by $I_g(t) = 1[Z(t) = g]$ (1 or 0 according as the policy is in the state g or not at time t) and $N_{gh}(t) = \sharp\{\tau; Z(\tau-) = g, Z(\tau) = h, \tau \in (0, t]\}$ (the number of transitions from state g to state h ($\neq g$) during the time interval $(0, t]$).

The payments B generated by an insurance policy are typically of the form

$$dB(t) = \sum_g I_g(t) dB_g(t) + \sum_{g \neq h} b_{gh}(t) dN_{gh}(t) \quad (13)$$

where each B_g is a payment function specifying payments due during sojourns in state g (a general *life annuity*) (see **Asset–Liability Management for Life Insurers; Equity-Linked Life Insurance; Insurance Applications of Life Tables; Longevity Risk and Life Annuities**), and each b_{gh} specifying lump sum payments due upon transitions from state g to state h (a general *life assurance*). When different from 0, $\Delta B_g(t)$ represents a lump sum (general *life endowment*) payable in state g at time t . Positive amounts represent benefits and negative amounts represent premiums. In practice premiums are only of annuity type.

Figure 1 shows a flow-chart for a policy on a single life with payments dependent only on survival and death. We list the basic forms of benefits: an n -year *term insurance* (see **Options and Guarantees in Life Insurance; Risk Classification/Life**) with sum insured 1 payable immediately upon death, $b_{01}(t) = 1[t \in (0, n)]$; an n -year *life endowment* with sum 1, $\Delta B_0(n) = 1$; An n -year *life annuity* payable annually in arrears, $\Delta B_0(t) = 1, t = 1, \dots, n$; An n -year *life annuity* payable continuously at rate 1 per year, $b_0(t) = 1[t \in (0, n)]$; An $(n - m)$ -year *annuity* deferred in m years payable continuously at rate 1 per year, $b_0(t) = 1[t \in (m, n)]$. Thus, an n -year *term insurance* with sum insured b against premium payable continuously at rate c per year is given by $dB(t) = b dN_{01}(t) - c I_0(t) dt$ for $0 \leq t < n$ and $dB(t) = 0$ for $t \geq n$.

The flow-chart in Figure 2 is apt to describe a single-life policy with payments that may depend on the state of health of the insured. For instance, an n -year *endowment insurance* (a combined term insurance and life endowment) with sum insured b , against premium payable continuously at rate c while active (waiver of premium during disability), is given by $dB(t) = b (dN_{02}(t) + dN_{12}(t)) - c I_0(t) dt$,

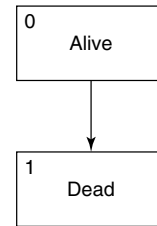


Figure 1 A single-life policy with two states

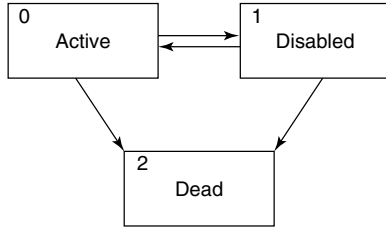


Figure 2 A single-life policy with three states

$0 \leq t < n$, $\Delta B(n) = b(I_0(n) + I_1(n))$, $dB(t) = 0$ for $t > n$.

The flow-chart in Figure 3 is apt to describe a multilife policy involving three lives called x , y , and z . For instance, an n -year insurance with sum b payable upon the death of the last survivor against premium payable as long as all three are alive is given by $dB(t) = b(dN_{47}(t) + dN_{57}(t) + dN_{67}(t)) - cI_0(t) dt$, $0 \leq t < n$, $dB(t) = 0$ for $t \geq n$.

The Markov Chain Model for the Policy History

The breakthrough of stochastic processes in life insurance mathematics was marked by Hoem's 1969 paper [3], where the process Z was modeled as a

time-continuous Markov chain. The Markov property means that the future course of the process is independent of its past if the present state is known: For $0 < t_1 < \dots < t_q$ and $j_1, \dots, j_q$ in $\mathcal{Z}$,

$$\begin{aligned} \mathbb{P}[Z(t_q) = j_q \mid Z(t_p) = j_p, \quad p = 1, \dots, q-1] \\ = \mathbb{P}[Z(t_q) = j_q \mid Z(t_{q-1}) = j_{q-1}] \end{aligned} \quad (14)$$

It follows that the *simple transition probabilities*,

$$p_{gh}(t, u) = \mathbb{P}[Z(u) = h \mid Z(t) = g] \quad (15)$$

determine the finite-dimensional marginal distributions through

$$\begin{aligned} \mathbb{P}[Z(t_p) = j_p, \quad p = 1, \dots, q] \\ = p_{0j_1}(0, t_1)p_{j_1j_2}(t_1, t_2) \cdots p_{j_{q-1}j_q}(t_{q-1}, t_q) \end{aligned} \quad (16)$$

hence they also determine the entire probability law of the process Z . It is moreover assumed that, for each pair of states $g \neq h$ and each time t , the limit

$$\mu_{gh}(t) = \lim_{u \searrow t} \frac{p_{gh}(t, u)}{u - t} \quad (17)$$

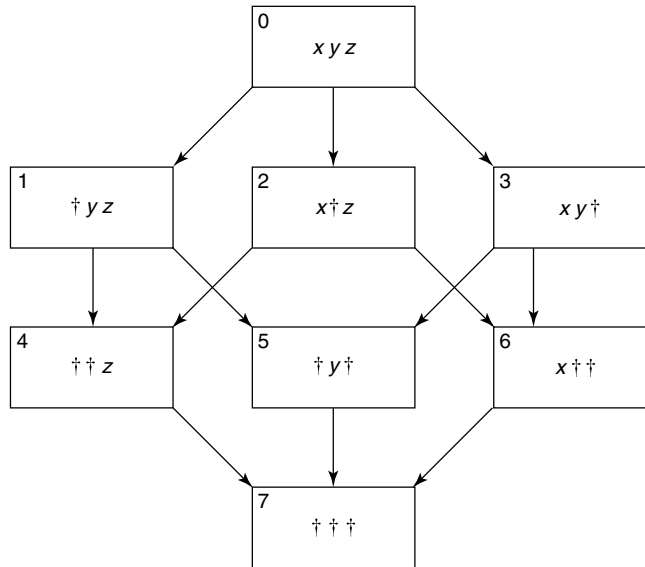


Figure 3 A policy involving three lives x , y , z . An expired life is replaced by a dagger $\dagger$

exists. It is called the *intensity of transition* from state g to state h at time t . In other words,

$$p_{gh}(t, u) = \mu_{gh}(t) dt + o(dt) \quad (18)$$

where $o(dt)$ denotes a term such that $o(dt)/dt \rightarrow 0$ as $dt \rightarrow 0$. The intensities, being one dimensional and easy to interpret as “instantaneous conditional probabilities of transition per time unit”, are the basic entities in the probability model. They determine the simple transition probabilities uniquely as solutions to sets of differential equations. The *Kolmogorov backward differential equations* for the $p_{jg}(t, u)$, seen as functions of $t \in [0, u]$ for fixed g and u , are

$$\frac{\partial}{\partial t} p_{jg}(t, u) = - \sum_{k: k \neq j} \mu_{jk}(t) (p_{kg}(t, u) - p_{jg}(t, u)) \quad (19)$$

with side conditions $p_{gg}(u, u) = 1$ and $p_{jg}(u, u) = 0$ for $j \neq g$. The *Kolmogorov forward equations* for the $p_{gj}(s, t)$, seen as functions of $t \in [s, n]$ for fixed g and s , are

$$\frac{\partial}{\partial t} p_{gj}(s, t) = \sum_{i: i \neq j} p_{gi}(s, t) \mu_{ij}(t) - p_{gj}(s, t) \mu_{j \cdot}(t) \quad (20)$$

with obvious side conditions at $t = s$. The forward equations are sometimes the more convenient because, for any fixed g, s, t , the functions $p_{gj}(s, t)$, $j = 0, \dots, J$, are probabilities of disjoint events and therefore sum to 1. A technique for obtaining such differential equations is sketched in the section titled “Actuarial Analysis of Standard Insurance Products”.

A differential equation for the *sojourn probability*,

$$p_{\overline{gg}}(t, u) = \mathbb{P}[Z(\tau) = g, \tau \in (t, u) | Z(t) = g] \quad (21)$$

is easily put up and solved to give

$$p_{\overline{gg}}(t, u) = e^{-\int_t^u \mu_{g \cdot}(s) ds} \quad (22)$$

where $\mu_{g \cdot}(t) = \sum_{h, h \neq g} \mu_{gh}(t)$ is the *total intensity of transition* out of state g at time t . To see that the intensities govern the probability law of the process Z , consider a fully specified path of Z , starting from the initial state $g_0 = 0$ at time $t_0 = 0$, sojourning there

until time t_1 , making a transition from g_0 to g_1 in $[t_1, t_1 + dt_1)$, sojourning there until time t_2 , making a transition from g_1 to g_2 in $[t_2, t_2 + dt_2)$, and so on until making its final transition from g_{q-2} to g_{q-1} in $[t_{q-1}, t_{q-1} + dt_{q-1})$, and sojourning there until time $t_q = n$. The probability of this elementary event is a product of sojourn probabilities and infinitesimal transition probabilities, hence a function only of the intensities:

$$\begin{aligned} & e^{-\int_{t_0}^{t_1} \mu_{g_0 \cdot}(s) ds} \mu_{g_0 g_1}(t_1) dt_1 e^{-\int_{t_1}^{t_2} \mu_{g_1 \cdot}(s) ds} \\ & \times \mu_{g_1 g_2}(t_2) dt_2 \dots e^{-\int_{t_{q-1}}^{t_q} \mu_{g_{q-1} \cdot}(s) ds} \\ & = e^{\sum_{p=1}^{q-1} \ln \mu_{g_{p-1} g_p}(t_p) - \sum_{p=1}^q \sum_{h: h \neq g_{p-1}} \int_{t_{p-1}}^{t_p} \mu_{g_{p-1} h}(s) ds} \\ & \times dt_1 \dots dt_{q-1} \end{aligned} \quad (23)$$

Actuarial Analysis of Standard Insurance Products

The bulk of existing life insurance mathematics deals with the situation where the functions B_g and b_{gh} depend only on the life history of the individual(s) covered under the policy. We refer to such products as *standard*. Moreover, interest rates and intensities of transition are assumed to be deterministic (known at time zero).

We consider, first, simple products with payments $dB_g(t)$ and $b_{gh}(t)$ depending only on the policy duration t (as the notation indicates). Then, with “memoryless” payments and policy process (the Markov assumption), the reserve in equation (12) is a function of the time t and the current policy state $Z(t)$ only. Therefore, we need only determine the state-wise reserves

$$V_j(t) = \mathbb{E} \left[\int_t^n e^{-\int_t^\tau r(s) ds} dB(\tau) \middle| Z(t) = j \right] \quad (24)$$

Inserting equation (13) into equation (24) and using the obvious relationships

$$\mathbb{E} [I_g(\tau) | Z(t) = j] = p_{jg}(t, \tau) \quad (25)$$

$$\mathbb{E} [dN_{gh}(\tau) | Z(t) = j] = p_{jg}(t, \tau) \mu_{gh}(\tau) d\tau \quad (26)$$

we obtain

$$V_j(t) = \int_t^n e^{-\int_t^\tau r(s) ds} \sum_g p_{jg}(t, \tau) \times \left(dB_g(\tau) + \sum_{h; h \neq g} \mu_{gh}(\tau) b_{gh}(\tau) d\tau \right) \quad (27)$$

It is an almost universal principle in continuous-time stochastic processes theory that conditional expected values of functions of the future, given the past, are solutions to certain differential equations. More often than not, these are needed in order to construct the solution. Therefore, the theory of differential equations and numerical methods for solving them are part and parcel of stochastic processes and their applications.

The state-wise reserves V_j satisfy the first-order ordinary differential equations (ODE)

$$\begin{aligned} \frac{d}{dt} V_j(t) &= r(t) V_j(t) - b_j(t) \\ &\quad - \sum_{k; k \neq j} \mu_{jk}(t) (b_{jk}(t) + V_k(t) - V_j(t)) \end{aligned} \quad (28)$$

valid at all times t where the coefficients r , μ_{jk} , b_j , and b_{jk} are continuous, and there are no lump sum annuity payments. The ultimo conditions

$$V_j(n-) = \Delta B_j(n) \quad (29)$$

$j = 1, \dots, J$, follow from the very definition of the reserve. Likewise, at times t where annuity lump sums are due,

$$V_j(t-) = \Delta B_j(t) + V_j(t) \quad (30)$$

The equation (28) are so-called backward differential equations since the solution is to be computed backwards starting from equation (29).

The differential equations can be derived in various ways. We will sketch a simple heuristic method called *direct backward construction*, which works due to the piecewise deterministic behavior of the Markov chain. Split the expression on the right of equation (5) into

$$\mathcal{V}(t) = dB(t) + e^{-r(t)dt} \mathcal{V}(t+dt) \quad (31)$$

(suppressing a negligible term $o(dt)$) and condition on what happens in the time interval $(t, t+dt]$. With probability $1 - \mu_{j\cdot}(t) dt$ the policy stays in state j and, conditional on this, $dB(t) = b_j(t) dt$ and the expected value of $\mathcal{V}(t+dt)$ is $V_j(t+dt)$. With probability $\mu_{jk}(t) dt$ the policy moves to state j and, conditional on this, $dB(t) = b_{jk}(t)$ and the expected value of value of $\mathcal{V}(t+dt)$ is $V_k(t+dt)$. One gathers

$$\begin{aligned} V_j(t) &= (1 - \mu_{j\cdot}(t) dt) (b_j(t)dt + e^{-r(t)dt} V_j(t+dt)) \\ &\quad + \sum_{k; k \neq j} \mu_{jk}(t) dt (b_{jk}(t) + e^{-r(t)dt} V_k(t+dt)) \\ &\quad + o(dt) \end{aligned} \quad (32)$$

Rearranging, dividing by dt , and letting $dt \rightarrow 0$, one arrives at equation (28). In the single-life model sketched in Figure 1, consider an endowment insurance with sum b against premium at level rate c under constant interest rate r . The differential equation for V_0 is

$$\frac{d}{dt} V_0(t) = r V_0(t) + c - \mu(t) (b - V_0(t)) \quad (33)$$

subject to $V_0(n-) = b$. This is *Thiele's differential equation*, discovered in 1875.

The expression on the right of equation (28) shows how the reserve, seen as a debt, increases with interest (first term) and decreases with redemption of annuity type in the current state (second term), and of lump sum type upon transition to other states (third term). The quantity

$$R_{jk} = b_{jk}(t) + V_k(t) - V_j(t) \quad (34)$$

appearing in the third term is called the *sum at risk* in respect of transition from state j to state k at time t since it is the amount credited to the insured's account upon such a transition: the lump sum payable immediately plus the adjustment of the reserve. This sum multiplied with the rate $\mu_{jk}(t)$ is a rate of expected payments. Solving equation (28) with respect to $-b_j(t)$, which can be seen as a premium (rate), shows that the premium consists of a *savings premium* $\frac{d}{dt} V_j(t) - r(t) V_j(t)$ needed to maintain the reserve (the increase of the reserve less the interest it earns), and a *risk premium* $\sum_{k; k \neq j} \mu_{jk}(t) R_{jk}(t)$ needed to cover risk due to transitions.

The differential equations (28) are as transparent as the defining integral expression (27) themselves, but there are other and more important reasons why they are useful.

Firstly, the easiest (and sometimes the only) way of computing the values of the reserves is by solving the differential equations numerically (e.g., by some finite difference method). The coefficients in the equations are precisely the elemental functions that are specified in the model and in the contract. Thus all values of the state-wise reserves are obtained in one run. The integrals (equation 27) might be computed numerically, but that would require separate computation of the transition probabilities as functions of τ for each given t . In general, the transition probabilities are themselves compound quantities that can only be obtained as solutions to differential equations.

Secondly, the differential equations are indispensable constructive tools when more complicated products are considered. For instance, if the life endowment contract behind equation (33) is modified such that 50% of the reserve is paid out upon death in addition to the sum insured, then its differential equation becomes

$$\frac{d}{dt} V_0(t) = r V_0(t) + c - \mu(t) (b - 0.5 V_0(t)) \quad (35)$$

which is just as easy as equation (33). Another case in point is administrative expenses, which are treated as benefits and covered by charging the policyholder an extra premium in accordance with the equivalence principle. Such expenses may incur upon the inception of the policy (included in $\Delta B_0(0)$), as annuity type payments (included in the $b_g(t)$), and in connection with payments of death and endowment benefits (included in the $b_{gh}(t)$ and the $\Delta B_g(t)$). In particular, expenses related to the company's investment operations are typically allocated to the individual policies on a pro rata basis, in proportion to their individual reserves. Thus, for our generic policy, there is a cost element running at rate $\gamma(t) V_j(t)$ at time t in state j . Subtracting this term on the right-hand side of equation (28) creates no difficulty and, virtually, is just to the effect of reducing the interest rate $r(t)$.

The noncentral conditional moments

$$V_j^{(q)}(t) = \mathbb{E}[\mathcal{V}(t)^q | Z(t) = j] \quad (36)$$

$q = 1, 2, \dots$, do not in general possess explicit integral expressions. They are, however, solutions to

the backward differential equations

$$\begin{aligned} \frac{d}{dt} V_j^{(q)}(t) &= (qr(t) + \mu_j(t)) V_j^{(q)}(t) \\ &\quad - q b_j(t) V_j^{(q-1)}(t) - \sum_{k; k \neq j} \mu_{jk}(t) \\ &\quad \times \sum_{p=0}^q \binom{q}{p} (b_{jk}(t))^p V_k^{(q-p)}(t) \end{aligned} \quad (37)$$

subject to the conditions $V_j^{(q)}(n-) = \Delta B_j(n)^q$ (plus joining conditions at times with annuity lump sums). The backward argument goes as for the reserves, only with a few more details to attend to, starting from

$$\begin{aligned} \mathcal{V}(t)^q &= (dB(t) + e^{-r(t)dt} \mathcal{V}(t+dt))^{q-p} \\ &= \sum_{p=0}^q \binom{q}{p} dB(t)^p e^{-r(t)dt(q-p)} \mathcal{V}(t+dt)^{q-p} \end{aligned} \quad (38)$$

Higher-order moments shed light on the risk associated with the portfolio. Recalling equation (9) and using the notation in the section titled “Valuation of Life Insurance Contracts by the Principle of Equivalence”, a solvency margin approximately equal to the upper ε -fractile in the distribution of the discounted outstanding net liability is given by $\sum_{i=1}^m \xi_i + c_{1-\varepsilon} s_m$, where $c_{1-\varepsilon}$ is the upper ε -fractile of the standard normal distribution. More refined estimates of the fractiles of the total liability can be obtained by involving three or more moments.

Path-Dependent Payments and Semi-Markov Models

The technical matters in the sections “The Markov Chain Model for the Policy History” and “Actuarial Analysis of Standard Insurance Products” become more involved if the contractual payments or the transition intensities depend on the past life history. We will consider examples where they may depend on the sojourn time $S(t)$ that has elapsed since entry into the current state, henceforth called the *state duration*. Thus, if $Z(t) = j$ and $S(t) = s$ at policy

duration t , the transition intensities and payments are of the form $\mu_{jk}(s, t)$, $b_j(s, t)$, and $b_{jk}(s, t)$. To ease exposition, we disregard intermediate lump sum annuities, but allow for terminal endowments $\Delta B_j(s, n)$ at time n . The state-wise reserve will now be a function of the form $V_j(s, t)$.

In simple situations (e.g., no possibility of return to previously visited states), one may still work out integral expressions for probabilities and reserves (they will be multiple integrals). The differential equation approach always works, however. The relationship (32) modifies to

$$\begin{aligned} V_j(s, t) = & (1 - \mu_{jj}(s, t) dt) \\ & \times (b_j(s, t) dt + e^{-r(t)dt} V_j(s + dt, t + dt)) \\ & + \sum_{k: k \neq j} \mu_{jk}(s, t) dt (b_{jk}(s, t) + V_k(0, t)) \\ & + o(dt) \end{aligned} \quad (39)$$

from which one obtains the first-order partial differential equations

$$\begin{aligned} \frac{\partial}{\partial t} V_j(s, t) = & r(t) V_j(s, t) - \frac{\partial}{\partial s} V_j(s, t) - b_j(s, t) \\ & - \sum_{k: k \neq j} \mu_{jk}(s, t) \\ & \times (b_{jk}(s, t) + V_k(0, t) - V_j(s, t)) \end{aligned} \quad (40)$$

subject to the conditions

$$V_j(s, n-) = \Delta B_j(s, n) \quad (41)$$

We give two examples of payments dependent on state duration. In the framework of the disability model in Figure 2, an n -year disability annuity payable at rate 1 only after a qualifying period of q , is given by $b_1(s, t) = 1[q < s < t < n]$. In the framework of the three-lives model sketched in Figure 3, an n -year term insurance of 1 payable upon the death of y if z is still alive and x is dead and has been so for at least q years, is given by $b_{14}(s, t) = 1[q < s < t < n]$.

Probability models with intensities dependent on state duration are known as a *semi-Markov models*. A case in support of their relevance is the disability model in Figure 2. If there are various forms of disability, then the state duration may carry information

about the severity of the disability and hence about prospects of longevity and recovery.

Managing Nondiversifiable Risk for Standard Insurance Products

Life insurance policies are typically long-term contracts, with time horizons wide enough to see substantial variations in interest, mortality, and other economic and demographic conditions affecting the economic result of the portfolio. The rigid conditions of the standard contract leave no room for the insurer to meet adverse developments of such conditions; he cannot cancel contracts that are in force, nor can he reduce their benefits or raise their premiums. Therefore, with standard insurance products there is associated a risk that cannot be diversified by increasing the size of the portfolio as in the section “Valuation of Life Insurance Contracts by the Principle of Equivalence”. The limit operation leading to equation (12) was made under the assumption of fixed interest. In an extended setup, with random economic and demographic factors, this amounts to conditioning on $\mathcal{G}_n$, the economic and demographic development over the term of the contract. Instead of equation (12), one gets

$$V(t) = \mathbb{E} \left[\int_t^n e^{-\int_t^\tau r(s) ds} dB(\tau) \middle| \mathcal{H}_t, \mathcal{G}_n \right] \quad (42)$$

At time t only $\mathcal{H}_t$ and $\mathcal{G}_t$ are known, so equation (42) is not a feasible reserve. In particular, the equivalence principle equation (7), recast as

$$\Delta B_0(0) + \mathbb{E} \left[\int_0^n e^{-\int_0^\tau r(s) ds} dB(\tau) \middle| \mathcal{G}_n \right] = 0 \quad (43)$$

is also infeasible since benefits and premiums are fixed at time 0 when $\mathcal{G}_n$ cannot be anticipated. The traditional way of managing the nondiversifiable risk is to charge premiums sufficiently high to cover, on the average in the portfolio, the contractual benefits under all likely economic-demographic scenarios. The systematic surpluses that (most likely) will be generated by such prudently calculated premiums belong to the insured and are paid back in arrears as the history $\mathcal{G}_t$ unfolds. Such contracts are called *participating policies* or *with-profit contracts*.

The repayments, called *dividends* or *bonus*, are represented by a payment function D . They should be controlled in such a manner as to ultimately restore equivalence when the full history is known at time n :

$$\Delta B_0(0) + \mathbb{E} \left[\int_0^n e^{-\int_0^\tau r(s) ds} (dB(\tau) + dD(\tau)) \middle| \mathcal{G}_n \right] = 0 \quad (44)$$

A common way of designing the prudent premium plan is, in the framework of the preceding five sections, to calculate premiums and reserves on a so-called technical basis with interest rate r^* and transition intensities μ_{jk}^* that represent a worst-case-scenario. Equipping all technical quantities with an asterisk, we denote the corresponding reserves by V_j^* , the sums at risk by R_{jk}^* etc. The surplus generated by time t is, quite naturally, defined as the excess of the factual retrospective reserve over the contractual prospective reserve,

$$S(t) = \mathcal{U}(t) - \sum_{j=0}^J I_j V_j^*(t) \quad (45)$$

Upon differentiating this expression, using equations (4), (13), and (28), and the obvious relationship $dI_j(t) = \sum_{k; k \neq j} (dN_{kj}(t) - dN_{jk}(t))$, one obtains after some rearrangement that

$$dS(t) = S(t) r(t) dt + dC(t) + dM(t) \quad (46)$$

where

$$dC(t) = \sum_{j=0}^J I_j(t) c_j(t) dt \quad (47)$$

$$c_j(t) = (r(t) - r^*) V_j^*(t) + \sum_{k; k \neq j} R_{jk}^*(t) (\mu_{jk}^*(t) - \mu_{jk}(t)) \quad (48)$$

$$dM(t) = - \sum_{j \neq k} R_{jk}^*(t) (dN_{jk}(t) - I_j(t) \mu_{jk}(t) dt) \quad (49)$$

The right-hand side of equation (46) displays the dynamics of the surplus. The first term is the interest earned on the current surplus. The last term, given

by equation (49), is purely erratic and represents the policy's instantaneous random deviation from the expected development. The second term, given by equation (47), is the systematic contribution to surplus, and equation (48) shows how it decomposes into gains due to prudent assumptions about interest and transition intensities in each state. One may show that equation (44) is equivalent to

$$\mathbb{E} \left[\int_0^n e^{-\int_0^\tau r(s) ds} (dC(\tau) - dD(\tau)) \middle| \mathcal{G}_n \right] = 0 \quad (50)$$

which says that, on the average over the portfolio, all surpluses are to be repaid as dividends.

The dividend payments D are controlled by the insurer. Since negative dividends are not allowed, it is possible that equation (50) cannot be ultimately attained if the contributions to surplus turn negative (the technical basis was not sufficiently prudent) and/or dividends were paid out prematurely. Designing the dividend plan D is therefore a major issue, and it can be seen as a problem in optimal stochastic control theory. For a general account of the point process version of this theory, see [4]. Various schemes used in practice are described in [5]. The simplest (and least cautious one) is the so-called contribution plan, whereby surpluses are repaid currently as they arise: $D = C$.

The much discussed issue of *guaranteed interest* takes a clear form in the framework of the present theory. Focusing on interest, suppose the technical intensities μ_{jk}^* are the same as the factual μ_{jk} so that the surplus emerges only from the excess of the factual interest rate $r(t)$ over the technical rate r^* :

$$dC(t) = (r(t) - r^*) V_{Z(t)}^*(t) dt \quad (51)$$

Under the contribution plan, $dD(t)$ must be set to 0 if $dC(t) < 0$, and the insurer will therefore have to cover the negative contributions $dC^-(t) = (r^* - r(t))_+ V_{Z(t)}^*(t) dt$. Averaging out the life-history randomness, the discounted value of these claims is

$$\int_0^n e^{-\int_0^\tau r(s) ds} (r^* - r(\tau))_+ \sum_{j=0}^J p_{0j}(0, \tau) V_j^*(\tau) d\tau \quad (52)$$

Mobilizing the principles of arbitrage pricing theory set out in the section titled "Valuation of

Financial Contracts,” we conclude that the interest guarantee inherent in the present scheme has a market price which is the expected value of equation (52) under the equivalent martingale measure. Charging a down premium equal to this price at time 0 would eliminate the downside risk of the contribution plan without violating the format of the with-profit scheme.

Unit-Linked Insurance

The dividends D redistributed to holders of standard with-profit contracts can be seen as a way of adapting the benefits payments to the development of the nondiversifiable risk factors of $\mathcal{G}_t$, $0 < t \leq n$. Alternatively, one could specify in the very terms of the contract, that the payments will depend, not only on life-history events, but also on the development of interest, mortality, and other economic-demographic conditions. One such approach is the unit-linked contract, which relates the benefits to the performance of the insurer’s investment portfolio. To keep things simple, let the interest rate $r(t)$ be the only uncertain nondiversifiable factor. A straightforward way of eliminating the interest rate risk is to let the payment function under the contract be of the form

$$dB(t) = e^{\int_0^t r(s) ds} dB^\circ(t) \quad (53)$$

where B° is a baseline payment function dependent only on the life history. This means that all payments, premiums and benefits, are index-regulated with the value of a unit of the investment portfolio. Inserting this into equation (43), assuming that life-history events and market events are independent, the equivalence requirement becomes

$$\Delta B_0^\circ(0) + \mathbb{E} \left[\int_0^n dB^\circ(\tau) \right] = 0 \quad (54)$$

This requirement does not involve the future interest rates and can be met by setting an equivalence baseline premium level at time 0.

Perfect unit-linked products of the form (53) are not offered in practice. Typically, only the sum insured (e.g., a term insurance or a life endowment) is index-regulated, while the premiums are not. Moreover, the contract usually comes with a guarantee that the sum insured will not be less

than a certain nominal amount. Averaging out over the life histories, the payments become purely financial derivatives, and pricing goes by the principles described in the section “Valuation of Financial Contracts”. If random life-history events are kept part of the model, one faces a pricing problem in an incomplete market. This problem was formulated and solved in [6] in the framework of the theory of risk minimization [7].

Defined Benefits and Defined Contributions

With-profit and unit-linked insurance schemes are just two ways of adapting benefits to the long-term development of nondiversifiable risk factors. The former does not include the adaptation rule in the terms of the contract, whereas the latter does. We mention two other arch-type insurance schemes that are widely used in practice. *Defined benefits* means literally that the benefits are specified in the contract, either in nominal figures as in the with-profit contract or in units of some index. A commonly used index is the salary (final or average) of the insured. In that case also, the contributions (premiums) are usually linked to the salary (typically a certain percentage of the annual income). Risk management of such a scheme is a matter of designing the rule for collection of contributions. Unless the future benefits can be precisely predicted or reproduced by dynamic investment portfolios, defined benefits leave the insurer with a major nondiversifiable risk. Defined benefits are gradually being replaced with their opposite, *defined contributions*, with only premiums specified in the contract. This scheme has much in common with the traditional with-profit scheme, but leaves more flexibility to the insurer as benefits do not come with a minimum guarantee.

Securitization

Generally speaking, and recalling the section “Valuation of Financial Contracts,” any introduction of new securities in a market helps to complete it (see **Securitization/Life**). *Securitization* means creating tradeable securities that may serve to make nontraded claims attainable. This device, well known and widely used in the commodities

markets, was introduced in nonlife insurance in the 1990s when exchanges and insurance corporations launched various forms of *insurance derivatives* aimed to transfer catastrophe risk to the financial markets. Securitization of nondiversifiable risk in life insurance, e.g., through bonds with coupons related to mortality experience, is conceivable. If successful, it would open new opportunities of financial management of life insurance risk by the principles described in the section “Valuation of Financial Contracts”. A work in this spirit is [8], where market attitudes are modeled for all forms of risk associated with a life insurance portfolio, leading to market values for reserves.

A View to the Literature

In this brief survey of contemporary life insurance mathematics, space has not been devoted to the wealth of techniques that are now mainly only of historical interest, and no attempt has been made to trace the origins of modern ideas and results. References to the literature are selected accordingly, their purpose being to add details to the broad picture drawn here. A key reference on the early history of life insurance mathematics is [11]. Textbooks covering classical insurance mathematics are [12–16]. An account of counting processes and martingale techniques in life insurance mathematics can be compiled from [5, 17].

References

- [1] Norberg, R. (2004). Life insurance mathematics, in *Encyclopaedia of Actuarial Science*, B. Sundt & J. Teugels, eds, John Wiley & Sons.
- [2] Björk, T. (1998). *Arbitrage Theory in Continuous Time*, Oxford University Press.
- [3] Hoem, J.M. (1969). Markov chain models in life insurance, *Blätter der Deutschen Gesellschaft für Versicherungsmathematik* **9**, 91–107.

- [4] Davis, M.H.A. (1993). *Markov Models and Optimization*, Chapman & Hall.
- [5] Norberg, R. (1999). A theory of bonus in life insurance, *Finance and Stochastics* **3**, 373–390.
- [6] Møller, T. (1998). Risk minimizing hedging strategies for unit-linked life insurance, *ASTIN Bulletin* **28**, 17–47.
- [7] Föllmer, H. & Sondermann, D. (1986). Hedging of non-redundant claims, in *Contributions to Mathematical Economics in Honor of Gerard Debreu*, W. Hildebrand & A. Mas-Collel, eds, North-Holland, pp. 205–223.
- [8] Steffensen, M. (2000). A no arbitrage approach to Thiele’s differential equation, *Insurance: Mathematics & Economics* **27**, 201–214.
- [9] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag.
- [10] Norberg, R. (1991). Reserves in life and pension insurance, *Scandinavian Actuarial Journal* **1991**, 1–22.
- [11] Hald, A. (1987). On the early history of life insurance mathematics, *Scandinavian Actuarial Journal* **1987**, 4–18.
- [12] Berger, A. (1939). *Mathematik der Lebensversicherung*, Verlag von Julius Springer, Vienna.
- [13] Bowers Jr, N.L., Gerber, H.U., Hickman, J.C. & Nesbitt, C.J. (1986). *Actuarial Mathematics*, The Society of Actuaries, Itasca.
- [14] De Vylder, F. & Jaumain, C. (1976). *Exposé Moderne de la Théorie Mathématique des Opérations Viagère*, Office des Assureurs de Belgique, Bruxelles.
- [15] Gerber, H.U. (1995). *Life Insurance Mathematics*, 2nd Edition, Springer-Verlag.
- [16] Jordan, C.W. (1967). *Life Contingencies*, The Society of Actuaries, Chicago.
- [17] Hoem, J.M. & Aalen, O.O. (1978). Actuarial values of payment streams, *Scandinavian Actuarial Journal* **1978**, 38–47.

RAGNAR NORBERG

Multistate Systems

All technical systems are designed to perform their intended tasks in a given environment. Some systems can perform their tasks with various distinguished levels of intensity usually referred to as *performance rates*. A system that can have a finite number of performance rates is called a *multistate system* (MSS) (see **Reliability Optimization**). Usually MSS is composed of elements that in turn can be multistate ones. An element is an entity in a system, which is not further subdivided. This does not imply that an element cannot be made of parts, but means only that, in a given reliability study, it is regarded as a self-contained unit and is not analyzed in terms of the reliability performances of its constituents.

Actually, a binary system is the simplest case of an MSS having two distinguished states (perfect functioning and complete failure).

There are different situations where a system is considered to be an MSS:

1. Any system consisting of different units that have a cumulative effect on the performance of the entire system has to be considered as an MSS. Indeed, the performance rate of such a system depends on the availability of its units, as the different numbers of the available units can provide different levels of the task performance.

The simplest example of such a situation is the well-known k -out-of- n systems. These systems consist of n identical binary units and can have $n + 1$ states depending on the number of available units. The system performance rate is assumed to be proportional to the number of available units. It is assumed that the performance rates corresponding to more than $k - 1$ available units are acceptable. When the contributions of different units to the performance rate of the cumulative system are different, the number of possible MSS states grows dramatically as different combinations of k available units can provide different performance rates for the entire system.

2. The performance rate of elements composing a system can also vary as a result of their deterioration (fatigue, partial failures) or because of variable ambient conditions. Element failures

can lead to the degradation of the entire MSS performance.

The performance rates of the elements can range from perfect functioning up to complete failure. The failures that lead to the decrease in the element performance are called *partial failures*. After partial failure, elements continue to operate at reduced performance rates, and after complete failure, the elements are totally unable to perform their tasks.

Consider the following two examples of MSSs:

- In the power supply system consisting of generating and transmitting facilities, each generating unit can function at different levels of capacity. Generating units are complex assemblies of many parts. The failures of different parts may lead to situations where the generating unit continues to operate, but at a reduced capacity. This can occur during the outages of several auxiliaries such as pulverizers, water pumps, fans, etc.
- In a wireless communication system consisting of transmission stations, the state of each station is defined by the number of subsequent stations covered in its range. This number depends not only on the availability of station amplifiers, but also on the conditions for signal propagation that depend on weather, solar activity, etc.

Generic Model of Multistate System

To analyze MSS behavior, one has to know the characteristics of its elements. Any system element j can have $k_j + 1$ different states corresponding to the performance rates, represented by the set $\mathbf{g}_j = \{g_{j0}, g_{j1}, \dots, g_{jk_j}\}$, where g_{jh} is the performance rate of element j in the state h , $h \in \{0, 1, \dots, k_j\}$. The performance rate G_j of element j at any time instant is a random variable that takes its values from $\mathbf{g}_j : G_j \in \mathbf{g}_j$. The probabilities associated with the different states (performance rates) of the system element j can be represented by the set

$$p_j = \{p_{j0}, p_{j1}, \dots, p_{jk_j}\} \quad (1)$$

where

$$p_{jh} = \Pr\{G_j = g_{jh}\} \quad (2)$$

Since the element's states compose the complete group of mutually exclusive events (meaning that the element can always be in one and only in one of $k_j + 1$ states)

$$\sum_{h=0}^{k_j} p_{jh} = 1 \quad (3)$$

Expression (2) defines the probability mass function (pmf) of a discrete random variable G_j . The collection of pairs $g_{jh}, p_{jh}, h = 0, 1, \dots, k_j$, completely determines the performance distribution of element j .

When the MSS consists of n independent elements, its performance rate is unambiguously determined by the performance rates of these elements. At each moment, the system elements have certain performance rates corresponding to their states. The state of the entire system is determined by the states of its elements. Assume that the entire system has $K + 1$ different states and that v_i is the entire system's performance rate in state $i \in \{0, \dots, K\}$. The MSS performance rate is a random variable V that takes values from the set $M = \{v_1, \dots, v_K\}$.

Let

$$L^n = \{g_{10}, \dots, g_{1k_1}\} \times \{g_{20}, \dots, g_{2k_2}\} \times \dots \times \{g_{n0}, \dots, g_{nk_n}\} \quad (4)$$

be the space of possible combinations of performance rates for all of the system elements and $M = \{v_0, \dots, v_K\}$ be the space of possible values of the performance rate for the entire system. The transform $\phi(G_1, \dots, G_n) : L^n \rightarrow M$, which maps the space of the elements' performance rates into the space of system's performance rates, is named the *system structure function* (see **Imprecise Reliability**).

The generic model of the MSS includes the pmf of performances for all the system elements and system structure function [1]:

$$g_j, p_j, \quad 1 \leq j \leq n \quad (5)$$

$$V = \phi(G_1, \dots, G_n) \quad (6)$$

From this model, one can obtain the pmf of the entire system's performance in the form

$$q_i, v_i, 0 \leq i \leq K, \text{ where } q_i = \Pr\{V = v_i\} \quad (7)$$

Measures of Risk

The acceptability of system state can be usually defined by the acceptability function $f(V, \theta)$ representing the desired relation between the system performance V and some limit value θ named *system demand* ($f(V, \theta) = 1$ if the system performance is acceptable and $f(V, \theta) = 0$ otherwise). The MSS reliability is defined as its expected acceptability (probability that the MSS satisfies the demand) [2]. Having the system pmf (equation (7)) one can obtain its reliability as expected value of the acceptability function $E[f(V, \theta)]$:

$$R(\theta) = E[f(V, \theta)] = \sum_{i=1}^K q_i f(v_i, \theta) \quad (8)$$

For example, in applications where the system performance is defined as a task execution time and θ is the maximum allowed task execution time, equation (8) takes the form

$$R(\theta) = \sum_{i=1}^K q_i 1(v_i < \theta) \quad (9)$$

and in applications where the system performance is defined as system productivity (capacity) and θ is the minimum allowed productivity, equation (8) takes the form

$$R(\theta) = \sum_{i=1}^K q_i 1(v_i > \theta) \quad (10)$$

where $1(\cdot)$ is unity function: $1(\text{TRUE}) = 1$, $1(\text{FALSE}) = 0$.

The risk measure named unreliability is defined as $1 - R(\theta)$.

Another important measure of the risk associated with system performance reduction is the expected unsupplied demand $W(\theta)$. It can be obtained as

$$W(\theta) = \sum_{i=1}^K q_i \cdot |v_i - \theta| \cdot (1 - f(v_i, \theta)) \quad (11)$$

To calculate the indices $R(\theta)$ and $W(\theta)$, one has to obtain the pmf of the MSS random performance in the equation (7) from the model (5) and (6). The reliability block diagram (RBD) (see **Systems Reliability; Markov Modeling in Reliability**) method for obtaining the performance distribution

for series–parallel MSS is based on the universal generating function (u -function) technique, which was introduced in [3] and proved to be very effective for the reliability evaluation of different types of MSSs [2].

Evaluating the Risk Associated with MSS Performance Reduction

Universal Generating Function (u -Function) Technique

The u -function representing the pmf of a discrete random variable Y_j is defined as a polynomial

$$u_j(z) = \sum_{h=0}^{k_j} \alpha_{jh} z^{y_{jh}} \quad (12)$$

where the variable Y_j has $k_j + 1$ possible values and $\alpha_{jh} = \Pr\{Y_j = y_{jh}\}$.

To obtain the u -function representing the pmf of a function of n independent random variables $\varphi(Y_1, \dots, Y_n)$, the following composition operator is used:

$$\begin{aligned} U(z) &= \otimes_{\varphi}(u_1(z), \dots, u_n(z)) \\ &= \otimes_{\varphi} \left(\sum_{h_1=0}^{k_1} \alpha_{1h_1} z^{y_{1h_1}}, \dots, \sum_{h_n=0}^{k_n} \alpha_{nh_n} z^{y_{nh_n}} \right) \\ &= \sum_{h_1=0}^{k_1} \sum_{h_2=0}^{k_2} \dots \sum_{h_n=0}^{k_n} \left(\prod_{i=1}^n \alpha_{ih_i} z^{\varphi(y_{1h_1}, \dots, y_{nh_n})} \right) \end{aligned} \quad (13)$$

The polynomial $U(z)$ represents all the possible mutually exclusive combinations of realizations of the variables by relating the probabilities of each combination to the value of function $\varphi(Y_1, \dots, Y_n)$ for this combination.

In the case of MSS, u -functions

$$u_j(z) = \sum_{h_j=0}^{k_j} p_{jh_j} z^{g_{jh_j}} \quad (14)$$

represents the pmf of random performances of independent system elements. Having a generic model of an MSS in the form of equations (5) and (6) one

can obtain the measures of the risk by applying the following steps:

1. Represent the pmf of the random performance of each system element j in the form of the u -function (equation (14)).
2. Obtain the u -function of the entire system $U(z)$ that represents the pmf of the random variable V by applying the composition operator $\otimes_{\phi}$ that uses the system structure function ϕ (equation (6)).
3. Calculate the MSS performance indices applying equations (10) and (11) over system performance pmf (equation (7)) represented by the u -function $U(z)$.

While steps 1 and 3 are rather trivial, step 2 may involve complicated computations. Indeed, the derivation of a system structure function for various types of systems is usually a difficult task. As shown in [2], representing the structure functions in the recursive form is beneficial from the viewpoint of both derivation clarity and computation simplicity. In many cases, the structure function of the entire MSS can be represented as the composition of the structure functions corresponding to some subsets of the system elements (MSS subsystems).

The u -functions of the subsystems can be obtained separately and the subsystems can be further treated as single equivalent elements with the performance pmf represented by these u -functions. The method for distinguishing recurrent subsystems and replacing them with single equivalent elements is based on a graphical representation of the system structure and is referred to as the *reliability block diagram* method. This approach is usually applied to systems with a complex series–parallel configuration.

Generalized RBD Method for Series–Parallel MSS

The structure function of a complex series–parallel system can always be represented as a composition of the structure functions of statistically independent subsystems containing only elements connected in a series or in parallel. Therefore, to obtain the u -function of a series–parallel system, one has to apply the composition operators recursively to obtain the u -functions of the intermediate pure series or pure parallel structures.

The following algorithm realizes this approach:

1. Find any pair of system elements (i and j) connected in parallel or in series in the MSS.
2. Obtain the u -function of this pair using the corresponding composition operator $\otimes_\varphi$ over two u -functions of the elements:

$$\begin{aligned} U_{\{i,j\}}(z) &= u_i(z) \otimes_\varphi u_j(z) \\ &= \sum_{h_i=0}^{k_i} \sum_{h_j=0}^{k_j} p_{ih_i} p_{jh_j} z^{\varphi(g_{ih_i}, g_{jh_j})} \end{aligned} \quad (15)$$

where the function φ is determined by the nature of interaction between the elements' performances.

3. Replace the pair with single element having the u -function obtained in step 2.
4. If the MSS contains more then one element, return to step 1.

The resulting u -function represents the performance distribution of the entire system.

The choice of the functions φ for series and parallel subsystems depends on the type of system.

For example, in flow transmission MSS, where performance is defined as capacity or productivity, the total capacity of a subsystem containing two independent elements connected in series is equal to the capacity of a bottleneck element (the element with least performance). Therefore, the structure function for such a subsystem takes the form

$$\phi_{\text{ser}}(G_1, G_2) = \min\{G_1, G_2\} \quad (16)$$

If the flow can be dispersed and transferred by parallel channels simultaneously (which provides the work sharing), the total capacity of a subsystem containing two independent elements connected in parallel is equal to the sum of the capacities of these elements. Therefore, the structure function for such a subsystem takes the form

$$\phi_{\text{par}}(G_1, G_2) = G_1 + G_2 \quad (17)$$

In the task processing MSS, where the performance of each element is characterized by its processing speed, different subtasks are performed by different components consecutively. Therefore, the time for the entire task completion (reciprocal of the processing speed) is equal to the sum of subtask execution times. In terms of the processing speeds, one

can determine the performance of a subsystem consisting of two consecutive elements as

$$\begin{aligned} \varphi_{\text{ser}}(G_1, G_2) &= \text{inv}(G_1, G_2) \\ &= \frac{1}{1/G_1 + 1/G_2} = \frac{G_1 G_2}{G_1 + G_2} \end{aligned} \quad (18)$$

Parallel elements perform the same task starting it simultaneously. The task is completed by a group of elements when it is completed by any element belonging to this group. Therefore, the performance of the group is equal to the performance of its fastest available element. Therefore, for a subsystem of two parallel elements

$$\phi_{\text{par}}(G_1, G_2) = \max\{G_1, G_2\} \quad (19)$$

MSS Survivability and Optimal Defense

A survivable system is one that is able to complete its mission in a timely manner, even if significant portions are incapacitated by attack or accident. This definition presumes two important things.

First, the impact of both external factors (attacks) and internal causes (failures), affects system survivability. Therefore, it is important to take into account the influence of the availability of system elements on the entire system survivability.

Second, a system can have different states corresponding to different combinations of failed or destroyed elements composing the system. The success of a system is defined as its ability to meet a demand.

The damage caused by the destruction of elements with different performance rates will be different. Therefore, the performance rates of system elements should be taken into account when the risk associated with both the failures and the attacks is estimated. The index $R(\theta)$ (equation (8)) for the systems exposed to the external attacks is named *survivability*. The risk measure $1 - R(\theta)$ is named *system vulnerability*.

Any external impact can be considered as a common cause failure since several elements, not separated one from another, can be destroyed by the same impact. To evaluate the system vulnerability, one has to incorporate the common cause failures probability into the RBD method as suggested in [2, 4].

To reduce the system vulnerability, the elements can be separated (to avoid the destruction of several elements by a single attack) and protected.

Protecting against intentional attacks is fundamentally different from protecting against accidents or natural cataclysms [5]. Adaptive strategy allows the attacker to target the most sensitive parts of a system. Choosing the time, place, and means of attacks, the attacker always has an advantage over the defender.

Therefore, the optimal policy for allocating resources among possible defensive investments should take into account the attacker's strategy.

Unlike protecting against accidents or natural cataclysms, defense strategies against intentional attacks can influence the adaptive strategy of the attacker. This can be achieved not only by separation and protection of system elements, but also by creating false targets.

The three components of the defense strategy play different roles: deployment of the false targets reduces the probability of the attack on each separated group of elements, protection reduces the probability of elements getting destroyed in case of attack, and separation reduces the damage caused by a single successful attack. To evaluate the influence of the defense strategy components on the system vulnerability, one has to use the multistate model.

Examples of defense strategy optimization based on the universal generating function and the generalized RBD method can be found in [6–9].

References

- [1] Lisnianski, A. & Levitin, G. (2003). *Multi-State System Reliability. Assessment, Optimization and Applications*, World Scientific.
- [2] Levitin, G. (2005). *Universal Generating Function in Reliability Analysis and Optimization*, Springer-Verlag, London.
- [3] Ushakov, I. (1987). Optimal standby problems and a universal generating function, *Soviet Journal of Computer Systems Science* **25**, 79–82.
- [4] Levitin, G. & Lisnianski, A. (2003). Optimizing survivability of vulnerable series-parallel multi-state systems, *Reliability Engineering and System Safety* **79**, 319–331.
- [5] Bier, V., Nagaraj, A. & Abhichandani, V. (2005). Protection of simple series and parallel systems with components of different values, *Reliability Engineering and System Safety* **87**, 315–323.
- [6] Levitin, G. (2003). Optimal multilevel protection in series-parallel systems, *Reliability Engineering and System Safety* **81**, 93–102.
- [7] Levitin, G., Dai, Y., Xie, M. & Poh, K.L. (2003). Optimizing survivability of multi-state systems with multi-level protection by multi-processor genetic algorithm, *Reliability Engineering and System Safety* **82**, 93–104.
- [8] Korczak, E., Levitin, G. & Ben Haim, H. (2005). Survivability of series-parallel systems with multilevel protection, *Reliability Engineering and System Safety* **90**(1), 45–54.
- [9] Levitin, G. (2007). Optimal defense strategy against intentional attacks, *IEEE Transactions on Reliability* **56**(1), 148–157.

GREGORY LEVITIN

Multivariate Reliability Models and Methods

There are several multivariate extensions of failure time distributions, hazard rate, positive and negative aging. In this paper, we generally consider those models that have no unique natural extension in the multivariate case. More specifically, for failure time distributions like exponential and Weibull, there is no unique natural extension. Exponential distribution plays an important role as failure time distribution with loss of memory property (LMP) and also in reliability theory. However, the multivariate extensions of exponential are not straightforward (*see Copulas and Other Measures of Dependency*). There are several versions of bivariate exponential (BVE) or multivariate exponential (MVE) distributions given by several authors. The main characteristics of BVE or MVE are (a) marginal exponential, (b) absolute continuity, and (c) LMP given by Marshall and Olkin [1]. At most, two of the three characteristics are satisfied by these BVE models. If all the three characters are satisfied, then it leads to independent exponentials. In a similar way, there are different generalizations of BVE to bivariate Weibull (BVW), Pareto, and extreme value distributions by transformation of variables.

Bivariate survival data arise when each study subject experiences two events. Failure times of paired human organs, e.g., kidneys, eyes, lungs, and the recurrence of a given disease are particular examples of such data. In a different context, the data may consist of time to diagnosis or hospitalization and the time to eventual death from a fatal disease. In industrial applications, these data types may come from systems whose survival depends on the survival of two very similar components. For example, the breakdown times of dual generators in a power plant or failure times of twin engines in a two-engine airplane are illustrations of bivariate survival data.

Marshall and Olkin [1] proposed a well-known BVE distribution which is widely accepted in the literature for reliability applications. The main characteristic feature of the Marshall–Olkin model (*see Mathematics of Risk and Reliability: A Select History*) is simultaneous failures occurring due to shock, which leads to dependence between the two components. In a similar manner, the characteristic feature

of the Freund [2] model is that the failure rate of the lifetime of a component changes when the other component fails, which is very common in real-life situations. When one component fails, there is a load on the other component and the failure rate of other survival time of the other component increases, which leads to the dependence between the two components. These two types of dependence structures are combined in the BVE of Proschan and Sullo [3].

The BVE models of Marshall and Olkin [1] and Freund [2] are the submodels of the BVE of Proschan and Sullo [3]. The BVE of Block and Basu [4] is the submodel of Freund [2] and also an absolutely continuous part of Marshall and Olkin [1]. The BVE of Marshall and Olkin [1] and Proschan and Sullo [3] are not absolutely continuous and they have a positive density on the diagonal $T_1 = T_2$. Lee [5] extended the BVE of Gumbel [6] to BVW by taking power transformation. Lu [7] extended the Weibull extensions to the Freund and the Marshall–Olkin BVE models. Later Hanagal [8] extended to the multivariate Weibull (MVW) distribution which is the extension of the BVW of Lu [7] and also an extension of the MVE of Marshall and Olkin [1]. Basu [9] extended the univariate failure rate to the bivariate case.

The LMP in the multivariate set up given by Marshall and Olkin [1] is as follows:

$$\begin{aligned} P[T_1 > t_1 + s, \dots, T_k > t_k + s] \\ &= P[T_1 > t_1, \dots, t_k > T_k]P[T_1 > s, \dots, T_k > s] \\ &\quad \times S(t_1 + s, \dots, t_k + s) \\ &= S(t_1, \dots, t_k)S(s, \dots, s) \end{aligned} \quad (1)$$

where $S(\cdot, \dots, \cdot)$ is the multivariate reliability function or multivariate survival function. The above equation is a weaker version of the multivariate LMP with common s . The stronger version is

$$\begin{aligned} P[T_1 > t_1 + s_1, \dots, T_k > t_k + s_k] \\ &= P[T_1 > t_1, \dots, T_k > t_k] \\ &\quad \times P[T_1 > s_1, \dots, T_k > s_k] \\ &\quad \times S(t_1 + s_1, \dots, t_k + s_k) \\ &= S(t_1, \dots, t_k)S(s_1, \dots, s_k) \end{aligned} \quad (2)$$

In the stronger version of LMP, only independent exponentials satisfy this property.

The reliability and risk assessment of the system is an integral element of the design process. To achieve greater reliability of the system, one has to minimize the risk with respect to quality of raw material, skilled manpower, precaution in event of catastrophe, robustness of the system, and several other factors. The risk function increases as the lifetime of the system increases or the reliability of the system decreases. The risk function is directly proportional to the lifetime and inversely proportional to the reliability of the system. The multivariate risk function (or multivariate hazard function) is related to the multivariate reliability function i.e., $H(t_1, \dots, t_k) = -\log S(t_1, \dots, t_k)$ where $H(t_1, \dots, t_k)$ is the multivariate risk function.

In the section titled “Frailty Models”, we discuss the frailty models in multivariate reliability models (see **Risk Classification/Life**). Frailty means unknown covariates. Frailty models are random-effect models and they follow certain distributions. They are used to reduce heterogeneity in the error term.

Frailty Models

The shared gamma frailty models were suggested by Clayton [10] for the analysis of the correlation between clustered survival times in genetic epidemiology (see **Insurance Applications of Life Tables; Default Correlation**). In a frailty model, it is absolutely necessary to be able to include explanatory variables with frailty or random effect. The reason is that the frailty describes the influence of common unknown factors. If some common covariates are included in the model, the variation owing to unknown covariates should be reduced.

Common covariates are common for all members of the group. For monozygotic twins, examples are sex and any other genetically based covariate. Both monozygotic and dizygotic twins share date of birth and common prebirth environment. By measuring some potentially important covariates, we can examine the influence of the covariates, and we can examine whether they explain the dependence, that is, whether the frailty has no effect (or more correctly, no variation), when the covariate is included in the model.

It is not possible, in practice, to include all relevant covariates. For example, we might know that some

given factor is important, but if we do not know the value of the factor for each individual, we cannot include the variable in the analysis. For example, it is known that excretion of small amounts of albumin in the urine is a diagnostic marker for increased mortality, not only for diabetic patients, but also for the general population. However, we are unable to include this variable, unless we actually obtain urine and analyze samples for each individual under study. It is furthermore possible that we are not aware that there exist variables that we ought to include. For example, this could be a genetic factor, as we do not know all possible genes having influence on survival. This consideration is true for all regression models and not only survival models. If it is known that some factor is important, it makes sense to try to obtain the individual values, but if it is not possible, the standard is to ignore the presence of such variables. In general terms, we let the heterogeneity go into the error term. This, of course, leads to an increase in the variability of the response compared to the case, when the variables are included. In the survival data case, however, the increased variability implies a change in the form of the hazard function, as is illustrated by some more detailed calculations.

The regression model is derived conditionally on the shared frailty (Y). Conditionally on Y , the hazard function for individual j is assumed to be of the form

$$Y\mu_j(t)$$

where $\mu_j(t)$ is the hazard function of the j th individual and it is assumed that the value of Y is common to several individuals in a group. Independence of lifetimes corresponds to no variability in the distribution of Y , implies that Y has a degenerate distribution. When the distribution is not degenerate, the dependence is positive. The value of Y can be considered as generated from unknown values of some explanatory variables. Conditional on $Y = y$, the multivariate survival function (see **Lifetime Models and Risk Assessment**) is

$$S(t_1, \dots, t_k|y) = \exp[-y\{M_1(t_1) + \dots + M_k(t_k)\}] \quad (3)$$

where $M_j(t) = \int_0^t \mu_j(u) du$, $j = 1, 2, \dots, k$ are the integrated conditional hazards. From this, we immediately derive the bivariate survival function by integrating Y out

$$\begin{aligned} S(t_1, \dots, t_k) &= E \exp[-Y\{M_1(t_1) + \dots + M_k(t_k)\}] \\ &= L(M_1(t_1) + \dots + M_k(t_k)) \end{aligned} \quad (4)$$

where $L(\cdot)$ is the Laplace transform of the distribution of Y . Thus, the bivariate survivor function is easily expressed by means of the Laplace transform of the frailty distribution, evaluated at the total integrated conditional hazard.

Gamma distributions have been used for many years to generate mixtures in exponential and Poisson models. From a computational point of view, gamma models fit very well to survival models, because it is easy to derive the formulas for any number of events. This is because of the simplicity of the derivatives of the Laplace transform. This is also the reason why this distribution has been applied in most of the applications published until now. For many calculations, it makes sense to restrict the scale parameter (θ) and shape parameter (α). The standard restriction is $\theta = \alpha$, as this implies a mean of one for Y , when $\alpha \rightarrow \infty$, the distribution becomes degenerate. The probability density function (PDF) of a gamma distribution with a single parameter (α) is as follows:

$$f(y) = \alpha^\alpha y^{\alpha-1} \exp(-\alpha y) / \Gamma(\alpha) \quad (5)$$

The Laplace transform of gamma is

$$E\{e^{-sY}\} = \left[1 + \frac{s}{\alpha}\right]^{-\alpha} \quad (6)$$

Hanagal [11] proposed the BVW regression models with gamma distribution as a frailty for the survival data and estimated the frailty parameter and also the parameters of BVW model.

In some situations, the gamma frailty model may not fit well. Hougaard [12] proposed a positive stable model as a useful alternative, in part because it has the attractive feature that predictive hazard ratio decreases to one over time. The property is observed in familial associations of the ages of onset of diseases with etiologic heterogeneity, where genetic cases occur early and long-term survivors are weakly correlated. The gamma model has predictive hazard

ratios, which are time invariant and may not be suitable for these patterns of failures.

The pdf of positive stable distribution with single parameter (α) is given by

$$\begin{aligned} f(y) &= \frac{1}{\pi} \sum_{k=1}^{\infty} \frac{\Gamma(k\alpha + 1)}{k!} \left(-\frac{1}{y}\right)^{\alpha k + 1} \sin(\alpha k \pi), \\ y &> 0, \quad 0 < \alpha < 1 \end{aligned} \quad (7)$$

with Laplace transform

$$E\{e^{-sY}\} = \exp(-s^\alpha) \quad (8)$$

Hanagal [13] proposed the BVW regression models with positive stable distribution as a frailty for the survival data and estimated the frailty parameter and also the parameters of the BVW model.

The power variance function (PVF) distribution is used as another alternative for frailty. It is a three-parameter family uniting gamma and positive stable distributions. Inverse Gaussian and noncentral gamma distributions are also submodels of the PVF distribution.

Yashin and Iachine [14] developed correlated frailty models for the analysis of bivariate failure time data, in which two associated random variables are used to characterize the frailty effect for each pair. For example, one random variable is assigned for partner 1 and one for partner 2 so that they would no longer be constrained to have a common frailty. These two variables are associated and have a joint distribution. Knowing one of them does not necessarily imply knowing other. Examples of the use of multivariate frailty models are various and are listed below.

1. Aalen and Tretli [15] analyzed the incidence of testis cancer by means of a frailty model based on data from the Norwegian Cancer Registry collected during 1953–1993. The incidence of testicular cancer is greatest among younger men, and then declines from a certain age. The frailty is considered to be established by birth, and due to a mixture of genetic and environmental effects. The idea of the frailty model is that a subgroup of men is particularly susceptible to testicular cancer, which results in selection over time. This would explain why testis cancer is primarily a disease of young men. Aalen and Tretli [15]

applied the compound Poisson frailty distribution to testicular cancer data.

2. Malignant melanoma (skin cancer) was studied at the University Hospital of Odense in Denmark. Hougaard [16] compared the traditional Cox regression model with a gamma frailty and PVF frailty model, respectively to these data.
3. An example deals with the time from insertion of a catheter into dialysis patients until it has to be removed because of infection. A subset of the complete data, including the first two infection times of 38 patients, was published by McGilchrist and Aisbett [17] who used lognormal frailty model. To account for the heterogeneity within the data, Hougaard [16] used a gamma frailty model.

References

- [1] Marshall, A.W. & Olkin, I. (1967). A multivariate exponential distribution, *Journal of the American Statistical Association* **62**, 30–44.
- [2] Freund, J.E. (1961). A bivariate extension of exponential distribution, *Journal of the American Statistical Association* **56**, 971–977.
- [3] Proschan, F. & Sullo, P. (1974). Estimating the parameters of bivariate exponential distributions in several sampling situations, in *Reliability and Biometry*, F. Proschan & R.J. Serfling, eds, SIAM, Philadelphia, pp. 423–440.
- [4] Block, H.W. & Basu, A.P. (1974). A continuous bivariate exponential extension, *Journal of the American Statistical Association* **69**, 1031–1037.
- [5] Lee, L. (1979). Multivariate distributions having Weibull properties, *Journal of Multivariate Analysis* **9**, 267–277.
- [6] Gumbel, E.J. (1960). Bivariate exponential distribution, *Journal of the American Statistical Association* **55**, 698–707.
- [7] Lu, J.C. (1989). Weibull extensions of the Freund and Marshall-Olkin bivariate exponential models, *IEEE Transactions on Reliability* **38**, 615–619.
- [8] Hanagal, D.D. (1996). A multivariate Weibull distribution, *Economic Quality Control* **11**, 193–200.
- [9] Basu, A.P. (1971). Bivariate failure rate, *Journal of the American Statistical Association* **66**, 103–105.
- [10] Clayton, D.G. (1978). A model for association in bivariate life tables and its applications to epidemiological studies of familial tendency in chronic disease incidence, *Biometrika* **65**, 141–151.
- [11] Hanagal, D.D. (2006). A gamma frailty regression model in bivariate survival data, *IAPQR Transactions* **31**, 73–83.
- [12] Hougaard, P. (1986). A class of multivariate failure time distributions, *Biometrika* **73**, 671–678.
- [13] Hanagal, D.D. (2005). A positive stable frailty regression model in bivariate survival data, *Journal of Indian Society for Probability and Statistics* **9**, 35–44.
- [14] Yashin, A.I. & Iachine, I.A. (1995). Genetic analysis of durations: correlated frailty model applied to survival of Danish twins, *Genetic Epidemiology* **12**, 529–538.
- [15] Aalen, O.O. & Tretli, S. (1999). Analysing incidence of testis cancer by means of a frailty model, *Cancer Causes and Control* **10**, 285–292.
- [16] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer, New York.
- [17] McGilchrist, C.A. & Aisbett, C.W. (1991). Regression with frailty in survival analysis, *Biometrics* **47**, 461–466.

DAVID D. HANAGAL

Natural Resource Management

Risk assessment for natural resource management (NRM) deals with the *uncertainty* of our *impact* on the natural environment. Economic growth, development, and individual prosperity generally occur at the expense of environmental conservation and protection. The world's population is about 6.5 billion [1] and at its present rate of growth, will hit 10 billion by 2040. This will create unprecedented pressures on an already stressed environment. The realization that rates of consumption of even so-called renewable resources outstrip rates of production has been slow in coming. Actions to halt or reverse the trend have been even slower. The Central Intelligence Agency (CIA) [1] identifies rapid depletion of nonrenewable mineral resources, depletion of forests and wetlands, plant and animal extinctions, and the deterioration of air and water quality as among the most serious long-term environmental problems (*see* **Air Pollution Risk; Water Pollution Risk**). Human-induced climate change ranks as perhaps the single most pressing issue confronting mankind (*see* **Global Warming**). While degrading the environment, we simultaneously seek continuity of "ecosystem services" from it. Our list is long. We want clean air, rivers, streams, oceans, estuaries lakes, coasts, aquifers, and soil. We want safe, reliable, and clean supplies of drinking water and food. And we want our human, industrial, and household wastes to be disposed of safely (but "not in my backyard") while ensuring habitat and species conservation, environmental restoration, long-term sustainability, no adverse impacts on human health, and a full and comprehensive assessment of risks (*see* **Hazardous Waste Site(s); What are Hazardous Materials?**). Clearly, there are trade-offs that need to be acknowledged and balances that need to be struck.

Strategies that aim to reconcile economic growth and development with sustainability and environmental impact are characterized by risk. There are a number of facets to the risk equation: (a) uncertainty in all its forms (epistemic, linguistic, model); (b) model incertitude – how do we formulate and calibrate models to describe events that have never occurred and for

which no data are available? – and (c) natural variability. Furthermore, as with any modeling approach, risk assessments may be compromised by an inability or reluctance to enumerate all possible consequences (e.g., the effects of smoking on humans was known by tobacco companies, but not revealed, while the effect of polychlorinated biphenyls (PCBs) in the environment only became evident years after they were first used). Other issues also arise, such as how to assess the quality of different risk assessors and different estimates and predictions (e.g., in 1968 Paul Ehrlich claimed millions of people would starve to death because of overpopulation [2], while Galtung [3] concluded that "experts . . . were remarkably wrong" in their predictions) and the difficulty in answering the question "how well did we do" after adopting a risk-based approach to environmental decision making.

To manage natural resources and assets is to manage risk. For example, we build dams to guarantee water supply, but do so at the risk of decreasing biodiversity, altering in-stream species composition, and affecting riparian vegetation further downstream. The omnipotent threat of climate change and climate variability only serves to increase uncertainties and exacerbate consequences. Thus, in countries such as Australia where climate change has resulted in prolonged droughts, the level of uncertainty in water resource management is extraordinarily high. The accurate quantification of natural resources is central to a risk-based approach to NRM, as is the provision of reliable estimates of variability and uncertainty in these estimates. This, it is acknowledged, is easier said than done because reliable quantification, estimation, and prediction for natural resource management is rendered particularly difficult by virtue of two important characteristics of natural systems: complexity and uncertainty. Complexity arises from highly nonlinear relationships, feedback loops, hysteresis, multiple causes and effects, and timescales ranging from fractions of a second to millions of years. On the other hand, uncertainty in natural systems is often the result of one or more of the following factors: an event whose consequences are predictable (e.g., flood, drought, earthquake), but whose timing and magnitude are not; a discontinuity in an otherwise smooth trend (e.g., rapid change in pH when the buffering capacity of a water body is exceeded); or, as we have

already noted, unanticipated consequences of deliberate actions.

Traditional compliance-based approaches to pollution control focus on setting environmental standards whereby “safe” concentrations, loads, or exposure levels were identified and used as a not to be exceeded, regulatory limit. The approach focused on limiting public exposure to specific pollutants and did very little to identify pollution prevention measures that prohibit whole families of dangerous pollutants from being produced in the first place [4]. As noted by Fox [5], this resulted in data rich–information poor environment protection agencies that were ill equipped to make more comprehensive and holistic assessments of the environment. During the 1980s, risk-based approaches to NRM became increasingly popular, although as noted elsewhere in this encyclopedia (*see Ecological Risk Assessment*), these approaches were hampered by a lack of agreement as to what actually constituted a risk assessment, imprecise language, and inconsistent modes of analysis and application of risk methodologies.

Setting Environmental Standards for NRM

NRM encompasses environment protection, but has a broader scope. It is about balancing competing risks and demands, while ensuring the long-term sustainability of the biosphere or some part of it. The duality between *management* and *protection* has resulted in the establishment of environmental standards. The “fixed lines in the sand” that characterized the command-and-control approach to environmental regulation in some areas of NRM have given way to more flexible “standards”, “guidelines”, and “trigger values”. This has been necessary because high levels of complexity, uncertainty, and background variation mean that transgressions of fixed lines will occur even in the absence of anthropogenic influences. For example, storm events often result in exceedingly high concentrations and loads of sediments and nutrients in rivers and streams; upwelling in oceans may be responsible for abnormally high chlorophyll levels at the surface; while complex population dynamics may result in uncharacteristic abundances of some (nuisance) species that are unrelated to human impact. The important feature of a standard, guideline, or trigger value is that it is not so much the “violation” of the numerical target that is important,

but rather the *frequency* with which such violations occur. This notion immediately moves us away from prosaic “in compliance–out of compliance” assessments and more into the realm of *risk*. As noted by Barnett and O’Hagan [6], management standards need to be set in the context of an understanding of the processes involved and not the result of a somewhat arbitrary scaling of a result adapted from elsewhere. For example, prior to the release of revised water quality guidelines [7], many threshold values for pollutants and toxicants in marine and freshwaters were established using the “assessment factor” technique. Using this procedure, an endpoint (i.e., some outcome of a toxicity experiment) concentration for an organism or animal was scaled by a “factor” – typically orders of magnitude (10, 100, 1000 etc.) – so as to provide a margin of safety for humans. Notwithstanding the arbitrariness associated with the choice of the numerical factor value, no consideration was given to how relevant, meaningful, or achievable the resulting concentration was (*see Cumulative Risk Assessment for Environmental Hazards*). Nor did it provide any quantification of the level of protection afforded to humans and other species. In recognition of these deficiencies, the Australian guidelines adapted the risk-based approach of Aldenberg and Slob [8]. The idea was to use a theoretical distribution fitted to a sample of no observable effect concentrations (*NOECs*). This is the smallest concentration at which an “effect” (in terms of a designated endpoint) is observed from which a small percentile (typically 5%) was estimated. The resulting figure is referred to as a *trigger value* since an exceedance does not result in punitive action but rather it triggers a next-level investigation to explain why such a result was observed. A distinguishing feature of the trigger value is its attempt to tie the concentration to an effect in the population. The underlying assumption that a trigger value obtained as the p th percentile from the *NOEC* distribution will be protective of $p\%$ of all species in the environment is, nevertheless, contentious [9].

NRM, the Precautionary Principle, Neyman–Pearson Hypothesis Testing and Statistical Power

Countries around the world have, to varying degrees, embraced the *Precautionary Principle* (PP) as a guiding philosophy for risk-based NRM. A number of

variants of the PP exist; however, the basic tenet is that it seeks to avert or limit the risk of serious or irreversible harm to humans or the environment in the absence of full scientific certainty about that harm [10]. The critical issue is not whether we should be precautionary, but how precautionary we need to be on a case-by-case basis [11]. Poor risk management arises not only from insufficient levels of precaution, but equally from the pursuit of excessively high and unwarranted levels of precaution [11]. As we have seen in earlier sections of this article, uncertainty is a hallmark of NRM. The PP is one approach to dealing with uncertainty. A number of others exist such as prudent reduction, principle of prevention, best available techniques not entailing excessive cost (BATNEEC), best practicable environmental option (BPEO), and as low as reasonably achievable (ALARA). A distinguishing (and contentious) feature of the PP is that it shifts the onus of proof. Under the PP, proponents of a proposed activity (e.g., construction of a dam, erection of a mobile phone tower, deepening of a shipping channel, logging of a forest, construction of a housing estate) may be required to demonstrate that the activity has no serious or unacceptable environmental or human-health implications. Historically, the burden of proof has required those *opposing* the proposed activity to demonstrate that the activity is harmful. It is at this juncture, that risk assessments are usually produced in support of diametrically opposed conclusions. Statistical models can assist, although the use of different data and different modes of analysis are just two reasons behind the confusion and controversy that often accompanies a formal quantitative risk assessment (QRA). Conventional statistical hypothesis testing (or *null hypothesis testing* (NHT), as it is sometimes referred to), commences with a pair of hypotheses. The *null hypothesis* (“null” because it is meant to assume nothing or status quo conditions) is initially assumed to be true; the *alternative hypothesis* is generally the compliment or negation of the null hypothesis and will only be adopted if the evidence (in the form of data) provides extremely low support for the null hypothesis. This level of support is embodied in the ubiquitous (and widely misunderstood and abused) *p value*. The declaration of a “statistically significant result” occurs when a statistical test procedure returns a “small” *p value*. In the context of NRM, many of these features of hypothesis testing are troublesome. For a start, how small is small when

it comes to *p values* and who decides? why is the minimization of a Type I error (the probability that the test incorrectly rejects a true null hypothesis) seen to be overwhelmingly more important than the minimization of a Type II error (the probability that a false null hypothesis is accepted)? is it sensible to adopt the convention of couching a null hypothesis in terms of “no effect” when applied to NRM? and finally, is the binary view of the world that the hypothesis testing framework imposes on us the best way to assess environmental condition? Notwithstanding these issues, the PP imposes a role reversal on the null and alternative hypotheses. A precautionary approach to hypothesis testing would commence with an hypothesis that asserts that the proposed action has deleterious consequences. The alternative hypothesis then states that the consequences are not deleterious. As with the PP, this hypothesis testing schism has no clear-cut solution or recommendations although Fox [12] provided details of a hybrid approach. Equally problematic is the issue of “ecological significance” – that is, quantifying the magnitude of a change in condition that would be ecologically important. Environmental scientists and natural resource managers often struggle with the specification of an ecologically important effect size, but this is mandatory if, as is increasingly being demanded by natural resource management agencies, the issue of *statistical power* is to be investigated.

Statistical power and sample size analyses are useful adjuncts in the *planning* stage of any investigation or study. An *a priori* power analysis requires the investigator to nominate an ecologically significant effect size, a measure of intrinsic variation in the parameter under study (usually a mean response), a sample size and a level of significance (the Type I error rate). Armed with this information, it is a fairly straightforward task to compute the probability that the hypothesis-test procedure will correctly reject a false null hypothesis. This probability is referred to as the *power* of the statistical test. Obviously, a test with high power is preferred over a low-powered test. Unfortunately, further confusion exists about how to calculate and interpret the results of a power analysis. This is hardly surprising, given the lack of consensus in the literature and the publication of flawed advice [13, 14]. In fact, Hoenig and Heisey [15] identified 19 independent articles published between 1983 and 1997 advocating the

use of *postexperiment* power analysis. Postexperiment power analysis often involves the computation of “observed power” – a flawed quantity which will invariably “explain away” a nonsignificant statistical result as an outcome of a low-powered statistical design.

Modes of Analysis: Frequentist or Bayesian

Until recently, quantitative risk analyses were underpinned by “conventional” modes of statistical analysis. Frequentist statistical inference makes up the bulk of all statistical methodology and undergraduate statistics courses devote considerable time to it. Neyman–Pearson hypothesis testing and regression techniques are the mainstays of environmental assessments that support natural resource management decisions. There is nothing intrinsically wrong with these methods. Provided the attendant assumptions are satisfied, conventional *t-tests* and analysis of variance (ANOVA) techniques are optimal (uniformly most powerful) – in other words, one can essentially do no better with the available resources. Unfortunately, although diagnostic analyses are available to help assess the adequacy of the statistical method employed, these are oftentimes overlooked or ignored. This is seen as a major shortcoming of current NRM risk analyses. The inappropriate use of statistical models can result in fatally flawed natural resource management decisions.

Frustrated with actual and perceived limitations of the frequentist dogma, many environmental scientists are turning to information theory [16] and *Bayesian statistics* [17] (see **Bayesian Statistics in Quantitative Risk Assessment**) as alternative paradigms that are more aligned with the goals of risk-based NRM. Bayesian methods are seen as attractive if for no other reason than an explicit recognition of the legitimacy of “subjective” probability in the form of a *prior* probability distribution. In the context of NRM, a prior probability distribution might, in fact, represent the diversity of (personal) belief about some proposition that is held by a range of stakeholders. Although this does not necessarily mean that a Bayesian approach is necessarily superior to other competing methodologies, it does have a certain appeal, inasmuch as it provides a transparent means of elicitation and incorporation of personal belief and/or expert opinion. To this extent, Bayesian

methods are useful and are being increasingly used as part of a risk-based approach to NRM decision making.

As noted by Schipper and Meelis [18], data used to underpin decisions about environmental condition are invariably collected successively in time and so it makes sense to analyze those data sequentially. Apart from their own work, relatively little seems to have been done along these lines.

Future Directions

An examination of the literature identifies follow-up monitoring as a significant gap in risk-based approaches to NRM. Thus, while we argue over the merits of the precautionary principle and the commensurate shift toward prudent foresight, we spend comparatively little time on asking “how well did we do” [19, 20]. As noted by Suter [21], “a common criticism of risk-based environmental management is that we do not know whether it is effective”. Further research is required to develop companion tools and techniques that allow risk assessors to gauge the effectiveness of individual risk-based NRM programs.

As noted in this article, Bayesian methods for risk-based NRM are becoming increasingly popular. Protocols for the elicitation of “expert opinion” require further development and refinement as do methods for ranking those opinions. This may require closer collaboration between psychologists and statisticians to develop methods for reliably extracting opinions to improve the quality of the risk analysis and increase stakeholder trust in the process.

Acknowledgment

The author acknowledges the helpful advice of an anonymous referee and Prof. Mark Burgman at the University of Melbourne for his thoughtful comments on an earlier draft of this article.

References

- [1] Central Intelligence Agency (2007). *World Factbook 2007*, Washington, DC.
- [2] Ehrlich, P.R. (1968). *The Population Bomb*, Ballantine Books, New York.
- [3] Galtung, J. (2002). What did the experts predict? *Futures* 35, 123–145.

- [4] Faber, D. (1998). The political ecology of American capitalism: new challenges for the environmental justice movement, in *The Struggle for Ecological Democracy*, D. Faber, ed, Guilford Press, New York, pp. 27–59.
- [5] Fox, D.R. (2006). Statistical issues in ecological risk assessment, *Human and Ecological Risk Assessment* **12**, 120–129.
- [6] Barnett, V. & O'Hagan, A. (1997). *Setting Environmental Standards: The Statistical Approach to Handling Uncertainty and Variation*, Chapman & Hall.
- [7] Australian and New Zealand Environment and Conservation Council, Agriculture and Resource Management Council of Australia and New Zealand (2000). *National Water Quality Management Strategy Paper No. 4, Australian and New Zealand Guidelines for Fresh and Marine Water Quality*, Canberra.
- [8] Aldenberg, T. & Slob, W. (1993). Confidence limits for hazardous concentrations based on logistically distributed NOEC toxicity data, *Ecotoxicology and Environmental Safety* **25**, 48–63.
- [9] Fox, D.R. (1999). Setting water quality guidelines – a statistician's perspective, *SETAC News*, Society for Environmental Toxicology and Chemistry, May 1999, pp. 17–18.
- [10] Cooney, R. (2004). *The Precautionary Principle in Biodiversity Conservation and Natural Resource Management: An Issues Paper for Policymakers, Researchers and Practitioners*, IUCN, Cambridge.
- [11] Hrudey, S.E. & Lewis, W. (2003). Risk management and precaution: insights on the cautious use of evidence, *Environmental Health Perspectives* **111**(13), 1577–1581.
- [12] Fox, D.R. (2001). Environmental power analysis – a new perspective, *Environmetrics* **12**, 437–449.
- [13] Reed, J.M. & Blaunstein, A.R. (1995). Biologically significant population declines and statistical power, *Conservation Biology* **11**(1), 281–282.
- [14] Thomas, L. & Juanes, F. (1996). The importance of statistical power analysis: an example from animal behaviour, *Animal Behaviour* **52**, 856–859.
- [15] Hoenig, J.M. & Heisey, D.M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis, *The American Statistician* **55**(1), 19–24.
- [16] Burnham, K.P. & Anderson, K.P. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*, Springer, New York.
- [17] Clark, J.S. (2005). Why environmental scientists are becoming Bayesians, *Ecology Letters* **8**, 2–14.
- [18] Schipper, M. & Meelis, E. (1997). Sequential analysis of environmental monitoring data: refined SPRT's for testing against a minimal relevant trend, *Journal of Agricultural, Biological and Environmental Statistics* **2**(4), 467–489.
- [19] Caughlan, L. & Oakley, K.L. (2001). Cost considerations for long-term ecological monitoring, *Ecological Indicators* **1**, 123–134.
- [20] Vos, P., Meelis, E. & Ter Keurs, W.J. (2000). A framework for the design of ecological monitoring programs as a tool for environmental and nature management, *Environmental Monitoring and Assessment* **61**, 317–344.
- [21] Suter II, G.W. (2001). Applicability of indicator monitoring to ecological risk assessment, *Ecological Indicators* **1**, 101–112.

Related Articles

Ecological Risk Assessment

Environmental Monitoring

DAVID R. FOX

Near-Miss Management: A Participative Approach to Improving System Reliability

In this article, the authors discuss the “near-miss” concept, present various definitions used in different settings, and illustrate the utility of the concept by giving examples of where near-misses are used for evaluation of a system’s risk potential. We then identify the important constituents of a near-miss management system (NMMS) to achieve a high level of effectiveness.

The “Near-Miss Concept”

The notion of a “near-miss” (*see Operational Risk Modeling*) can best be explained by observations made in the process industries. When adverse incidents have occurred, it is observed that for every serious accident, a large number of related prior incidents occurred that resulted in limited impact and an even larger number of related prior incidents happened that resulted in no loss or damage. This observation is captured in the well-known safety pyramid shown in Figure 1 [1], which has been substantiated by many examples.

Incidents at the pyramid pinnacle, referred to in this chapter as “accidents”, may result in injury and loss, major environmental impact, and/or significant business disruption. These incidents are often obvious, are brought forcefully to the attention of management, and are reviewed according to site protocols. “Near-misses” comprise the lower portion of the pyramid. These incidents, depending on the circumstances, either result in minor losses or have the potential for but do not result in a loss. Near-misses are often less obvious than accidents and are defined as having little, if any, immediate impact on individuals or processes. Near-miss concept covers a broad range of incidents. Although in most cases it is not difficult to differentiate between a near-miss and an accident, under some conditions, such as an adverse event with a considerable negative impact, an incident can be classified either as an accident or as a near-miss. Such uncertain conditions are indicated by

the overlap region in Figure 1 and can be handled by procedures designed either for accidents or for near-misses.

Despite their limited direct impact, near-misses provide insight into potential accidents that could happen. Therefore, near-misses can be considered “indicators of potential problems”, including those of system flaws. The following are examples of near-misses signaling system issues:

1. Identification and reporting of O-ring degradation and its direct link to takeoff temperature of the Space Shuttle Challenger were noted as early as 1982, well before its explosion in 1986 [2].
2. Identification and written reporting of corroded and leaking pipes were noted well before they caused the 1997 Hindustan refinery explosion, resulting in the death of 60 people and atmospheric release of 10 000 metric tons of petroleum-based products [3].
3. Reporting of eight “signals passed at danger” (SPAD) between 1993 and 1999 for Signal 109 was recorded before the Paddington train crash in England in which 31 people died [4].

Within the last decade, several new disciplines have accepted the “near-miss” concept as a tool to improve the reliability of risk management systems. Tracking near-misses enables the early identification of “risk” and provides an opportunity to mitigate potential losses. In this work, near-misses across different industries, and a variety of service operations and natural disasters, have been considered. In the process, differences in near-miss definitions and the utilization of near-miss approaches to assess and manage a system’s risk factors have emerged and have been evaluated. Industrial near-misses are defined as events with minimal consequences. In case of natural disasters, such as seismic events, hurricanes, floods, drought, heat waves, etc., near-misses are defined as events that nearly gave rise to a major catastrophe. For example, a hurricane that missed a highly populated region and damaged the environment in a neighboring unpopulated area can be considered as a near-miss from which regional governments can learn the damage potential and take necessary precautions to minimize consequences in the future.

Some of the events that near-misses are associated with are more controllable, such as designing a

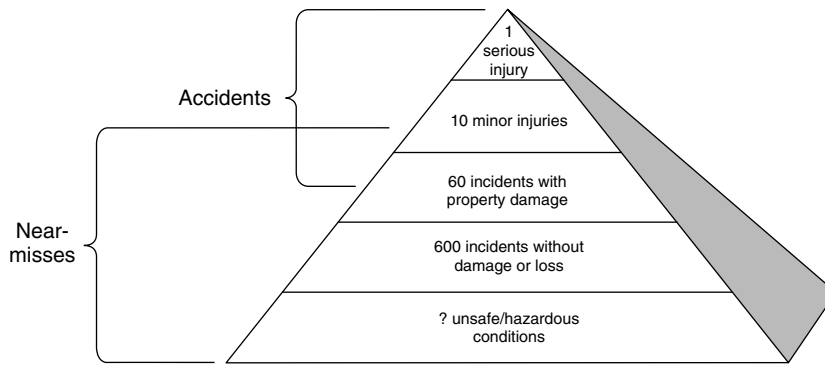


Figure 1 Safety pyramid (the lowest strata, unsafe conditions, is not shown by Bird and Germain) [Reproduced from [1] with permission from Det Norske Verita. © 1996.]

safer chemical process or a toy for a child, while others, such as hurricanes and earthquakes, cannot be controlled, but can be mitigated through various precautions designed and implemented in advance.

To reduce the likelihood of future accidents and to reduce system risk and improve process reliability, management and engineering systems that recognize operational weaknesses are being developed to seek and utilize accident precursors [5]. These programs operate under the umbrella of NMMS [6].

Although intensity and degree of control of events vary among different fields, the results of comparative analysis conducted using a wide selection of near-miss/accident data have shown significant similarities. Duffey and Saull [7] carried out an extensive study to examine the national and international reported error, incident, and fatal accident rates for multiple modern technologies, including shipping losses, industrial injuries, automobile fatalities, aircraft events and fatal crashes, chemical industry accidents, train derailments and accidents, medical errors, nuclear events, and mining accidents. Data for a time span of 20–200 years were collected and it was shown that all of the error data of different technologies follow a universal learning curve. That is to say, high error rates decline and are reduced exponentially to a minimum level as learning experiences (i.e., data) are accumulated.

The following examples are used to (a) demonstrate a variety of applications where near-misses are utilized, (b) identify influence of various stressors on the near-misses, and (c) understand the culture and functionality of the field to alter the key parameters that influence frequencies and consequences of

near-misses. Ashcroft *et al.* [8] conducted an analysis of dispensing errors and near-misses occurring in 35 community pharmacies in the United Kingdom, and concluded that for every 10 000 items dispensed, there are around 22 near-misses, and 4 dispensing errors. In another healthcare study, McDonald [9] analyzed near-misses arising from bar-coding systems in hospitals to improve patient safety. A national reporting system has been launched to examine the near-misses faced by fire fighters, and to improve their education, operations, and training [10]. Clarke *et al.* [11] performed a survey of 2287 nurses of medical surgical unit in 22 hospitals of United States and concluded that high workloads and poor organizational climates in hospitals cause increases in needle-stick injuries and other near-miss incidents affecting hospital nurses. Goldenhar *et al.* [12] analyzed near-miss and injury data for construction workers and examined their relationship with various job stressors (for example, job demand, training, safety climate, and exposure hours). Lilley *et al.* [13] examined the impact of organizational changes in the New Zealand Forestry Industry on the near-misses and accidents faced by forestry workers.

An NMMS helps to improve process reliability and reduce business risk. Muermann and Oktem [14] proposed using the “near-miss” concept as an advanced management approach for operational risk in financial institutions. In another study, Tucker and Spear [15] studied six US hospitals to examine the frequency of work system failures and their impact on nurse productivity. The five most common failures involved medications, orders, supplies, staffing, and equipment. They suggested that nurses’ effectiveness

can be increased by creating improvement processes to remediate the work system failures identified by near-misses.

A recent example that demonstrates how identification and management of near-miss could have saved significant business losses is the Sony battery recall case. Although Sony was made aware that there were some overheating issues in laptop computers as early as October 2005 and February 2006, they did not order a recall and continued using previously stocked batteries. This was clearly a near-miss, which was ignored by Sony. Later, Sony batteries were identified for potential incidents in Dell laptops, which then led to battery recalls, causing a 96% decrease in Sony's reported quarterly net profit. Even at this point, they could have treated the Dell incident as a near-miss (although clearly with more significant impact) and recalled the batteries from other companies like Fujitsu, Toshiba, Gateway, Dell, and Apple, all of whom were users of Sony batteries. However, the company continued to miss the signals from these early near-miss incidents, and, ultimately, the problems were revealed in other companies as well. Estimates indicate that this problem will cost Sony at least \$429 million before it is resolved, and possibly twice as much when the results of legal actions are considered. Sony could have avoided much of this expense and business risk had they recalled the batteries when they originally observed the overheating issues (near-misses). The dynamics of this case study illustrate the power and advantages of having an NMMS in place and demonstrate the tremendous negative outcome of not dealing with near-misses in a systematic way early on [16–18].

The identification and management of near-misses have many dimensions. Some events clearly test the limits of the near-miss concept. For example, Cunningham [19] analyzed a chemical spill of cyanide and heavy metals from a gold processing plant in 2000, which quickly moved from one river to the next through Romania, Hungary, the Federal Republic of Yugoslavia, and Bulgaria, killing tens of thousands of fish and other forms of wildlife, as well as poisoning drinking-water supplies. This case demonstrates the importance of information flow and interpretation of events. While the spill was considered a disaster by some, it was seen as a nonevent by others. This suggests that precursor events can go unnoticed or be viewed as catastrophic, or anything in between,

depending upon political, historical, and personal perspectives and awareness.

In the literature, there are only a limited number of publications that quantitatively demonstrate the impact of an NMMS to reliably identify risks and reduce the number of “accidents”. Jones *et al.* [20] provided an account of NMMSs successfully applied in the European chemical industries. Two examples of “near-miss” programs applied at Norsk Hydro's offshore and onshore facilities were studied. In both cases, their results suggest that an increase in near-miss reports can yield improved safety performance. In offshore drilling, over 7 years, a 10-fold increase in near-miss reporting corresponded to a 60% reduction in lost-time injuries. In onshore activities, over a 13-year span, an increase in reporting rates from zero to one-half per person per year corresponded to a 75% reduction in lost-time injuries [21]. When the authors discussed this issue with different corporate representatives, they found that in general there is hesitation toward reporting near-miss data due to possible legal ramifications.

Near-Miss Management Systems

NMMSs are designed to enable people and institutions to learn from high-frequency, low-impact incidents (near-misses) to prevent low-frequency, high-impact incidents (accidents). Near-misses are considered “weak signals”, some of which contain the signature of more serious, adverse effects. Hence, near-misses provide insight into potentially major, harmful conditions and business disruptions. Therefore, addressing near-misses in a timely and disciplined manner reduces the frequency and consequences of larger problems [20].

Phimister *et al.* [21] proposed a seven-step NMMS on the basis of an extensive research in the chemical industry. This was modified further by Muermann and Oktem [14] into an eight-step process by separating one of the important functions, “prioritization”, from the others. These eight steps are as follows:

1. identification (recognition) of a near-miss;
2. disclosure (reporting) of the identified information/incident;
3. prioritization and classification of the information for future actions;
4. distribution of the right level of information to the proper channels;

5. analysis of causes of the problem;
6. identification of solutions (remedial actions);
7. dissemination of actions to the implementers and (optional) to a larger group for their knowledge;
8. resolution (wrap-up) of all open actions and completion of reports.

Implicit in the above structure of near-miss reporting and analysis is the value associated with NMMS arising at two levels. First is the sequencing itself. The sequential structure of the above eight-step process underlines the fact that achieving maximum potential learning from near-misses requires the successful execution of each progressive stage. Thus, logically, near-misses that are not identified will not yield any value. Near-misses that are not disclosed may lead to some value for local operators, but will not yield any broader lessons unless these near-misses are analyzed and the information is distributed. This chain of implications led to the eight-step model that reflects the fact that early steps in the NMMS have “conjunctive” or multiplier effects on later steps. Second level of value associated with NMMS arises from the proper execution of each step itself. Such that near-misses that are recognized can have an immediate benefit at that site, as increasing awareness and initiating precautionary measures to prevent similar events, even if they are not reported, analyzed, and disseminated further. The following discussion is provided to clarify the nature and value of each of the eight steps.

Step 1: Identification

The first step of any NMMS is the recognition of a near-miss. As mentioned previously, near-misses are defined in a variety of ways by different authors [14, 21, 22]. Some definitions are very focused and are based on the extent of the potential negative consequences, such as “a near-miss is an undesired event or sequence of events with potential to cause serious damage”. To make the near-miss experience an occasion to learn from and to improve, rather than a punishable event, it is better to use a broader definition which balances the negative side of near-misses by focusing also on their positive contributions to risk reduction and system reliability. The following is a definition, similar to the ones used by Phimister *et al.* [21] and Muermann and Oktem [14] which fits to the above criteria of emphasizing the positive contribution of a near-miss:

A near-miss is an event, a sequence of events, or any other opportunity, such as an observation, that has the potential to improve system reliability.

This definition contains three important features of a near-miss:

- It views near-misses as improvement opportunities, which are positive experiences encouraging employees to report rather than to hide these events.
- It not only captures events but also includes observations.
- It includes all disturbances, from those which have the potential to cause serious damage to others that are inconveniences or elements of system operation or configuration that mainly cause inefficiencies.

It is important that each company develop its own near-miss definition and clearly communicate it to its employees to eliminate any confusion. The authors’ earlier studies indicate that not being able to recognize an incident as a near-miss is one of the primary reasons for lack of reporting. While the above broad definition is appropriate for certain risks, such as operational risk in financial institutions or in information technology, a much more focused definition may be necessary for a specific process industry. Tamuz *et al.* [23] examined the effect of definition and classification of safety-related events in hospitals on the organizational routines for gathering information, allocating incentives, and analyzing reported data. Their study suggests that better definition and classification procedures may help in improving patient safety, given the processes specific to patient management.

Step 2: Disclosure

The reporting of a near-miss should be made very simple to encourage everyone who observes or experiences a problem to fill out a report without spending much time and effort. Bridges [24] focused on barriers that inhibit disclosure. To increase report rates, Bridges advocates that nine barriers be overcome. These can be grouped into the four categories listed below:

1. potential recriminations for reporting (fear of disciplinary action, fear of peer teasing, and investigation involvement concerns);

2. motivational problems (lack of incentive and management discouraging near-miss reports);
3. lack of management commitment (sporadic emphasis and management fear of liability);
4. individual confusion (confusion as to what constitutes a near-miss and how it should be reported).

Classification and interpretation of near-misses during disclosure is very important for overall effectiveness of an NMMS and should be given special attention during its development. For example, a shared near-miss event reporting system has existed in the aviation industry for over 25 years [25] and was signaled as a model to improve healthcare safety. At first glance, both industries have a lot in common, such as highly trained professionals who work in teams and use technology that can also be a factor in undesirable outcomes. But, Tamuz and Thomas [26] argue that although some aspects of aviation safety systems are readily transferable to healthcare, other aspects, including professional controls and organizational complexities, can differ significantly, thereby influencing event classification and interpretation. If these differences are not properly accounted for, potentially dangerous events may not be recognized even if a shared system is established.

The structure of reporting systems plays a very important role in implementation of this step. During our studies of the chemical industry, the authors observed that having two separate reporting mechanisms, one for near-misses and another for accidents, causes confusion, and hence, misreporting of incidents that could potentially be classified in both categories (see Figure 1). On the other hand, having a single reporting mechanism, which is usually geared for accidents and requires extensive information, becomes too cumbersome for most near-misses, discouraging people from reporting small incidents. A possible solution is to have a single reporting system which enables the reporter to differentiate between various levels of information requirements based on the nature of the incident. This can be achieved through guidelines and reporting criteria defined by the risk management group of an institution.

Step 3: Prioritization

Although investigations have shown that almost all major incidents had precursors with minor or no consequences, not all minor incidents have the potential

to cause major accidents. Yet the abundance of minor incident reports creates a prioritization dilemma. An effective near-miss system should average at least one or more near-miss reporting per person per year. In a large organization or an industrial site, this may result in thousands of near-misses that need management's consideration. If the same intense attention is paid to every detail of each near-miss, then the cost of the NMMS will be overwhelming and unsustainable. In contrast, if very little attention is given to each near-miss, the major problems brewing in the system will be missed. Therefore, prioritization is a very critical step in establishing an effective near-miss system since this step determines, out of the large number of near-miss reports, which ones will require, and to what extent, the attention of the limited resources assigned to near-miss management.

NMMSs must identify criteria to enable such classification. Guiding principles for such a classification must be determined at the corporate level and further defined at application sites and local levels. It is important to monitor the system initially until a thorough understanding of the prioritization criteria is established among the employees, since they will, most likely, determine the priority at the time of disclosure. Training programs can play a very important role in accomplishing this objective.

Step 4: Distribution

There are two objectives of this step:

1. to inform those who will provide the right level of follow-up and
2. to alert a broader audience of the hazard.

Once near-miss information is reported into the NMMS, it should be directed to those who can act upon it. Identification of the right recipients strongly depends on the design of the "priority" system mentioned above.

Common obstacles that limit the effectiveness of this stage include unwillingness of supervisors to share near-miss information and systems that do not specify a time frame for review.

Step 5: Analysis

Once a near-miss is reported on the basis of the given priority, the reporting party, a supervisor, or a group

of experts related to the subject matter should identify direct and root causes of the problem to arrive at action(s) to eliminate the recurrence of this or similar incidents. The objective in root-cause analysis (*see Canonical Modeling; Reliability Integrated Engineering Using Physics of Failure*) is to determine the direct and underlying factors of an incident or unsafe condition. Short-term solutions resolve direct causes, while more permanent solutions rectify root causes. When structuring near-miss programs, it is important to recognize both the interaction between root-cause analysis and solution determination, and the separateness of these activities. Following are examples of analysis found in the literature.

Uth and Wiese [27] studied the accident reports collected by a German central organization created in 1993 which is called as a central body for collecting and evaluating major accidents (ZEMA). In 1994, they extended their charter to analyze minor and near-miss events as well, realizing that the analysis of incidents, and not just accidents, was also important. The root-cause analysis of the data revealed the importance of maintenance, detailed knowledge of chemical properties, and human factor issues as possible areas upon which to focus to improve system safety and reliability.

Ebright *et al.* [28] carried out an analysis of near-misses and adverse events faced by novice nurses. In some cases, novice nurses were unable to prevent near-miss/adverse events because of the lack of information to manage the entire situation and their inability to sort out the relevant data to act upon.

Weeks *et al.* [29] analyzed 30 near-miss maternal cases in Uganda. They found that financial barriers, problems with transportation, medical mistakes, and delays in referral in rural health centers were the socioeconomic causes behind maternal problems that, in some cases, led to mortalities. Analysis of near-misses is especially valuable in this context because of the difficulty in understanding the facts that led to the mortalities. This is unlike the process industries where electronic information is often available, describing the sequence of events, before an accident.

Step 6: Solution Identification

Once the causes of a problem are identified, the next step is to find viable solutions for each cause. Important considerations to factor include the following:

1. Solutions should be matched to causes, thus making sure that each cause has been addressed.
2. If possible, a member of the department who will implement the solution under discussion should be included in solution identification.
3. Identified solutions should be reviewed to ensure that they will not cause new problems.

This third item, which is called *change management* in some industries, is of particular relevance. For example, McDonald [9] showed that bar-coding systems established in hospitals to improve patient safety can create new errors, when not accompanied by well-designed cross-check processes and second-level human confirmations.

The overall goal of this step is to ensure that the right solutions are identified for the differing organizational needs. For example, Yacavone [30] analyzed mishaps in naval aviation from 1986 to 1990. A significant fraction of mishaps were due to aircrew errors and were related to human failure. As a result, aircrew coordination training (ACT) was instituted to address issues of human error (*see Reliability Data; Human Reliability Assessment; Probabilistic Risk Assessment*). Wiegmann and Shappell [31] examined the role of human error and crew–resource management failures in the United States. Naval aviation mishaps between 1990 and 1996. Approximately 56% of mishaps involved at least one crew–resource management failure and a large fraction of these occurred during nonroutine or extreme flight. These trends did not change even after implementation of ACT, indicating that the solutions have to be tailored and, if necessary, modified to specific problems that are faced in order to be most effective.

Step 7: Information Dissemination

Once an action (or set of actions) is determined,

1. it should be approved by the proper channels and assigned to a group to implement;
2. a larger group, including other divisions of the organization, customers, etc., may need to be informed of the incident and the actions taken.

An example of the latter is “Dear Doctor” letters that pharmaceutical companies send to physicians when they find a potentially negative effect associated with an already marketed drug and seek to warn them

about changing their practice (its dose, combined use with another drug, etc.) with respect to this particular application.

Step 8: Resolution

In the final step, all actions are completed including follow-up with proper departments and individuals. This means identification and tracking of all open actions and their follow-up with the right people for their closure. These activities may involve seeking approval from higher management ranks to obtain priority for implementation procedures. It is also highly desirable and rewarding to give feedback to the individual who identified the near-miss and/or reported it. Action-tracking software packages are available on the market and can be purchased separately, if tracking is not already built into the existing NMMS.

There are several variations of NMMSs that exist in different industries. For example, Dye and van der Schaaf [32] discuss a near-miss reporting system to manage the risk of customer dissatisfaction in the telecommunication industry, titled “Prevention and Recovery Information System for Monitoring and Analysis (PRISMA)”. In this context, events leading to customer loss due to dissatisfaction are considered as accidents that are injurious or damaging to the business relationship, while customer complaints are considered to be near-misses. The seven steps of PRISMA are detection, selection, description, classification, computation, interpretation and implementation, and evaluation. PRISMA’s reported financial benefits to companies include lower customer turnover, fewer failures, and a reduction in employee hours allocated to resolving customer problems.

Management, Implementation, and Continuous Improvement

Near-misses provide insight into (a) potential failure points within the system being analyzed and (b) weaknesses in the management system itself. As such new incidents, whether near-misses or accidents, can provide an opportunity to address and remedy potential weaknesses. This is particularly the case when new incidents appear to be “repeats” of previous events, for this indicates that one of the eight

stages may have failed to ensure hazards exposed by the previous incident were fully remedied. Hall [33] presents an example in arguing that similar to the Space Shuttle Challenger incident, the Space Shuttle Columbia disintegration in 2003 also displayed characteristics of organizational failure, even though significant improvements had been made since the previous incident. Similarly, Hallstrom and Smith [34] noted that hurricane ANDREW was a near-miss for hurricane RITA.

Although it is highly desirable to implement an eight-step NMMS, various conflicts can arise between management and those involved with reporting when selecting an NMMS that implements one or more of these eight steps. In this regard, Meel *et al.* [35] modeled the conflicts and trade-offs of management and engineering preferences as compared with process operator preferences using game theory (*see Risk Measures and Economic Capital for (Re)insurers; Mathematics of Risk and Reliability: A Select History; Group Decision*) to demonstrate the complexity of selecting an NMMS to be implemented in a chemical plant. Given various favorable and unfavorable views, the benefit of selecting an NMMS that satisfies the preferences of both is emphasized.

The auditing of an NMMS could provide significant insights into the benefits and weaknesses of the eight-step process. For example, Filippi *et al.* [36] carried out an in-depth analysis of near-misses in obstetrical emergencies among 12 hospitals in Benin, Cote d’Ivoire, Ghana, and Morocco by conducting near-miss audits. Steps of an audit process included (a) identification of near-miss cases, (b) selection of case(s) for audit, (c) interview(s) with the women involved, (d) interviews of persons involved in the management of case(s), (e) preparation of a medical summary for each case, (f) assessment of the quality of care provided, and (g) follow-up recommendations for improvement. These audits resulted in a wide range of positive changes in the procedures and resources available for the management of obstetric complications.

Quantitative Analysis of Near-Misses

In the future, as the benefits of NMMS are realized, large amounts of near-miss data sets are likely to be available. However, for now, many disciplines have not yet reached that level of data availability. The lack of near-miss data in industries could be due to

either the lack of efforts by management and safety personnel in collecting near-misses or the lack of availability of such data in the discipline of interest. For example, the health care industry has significantly more data than the process industries, which, in turn, have more data than natural disaster agencies.

The collection of data and their analysis are critical to improve system reliability and should, therefore, be encouraged. In their study, Rosenthal *et al.* [37] discuss the potential role of a facility's process safety management system (PSMS) in preventing low-probability, high-consequence (LP-HC) events. They note that a lack of process incident/accident data hinders predicting the characteristics of a facility's PSMS that can help in reducing "low-probability, high-consequence" accidents.

Having a rich database can help identify key issues. Kwon *et al.* [38] analyzed 5500 near-miss cases of a Korean chemical plant using KOSHA Accident Causation Management System (KACMS) and found that removal of human errors is the key issue that is necessary to prevent these near-miss cases and improve the process reliability.

Data on accident precursors are used to estimate the failure probabilities of the safety systems. Kaplan [39] introduced the concept of precursors and near-misses in quantitative risk assessment of space shuttles using Bayesian theory (*see Insurance Pricing/Nonlife; Imprecise Reliability; Lifetime Models and Risk Assessment*). Yi and Bier [40] developed a systematic approach using copula to extract maximum information from near-misses at a nuclear plant.

The complex and nonlinear interactions in chemical plants lead to occurrence of near-misses with varying consequences, some leading to accidents. This necessitated process engineers to seek robust and improved risk assessment tools to improve reliability. In search of more comprehensive accident prediction power, Meel and Seider [41] introduced an extended classification approach for abnormal events in chemical industries based on the safety pyramid in Figure 1. For systematic calculations with improved understanding of event propagation, they classified abnormal events into four categories in increasing order of severity of their consequences: (a) incipient faults, (b) near-misses, (c) incidents, and (d) accidents, and assumed that in a typical accident scenario abnormal events propagate in the above order. They used Bayesian theory to conduct statistical analysis of abnormal events (a) to improve system reliability and

(b) to perform dynamic risk assessment of continuous processes. More specifically, they developed the following tools to be implemented as part of an effective NMMS [41]:

1. Forecasting analyzer

Predicts the rate of occurrence of events in a time interval.

2. Reliability analyzer

Predicts human and equipment reliability by estimating the failure probabilities of human actions and equipment associated. Furthermore, performance of safety systems, which are crucial in preventing accidents, is assessed under abnormal events to identify the systems where reliability is lacking.

3. Accident closeness analyzer

Estimates the proximity of plant state with time to near-misses incidents and accidents using fuzzy clustering (*see Reliability Optimization*).

Summary

NMMSs can be a powerful tool to reduce the risk and improve the reliability of an organization, or a production, or a service process. Having a broad definition, but specific to the application environment, is important to achieve the best results from an NMMS. Other essential features of an effective NMMS include (a) encouraging the reporting of as many near-misses as possible (a guideline of 1 near-miss per person per year is often suggested), (b) management owning the program and utilizing the information to improve the system reliability, and (c) using both individual incident information and statistical analysis of collective data to improve system performance. As in other safety and risk areas, the central factor is the participation of all affected parties in the management of risk. This is perhaps the most important feature of an effective NMMS.

References

- [1] Bird, F.E. & Germain, G.L. (1996). *Practical Loss Control Leadership*, Det Norske Verita, Loganville.
- [2] Vaughan, D. (1996). *The Challenger Launch Decision: Risk Technology, Culture and Deviance at NASA*, University of Chicago Press, Chicago.
- [3] Khan, F.I. & Abbasi, S.A. (1999). The world's worst industry accident of the 1990s, *Process Safety Progress* **18**, 135–145.

- [4] Cullen, W.D. (2000). *The Ladbroke Grove Mail Inquiry*, Her Majesty's Stationary Office, Norwich.
- [5] March, J.G., Sproull, L.S. & Tamuz, M. (1991). Learning from samples of one or fewer, *Organization Science* **2**, 1–13.
- [6] Oktem, U. (2003). Near-miss: a tool for integrated safety, health, environmental and security management, *AIChE Spring Meeting, Loss Prevention Symposium*, New Orleans.
- [7] Duffey, R.B. & Saull, J.W. (2003). Errors in technological systems, *Human Factors and Ergonomics in Manufacturing* **13**(4), 279–291.
- [8] Ashcroft, D.M., Quinlan, P. & Blenkinsopp, A. (2005). Prospective study of the incidence, nature and causes of dispensing errors in community pharmacies, *Pharmacoepidemiology and Drug Safety* **14**(5), 327–332.
- [9] McDonald, C.J. (2006). Computerization can create safety hazards: a bar-coding near miss, *Annals of Internal Medicine* **144**(7), 510–516.
- [10] Dickinson, C. (2005). *National Fire Fighter Near-Miss Reporting System*, at <http://www.firefighternearmiss.com/home.do> (accessed 2007).
- [11] Clarke, S.P., Rockett, J.L., Sloane, D.M. & Aiken, L.H. (2002). Organizational climate, staffing, and safety equipment as predictors of needlestick injuries and near-misses hospital nurses, *American Journal of Infection Control* **30**(4), 207–216.
- [12] Goldenhar, L.M., Williams, L.J. & Swanson, N.G. (2003). Modeling relationships between job stressors and injury and near-miss outcomes for construction laborers, *Work and Stress* **17**(3), 218–240.
- [13] Lilley, R., Feyer, A.M., Kirk, P. & Gander, P. (2002). A survey of forest workers in New Zealand – do hours of work, rest, and recovery play a role in accidents and injury? *Journal of Safety Research* **33**(1), 53–71.
- [14] Muermann, A. & Oktem, U. (2002). The near miss management of operational risk, *Journal of Financial Risk* **4**(1), 25–36.
- [15] Tucker, A.L. & Spear, S.J. (2006). Operational failures and interruptions in hospital nursing, *Health Services Research* **41**(3), 643–662.
- [16] Oswald, E. (2006). *Sony's Battery Recall Expands Again*, http://www.betanews.com/article/Sonys_Battery_Recall_Expands_Again/1161701142 (accessed 2007).
- [17] Roberts, P.F. (2006). *Dell, Sony Discussed Battery Problem 10 Months Ago*, at http://www.infoworld.com/article/06/08/18/HNdellsnybattery_1.html (accessed 2007).
- [18] Wikipedia (2006). *Sony-Controversies-Batteries*, at <http://en.wikipedia.org/wiki/Sony#Batteries> (accessed 2007).
- [19] Cunningham, S.A. (2005). Incident, accident, catastrophe: cyanide on the Danube, *Disasters* **29**, 99–128.
- [20] Jones, S., Kirchsteiger, C. & Bjerke, W. (1999). The importance of near miss reporting to further improve safety performance, *Journal of Loss Prevention in the Process Industries* **12**(1), 59–67.
- [21] Phimister, J.R., Oktem, U., Kleindorfer, P.R. & Kunreuther, H. (2003). Near-miss incident management in the chemical process industry, *Risk Analysis* **23**(3), 445–459.
- [22] Barach, P. & Small, S.D. (2000). Clinical review–reporting and preventing medical mishaps: lessons from non-medical near-miss reporting systems, *British Medical Journal* **320**, 759–763.
- [23] Tamuz, M., Thomas, E.J. & Franchois, K.E. (2004). Defining and classifying medical error: lessons for patient safety reporting systems, *Quality and Safety in Health Care* **13**(1), 13–20.
- [24] Bridges, W.G. (2000). Get near-misses reported, process industry incidents: investigations protocols, case histories, lessons learned, *Centre for Chemical Process Safety International Conference and Workshop*, American Institute of Chemical Engineers, New York.
- [25] Reynard, W.D., Billings, C.E., Cheaney, E.S. & Hardy, R. (1986). *The Development of NASA Aviation Safety Reporting System*, NASA Reference Publication 1114, NASA Ames Research Center, Moffett Field.
- [26] Tamuz, M. & Thomas, E.J. (2006). Classifying and interpreting threats to patient safety in hospitals: insights from aviation, *Journal of Organizational Behavior* **27**(7), 919–940.
- [27] Uth, H.J. & Wiese, N. (2004). Central collecting and evaluating of major accidents and near-miss-events in the Federal Republic of Germany – results, experiences, perspectives, *Journal of Hazardous Materials* **111**(1–3), 139–145.
- [28] Ebright, P.R., Urden, L., Patterson, E. & Chalko, B. (2004). Themes surrounding novice nurse near-miss and adverse-event situations, *Journal of Nursing Administration* **34**(11), 531–538.
- [29] Weeks, A., Lavender, T., Nazziwa, E. & Mirembe, F. (2005). Personal accounts of 'near-miss' maternal mortalities in Kampala, Uganda, *British Journal of Obstetrics and Gynaecology: An International Journal of Obstetrics and Gynaecology* **112**(9), 1302–1307.
- [30] Yacavone, D.W. (1993). Mishap trends and cause factors in naval aviation – a review of naval-safety – center data, 1986–90, *Aviation Space and Environmental Medicine* **64**(5), 392–395.
- [31] Wiegmann, D.A. & Shappell, S.A. (1999). Human error and crew resource management failures in naval aviation mishaps: a review of US naval safety center data, 1990–96, *Aviation Space and Environmental Medicine* **70**(12), 1147–1151.
- [32] Dye, J. & van der Schaaf, T. (2002). PRISMA as a quality tool for promoting customer satisfaction in the telecommunications industry, *Reliability Engineering and System Safety* **75**(3), 303–311.
- [33] Hall, J.L. (2003). Columbia and Challenger: organizational failure at NASA, *Space Policy* **19**(4), 239–247.
- [34] Hallstrom, D. & Smith, V.K. (2005). Market responses to hurricanes, *Journal of Environmental Economics and Management* **50**, 541–561.

- [35] Meel, A., Seider, W.D. & Oktem, U. (2007). Analysis of management actions, human behavior, and process reliability in chemical plants, *Near-miss management system selection, Process Safety Progress* (In Press) **II**.
- [36] Filippi, V., Brugha, R., Browne, E., Gohou, V., Bacci, A., Brouwere, V.de., Sahel, A., Goufodji, S., Alihonou, E., & Ronsmans, C. (2004). Obstetric audit in resource-poor settings: lesson from a multi-country project auditing 'near miss' obstetrical emergencies, *Health Policy and Planning* **19**(1), 57–66.
- [37] Rosenthal, I., Kleindorfer, P.R. & Elliott, M.R. (2006). Predicting and confirming the effectiveness of systems for managing low-probability chemical process risks, *Process Safety Progress* **25**(2), 135–155.
- [38] Kwon, H., Hyungjoon, Y. & Moon, I. (2006). Industrial applications of accident causation management system, *Chemical Engineering Communications* **193**(8), 1024–1037.
- [39] Kaplan, S. (1990). On the inclusion of precursor and near miss events in quantitative risk assessments – a Bayesian point of view and a space-shuttle example, *Reliability Engineering and System Safety* **27**(1), 103–115.
- [40] Yi, W. & Bier, V.M. (1998). An application of copulas to accident precursor analysis, *Management Science* **44**(12), S257–S270.
- [41] Meel, A. & Seider, W.D. (2006). Plant-specific dynamic failure assessment using Bayesian theory, *Chemical Engineering Science* **61**, 7036–7056.

ULKU OKTEM AND ANJANA MEEL

No Fault Found

The No Fault Found (NFF) phenomenon occurs in systems where a system is tested after a failure is reported but no failure is found during the test.

The inability to isolate a system fault during a test may occur for a number of reasons that may be related to the state of the system itself, and the state of, and the methods of operation of, the test equipment. In trying to understand NFF all three areas have to be considered.

Systems are generally considered to be complex items that may contain people, hardware, and software and often contain subsystems. These subsystems may also contain a hierarchy of low-level subsystems until a basic component level is reached. At each system level there may well be a combination of components and lower level subsystems. In most cases the human interface and also the presence of a person interacting with the system are considered as part of the system. At each of these discrete levels and for each item in the system, a NFF situation may occur. These situations often occur across level boundaries where, for instance, a failure is detected at the top level – perhaps by built-in test (BIT) – which also identifies a faulty subsystem (*see Failure Modes and Effects Analysis; Reliability Integrated Engineering Using Physics of Failure*). This causes a repair to be performed, which swaps the identified subsystem for another one and the top-level system then works perfectly. When the removed subsystem (often called a *line replaceable unit, LRU or line replaceable item, LRI*) is tested, however, no fault is found. This is the classic NFF case.

Nomenclature

This basic phenomenon is known by a large number of terms, which are often abbreviated to three letter acronyms. The following names and acronyms are encountered in different industries and in the existing literature: No Fault Found (NFF), No Failure Found (NFF), No Trouble Found (NTF), Re-Test OK (RTOK), Trouble Not Identified (TNI), Cannot Duplicate (CND), No Fault Indicated (NFI), No Defect Found (NDF), No Problem Found (NPF), Bad Actor

(BA), Fault Not Found (FNF), Failure Not Found (FNF), False Alarm (FA), Cause Not Found (CNF), and similar acronyms [1–7]. In this article the term *NFF* will be used throughout.

Industrial Occurrence of NFF

A large number of organizations and companies have reported anecdotally that the level of NFF found in their systems is around 50% of removed items and this means that occurrence of NFF is a large problem for industry. Because of this high incidence, a number of formal studies have been performed that have attempted to quantify the extent of the problem. The aerospace industry has reported average figures as around 50% [1, 8–12] and in at least two specific cases 67% [5] and 85% [4]. The portable phone market, however, seems to suffer similar levels of NFF; Reference [13] states that *One in seven mobile phones are returned within the first year of purchase by subscribers as faulty. Further analysis into the nature of these returns reveals the even more disturbing statistic that 63% of the devices being returned are without fault*. Reference [14] states that this figure is 13% over the industry average for general consumer electronic devices, which makes this average the typical 50%. The train industry has also reported occurrence levels of around 50% [15] whereas one automotive manufacturer has reported NFFs as high as 80–90% with a particular device [2]. More generic, cross-sector work [16] reported figures from 0 to 53%, with an average – where there was at least one incidence of NFF – of around 45%. Other authors [17] have also reported averages between 25 and 50%, and one author [18] reports the averages to be as high as 80%.

Costs of NFF

Each time a system is repaired the same repair cost is encountered whether or not the repair is effective and the reported percentage suggests that up to 90% of the maintenance cost could be used up in dealing with NFF [4]. An actual estimate of the loss of money globally in the portable phone markets is given in [14] as \$4.5 billion in 2006. In the aerospace industry the F-16 program spent over \$10 million on NFF diagnostics in 2001 [5], and the air transport association reported that NFFs cost

member airlines at least \$100 million dollars per year, not including costs derived from network disruptions. In the automotive industry, because a reputation for intermittent faults can be so damaging to a company, often product recalls are performed to attempt to fix these sort of issues, and one manufacturer reported that the recall of over a million vehicles in 1987 for such an issue had cost them many millions of dollars [2].

Other effects of NFF

Quantitative measures of reliability (e.g., mean time between failures for systems (*see Design of Reliability Tests; Reliability Growth Testing*), mean time between removals) and availability (e.g., maintenance-free operating period) will be affected by the incidence of NFF, providing more pessimistic estimates that will bias analysis of the performance of operational systems and provide very conservative estimates as inputs to quantitative risk modeling.

Refitting a subsystem that potentially has an undiagnosed failure to a system may increase risks with implications for safety [4]. To avoid this issue, Thomas *et al.* [2] suggest that a company takes full responsibility for finding the root cause of all failures. This partly falls in line with the work reported by James *et al.* [9], which has developed guidance for the aerospace industry for dealing with NFF.

Possible Causes of NFF

It has been often suggested anecdotally that the NFFs in electronic and mechanical systems are caused by the volume of complex components present in a system. Complex components would include hybrid devices, microprocessors, large memories, and ICs that contain a large number of operational units. Often these devices will contain untested regions, which may contain faults and could be utilized during operation in special circumstances. Replication of the exact failure circumstances may not be possible during failure analysis due to time and cost constraints. The more complex a printed circuit board, the less likely it is that a diagnostic will be successful even if the components are in themselves quite simple [1, 2, 8, 9, 19–22]. This is particularly true in the aerospace and

other industries for complex control circuitry particularly where software forms an integral part of the system. However, the work done in [16] seems to indicate that this is not the case.

Another anecdotal reason proposed for the incidence of NFF is the presence of connectors on the circuit boards [9, 16]. Connectors can corrode and cause intermittent failures owing to increase in contact resistance caused by corrosion products. Often the removal of the board from the equipment, and later insertion of the board into test apparatus, may clean the connector contacts causing the board to function correctly again. Service engineers will often clean connector contacts as a routine before testing the boards and hence repair the intermittent fault. Again, the work done in [16] seems to indicate that this is not the case.

For the system that is not electronic, there has been no reported work to date investigating the potential causes of NFF although it is expected that these will have similar root causes, overly complex base units and the interface between such units.

A further possible cause of NFF in all cases is as a result of human interaction with the system [16]. It is possible that during the use of a new system, the operator's expectations of how the system should operate could possibly be different from the system's actual operating parameters. This could lead to the reporting of failures when no failure is in fact present. The examination of the data reported in [14] allows some investigation of these issues where it was found that, for returned failed units, 38% were from users abandoning devices after struggling to use them, 13% were from users who claimed that the device did not fulfill the purpose for which it had been purchased, and 49% were from users who diagnosed lack of service preconfiguration as a device fault.

Management of NFF

There are two reasons why a subsystem is classified as No Fault Found. Either there is no fault in the product or the system test fails to detect the fault. Where there is really no fault present in the subsystem, it is not surprising that the test fails to locate one; however, what is more surprising is the fact that the subsystem was rejected in the first instance. Where there is a fault present in the product

but the test fails to isolate the defect, it is likely that this subsystem will exhibit the same behavior when reinstalled on a system. In the latter case, it will be the effectiveness of the operator's recording and reporting systems that will determine whether the subsequent subsystem rejections can influence the level of diagnostic activity carried out during the test. If a particular subsystem is removed on a number of occasions for the same symptom, it is likely to be the cause of the failure. Rapid classification and good visibility of repeat removals or rogue units can provide an operator with important information. Besides providing an indication that the particular subsystem is actually faulty, it also highlights the ineffectiveness of the system test in this particular area.

There is a third reason why a subsystem may be classified as No Fault Found in testing (although it is actually a subset of those events with "no fault present"). Subsystems will be removed during scheduled maintenance. A scheduled removal will be tested to ensure that there are no subsidiary faults present. Subsystems that pass a test following a scheduled removal and any upgrade have been sometimes classified as NFF along with those subsystems removed during unscheduled maintenance.

Where there is really no failure contained within the subsystem, then the reason for its rejection must be reviewed in detail if the root cause of the system malfunction is to be located. Although most workshop facilities will carry out an extensive test regimen without the benefit of environmental stressing, it is understandable that if a product contains an intermittent fault, it has a lower likelihood of being detected during test. If the workshop operatives are not familiar with the idiosyncrasies of each product's test regimen, then the interpretation of results may play a major part in the success of the workshop. To effectively reduce NFF removals, a subsystem supplier must have direct dialogue with relevant workshop staff to elicit the factors that may influence workshop effectiveness and then drive for methods to improve the situation for future products.

It is important to acknowledge that any complex system will, in its early life, contain uncovered design faults. As experience with a new system grows, similarities in the symptoms of NFF removal events will begin to develop. Failures may also begin to manifest themselves where a particular operator uses

the system in an area of its envelope not previously used by other operators.

In order to confirm that the defect isolated/repared within the subsystem is indeed the root cause of the system malfunction, a number of options are possible. Correlation of the subsystem failure with the maintenance message or reported anomalous behavior is the most commonly used. If a mechanism cannot be postulated to link the specific subsystem failure to the system behavior, then it may be that the subsystem simply contains a subsidiary fault, with the root cause still on the system. For this reason it is extremely important that a structured technique is used to confirm if the subsystem rejection was responsible for the system malfunction, particularly where it affects safety. If the system malfunction disappears following the replacement of a subsystem, it is likely that this will be the cause of the problem. However, this is not always the case and extreme caution should be given to the assumption that a subsystem is faulty because its replacement cured the problem.

In order to be confident that the reason for rejection is confirmed in the workshop, there must be good correlation between the system defect definition and the fault detected within the subsystem. This correlation should include test failure(s) within the workshop, an Failure Mode and Effect Analysis FMEA-type analysis to give potential system behavior for given subsystem faults, and the clearance of the fault upon the subsystem rejection.

References

- [1] Beniaminy, I. & Joseph, D. (2002). Reducing the 'no fault found' problem: contributions from expert-system methods, *IEEE Aerospace Conference, Proceedings*, Big Sky, pp. 6-2971–6-2973.
- [2] Thomas, D., Ayers, K. & Pecht, M. (2002). The "trouble not identified" phenomenon in automotive electronics, *Microelectronics Reliability* **42**, 641–651.
- [3] Baskoro, G., Rouvroye, J.L., Bacher, W. & Brombacher, A.C. (2003). Developing MESA: an accelerated reliability test, *Proceedings of Annual Reliability and Maintainability Symposium (RAMS 2003)*, Tampa, pp. 303–308.
- [4] Williams, R., Banner, J., Knowles, I., Dube, M., Natischan, M. & Pecht, M. (1998). An investigation of 'cannot duplicate' failures, *Quality and Reliability Engineering International* **14**(5), 331–337.

- [5] Steadman, B. & Sievert, S. (2005). Attacking “bad actor” and “no fault found” electronic boxes, *AUTOTESTCON 2005 Conference*, Orlando, September 26–29, 2005.
- [6] Lu, Y., Den Ouden, E., Brombacher, A., Geudens, W. & Hartman, H. (2005). Analysing highly uncertain user-product interaction in product development, *4th International Conference on Quality and Reliability (ICQR 2005)*, Beijing, August 9–11, 2005.
- [7] Koca, A., Yuan, L., Brombacher, A.C. & Hartmann, J.H. (2006). Towards establishing foundations for new classes of reliability problems concerning strongly innovative products, *2006 IEEE International Conference on Management of Innovation and Technology (ICMIT 2006)*, Vols 1 and 2, June 21–23, 2006, Singapore.
- [8] Sudolsky, M.D. (1998). The fault recording and reporting method, *AUTOTESTCON Proceedings*, Salt Lake City, pp. 429–437.
- [9] James, I., Lombard, D., Willis, I. & Goble, J. (2003). Investigating no fault found in the aerospace industry, *Proceedings of Annual Reliability and Maintainability Symposium (RAMS 2003)*, Tampa, pp. 441–446.
- [10] Lee, L. & Andrews, J. (1999). Satellite hierarchical system test using IEEE 1149.1-based COTS test tools: a case history with results and lessons, *AUTOTESTCON 1999 Proceedings*, San Antonio, pp. 193–201.
- [11] Steadman, B., Pombo, T., Shively, J. & Kirkland, L. (2002). Reducing no fault found using statistical processing and an expert system, *AUTOTESTCON 2002 Proceedings*, Huntsville, pp. 872–879.
- [12] Sorensen, B. (2006). New air force testing methods of No Fault Found (NFF) proves highly successful, *ERI News* **23**, 1–3.
- [13] WDSGlobal <http://whitepapers.silicon.com/0,39024759,60262206p39000410q,00.htm> (accessed Nov 2006).
- [14] http://pda.mobileeurope.co.uk/news/news_story.ehtml?o=2314 (accessed Nov 2006).
- [15] Turner, D.B. (1979). Reliability improvement of BART vehicle train control, *Proceedings of the 29th IEEE Vehicular Technology Conference, 1979*, Arlington Heights, pp. 279–288.
- [16] Jones, J. & Hayes, J. (2001). Investigation of the occurrence of no-faults-found in electronics, *IEEE Transactions on Reliability* **50**(3), 289–292.
- [17] Bothe, K.R. & Bothe, A.R. (2000). *World Class Quality, Using Design of Experiments to Make it Happen*, 2nd Edition, Amacon.
- [18] O'Connor, P.D.T. (2001). *Test Engineering, A Concise Guide to Cost-Effective Design, Development and Manufacture*, ISBN: 0-471-49882-3, Hardcover, John Wiley & Sons, p. 288.
- [19] Zavatto, D.M., Gibson, J. & Talbot, M. (1998). C-17 portable advanced diagnostic system, *AUTOTESTCON Proceedings*, Salt Lake City, pp. 536–541.
- [20] Patel, V.C., Kadiramanathan, V. & Thompson, H.A. (1996). A novel self-learning fault detection system for gas turbine engines, *IEE Conference Publication*, Exeter, pp. 867–872.
- [21] Austin, J.W. & Weimer, R.J. (1987). *Service And Support of Electronic Products in the Trucking Industry*, SAE Technical Papers Series, 1987, SAE, Seattle.
- [22] James, I. (1999). Learning the lessons from in-service rejection, *IEE Colloquium Digest* **189**, 21–24.

JEFFREY A. JONES

Non- and Semiparametric Models and Inference for Reliability Systems

Components and Systems

The assessment of the risk of failure and the quality of deployed systems is of great importance in many settings, such as in the engineering, biomedical, public health, military, economic, and business arenas. It is therefore imperative to have probabilistic models and statistical inference methods for assessing the risk and quality of systems. Most of these systems can be conceptually modeled as reliability systems.

A reliability system (see **Multistate Systems**) is composed of a finite number of components, with such components possibly subsystems themselves. Such a system could be a mechanical, an electronic, a military, a biomedical, an economic, or even a managerial decision-making system. Such systems usually exist for the purpose of performing certain specific tasks. For a reliability system with K components, denote by $\mathbf{x} = (x_1, x_2, \dots, x_K)$ the state vector of the components, with $x_i \in \{0, 1\}$, and such that $x_i = 1$ if and only if component i is functioning. The structure function of a reliability system is the function $\phi: \{0, 1\}^n \rightarrow \{0, 1\}$ such that $\phi(\mathbf{x})$ indicates whether the system is in a functioning state ($\phi(\mathbf{x}) = 1$) or is in a failed state ($\phi(\mathbf{x}) = 0$). If the structure function ϕ satisfies the conditions that (a) it is nondecreasing in each argument, and (b) each component is relevant in the sense that, for each $i \in \{1, 2, \dots, K\}$, there exists an $\mathbf{x} \in \{0, 1\}^n$ such that $0 = \phi((0_i, \mathbf{x})) < \phi((1_i, \mathbf{x})) = 1$, where $(0_i, \mathbf{x}) = (x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_K)$ and $(1_i, \mathbf{x}) = (x_1, \dots, x_{i-1}, 1, x_{i+1}, \dots, x_K)$, then the reliability system is said to be a coherent system (see **Reliability Optimization**). We refer the reader to the excellent monograph of Barlow and Proschan [1] for a comprehensive treatment of coherent reliability systems.

Two common examples of coherent reliability systems are the series system whose structure function is $\phi_{\text{ser}}(\mathbf{x}) = \min(x_1, x_2, \dots, x_K)$ and the parallel system whose structure function is $\phi_{\text{par}}(\mathbf{x}) =$

$\max(x_1, x_2, \dots, x_K)$ (see **Game Theoretic Methods**). A more general coherent structure function is the k -out-of- K system whose structure function is $\phi_{k:K}(\mathbf{x}) = I\{\sum_{j=1}^K x_j \geq k\}$ with $I(A) = 0$ or 1 depending on whether event A does not or does hold, respectively. Thus, a k -out-of- K system is functioning if and only if at least k of its K components are functioning. Clearly, a series system is a K -out-of- K system, whereas a parallel system is a 1 -out-of- K system. Another coherent system, which is used to illustrate the procedures described in the sequel, is the series-parallel system whose structure function is $\phi_{\text{sp}}(x_1, x_2, x_3) = \min\{x_1, \max\{x_2, x_3\}\}$. This system functions so long as component 1 and at least one of components 2 and 3 are functioning.

Denote by X_i the random variable indicating whether component i is in a functioning state, and let $p_i = \Pr\{X_i = 1\}$, $i = 1, 2, \dots, K$. Assume that $X_1, X_2, \dots, X_K$ are independent random variables, and define $\mathbf{X} = (X_1, X_2, \dots, X_K)$ and $\mathbf{p} = (p_1, p_2, \dots, p_K)$. The reliability function associated with the structure function ϕ is defined by

$$h_\phi(\mathbf{p}) = E\{\phi(\mathbf{X})\} = \Pr\{\phi(\mathbf{X}) = 1\} \quad (1)$$

This reliability function represents the probability that the system is functioning. The series system has the reliability function $h_{\text{ser}}(\mathbf{p}) = \prod_{j=1}^K p_j$, while the reliability function of the parallel system is $h_{\text{par}}(\mathbf{p}) = 1 - \prod_{j=1}^K (1 - p_j)$. For the more general k -out-of- K system, the reliability function is

$$h_{k:K}(\mathbf{p}) = \sum_{\{(x_1, x_2, \dots, x_K) \in \{0, 1\}^K; \sum_{j=1}^K x_j \geq k\}} \prod_{j=1}^K p_j^{x_j} \quad (2)$$

In the special case where $p_j = p$, $j = 1, 2, \dots, K$, this reliability function simplifies to $h_{k:K}(p) = \sum_{j=k}^K C(K, j) p^j (1 - p)^{K-j}$, where $C(K, j)$ is the combination of K items taken j at a time. Thus, in this case, $h_{\text{ser}}(p) = h_{K:K}(p) = p^K$ and $h_{\text{par}}(p) = h_{1:K}(p) = 1 - (1 - p)^K$. On the other hand, for the series-parallel system, the reliability function is $h_{\text{sp}}(p_1, p_2, p_3) = p_1[1 - (1 - p_2)(1 - p_3)]$, which simplifies to $h_{\text{sp}}(p) = p(1 - (1 - p)^2)$ when all three components have the same reliability p .

The structure function and the reliability function are characteristics of the reliability system at a specific point in time. Because it is usually of interest to determine if the system will be able to accomplish

a specific task within a given period of time, it is fruitful to consider the evolution of the system as time moves forward. This entails the examination of the component and system lifetimes (*see Reliability Data; Extreme Values in Reliability*).

Nonparametric Models

Denote by $\mathbf{T} = (T_1, T_2, \dots, T_K)$ the vector of component lifetimes and by S the system lifetime. For a given time t , the state vector of the components is given by

$$\mathbf{X}(t) = (I(T_1 > t), I(T_2 > t), \dots, I(T_K > t)) \quad (3)$$

hence the state of the system at time t is given by $\phi(\mathbf{X}(t))$. Consequently, $\{S > t\} = \{\phi(\mathbf{X}(t)) = 1\}$. The system lifetime survivor function is therefore

$$\bar{F}_\phi(t) = \Pr\{S > t\} = \Pr\{\phi(\mathbf{X}(t)) = 1\} = E\{\phi(\mathbf{X}(t))\} \quad (4)$$

If the component lifetimes are independent and their survivor functions are given by $\bar{F}_j(t) = \Pr\{T_j > t\}$, $j = 1, 2, \dots, K$, then the system survivor function could be expressed in terms of the system's reliability function *via*

$$\bar{F}_\phi(t) = h_\phi(\bar{F}_1(t), \bar{F}_2(t), \dots, \bar{F}_K(t)) \quad (5)$$

A special case, for instance, is when the component lifetimes are independent exponentially distributed random variables with the j th component having mean lifetime of $1/\lambda_j$. Then $\bar{F}_j(t) = I(t < 0) + \exp(-\lambda_j t)I(t \geq 0)$, hence the system lifetimes for the series and parallel systems have survivor functions given by, for $t \geq 0$,

$$\bar{F}_{\text{ser}}(t) = \exp \left\{ - \left(\sum_{j=1}^K \lambda_j \right) t \right\}$$

and
$$\bar{F}_{\text{par}}(t) = 1 - \prod_{j=1}^K [1 - \exp(-\lambda_j t)] \quad (6)$$

For the series-parallel system, the survivor function under the exponentially distributed component lifetimes becomes

$$\bar{F}_{\text{sp}}(t) = \exp\{-\lambda_1 t\} \{1 - [1 - \exp(-\lambda_2 t)] \times [1 - \exp(-\lambda_3 t)]\} \quad (7)$$

Other possible parametric distributional models for the component lifetimes are the Weibull family,

the gamma family, the lognormal family, and other classes of distributions for positive-valued random variables. Given a specification of the component lifetime distributions, a system lifetime distribution is induced. For example, with exponential component lifetimes, an exponential system lifetime distribution under a series system and a nonexponential system lifetime distribution under a parallel system arise.

In the engineering, reliability, and military settings, in contrast to biomedical and public health settings, it may be reasonable to assume a parametric model for the distributions of the component lifetimes, and consequently, the system lifetime distribution. However, if the assumed parametric forms are misspecified, then erroneous inference and decisions may ensue. Robust inference procedures are therefore obtained by not imposing potentially restrictive parametric forms for the distributions, but rather simply by assuming that the F_j 's belong to the class of continuous distributions, or to some more specific nonparametric class of lifetime distributions such as the class of increasing failure rate on average (IFRA) distributions (*cf.*, Barlow and Proschan [1] for discussions on such classes of lifetime distributions) (*see Imprecise Reliability*). In the next section, we examine methods of inference under this nonparametric setting.

Nonparametric Inference

Let us consider the situation in which n identical systems are followed over some observation window. We suppose that the i th system is monitored until time τ_i , where the τ_i could be an administrative time, or it could be time to failure due to other causes not related to component failures (e.g., an accident). Denote the system lifetimes by $S_1, S_2, \dots, S_n$. However, because monitoring of the i th system is terminated at time τ_i , if the system is still functioning at time τ_i , then we will only know that the system lifetime exceeded τ_i . The observables for the n systems will therefore be

$$(\mathbf{Z}, \delta) = ((Z_1, \delta_1), (Z_2, \delta_2), \dots, (Z_n, \delta_n)) \quad (8)$$

where $Z_i = \min(S_i, \tau_i)$ and $\delta_i = I\{S_i \leq \tau_i\}$. The observable vector $(\mathbf{Z}, \delta)$ is referred to as the right-censored data for the n systems (*see Competing Risks*). It is of interest to make inferences about the system lifetime survivor function $\bar{F}_\phi(\cdot)$ based on $(\mathbf{Z}, \delta)$.

A nonparametric estimator of $\bar{F}_\phi$ is the product-limit estimator (PLE), also called the *Kaplan–Meier estimator* (see **Detection Limits**) in honor of its developers (*cf.*, Kaplan and Meier [2]). To describe this estimator in a compact form, we introduce some stochastic processes. Define the processes $N = \{N(s) : s \geq 0\}$ and $Y = \{Y(s) : s \geq 0\}$ according to

$$N(s) = \sum_{i=1}^n I\{Z_i \leq s, \delta_i = 1\}$$

and $Y(s) = \sum_{i=1}^n I\{Z_i \geq s\}$ (9)

The PLE of $\bar{F}_\phi$ is given by

$$\hat{\bar{F}}_\phi(t) = \prod_{s=0}^t \left[1 - \frac{dN(s)}{Y(s)} \right] \quad (10)$$

In the event that there are no right-censored observations, so $\delta_i = 1, i = 1, 2, \dots, n$, the PLE reduces to

$$\hat{\bar{F}}_\phi(t) = \frac{1}{n} \sum_{i=1}^n I\{S_i > t\} \quad (11)$$

the empirical survivor function. If the τ_i 's have a common distribution function $G(t) = \Pr\{\tau_1 \leq t\}$, it is well known from Efron [3] and Breslow and Crowley [4] that when the τ_i 's are independent of the T_{ij} s and n is large, $\hat{\bar{F}}_\phi(t)$ is approximately normally distributed with mean $\bar{F}_\phi(t)$ and asymptotic variance given by

$$\text{Var}\{\hat{\bar{F}}_\phi(t)\} = \frac{1}{n} [\bar{F}_\phi(t)]^2 \int_0^t \frac{-d\bar{F}_\phi(u)}{\bar{F}_\phi(u)^2 \bar{G}(u-)} \quad (12)$$

The PLE presented above is entirely based on the right-censored system lifetimes. A more efficient estimator of $\bar{F}_\phi(t)$ may be obtained provided that the failure times of the components that failed can be determined, and by exploiting the relationship between the system's reliability function and the system lifetime in equation (5). The idea is to obtain individual estimates of the $\bar{F}_j$, $j = 1, 2, \dots, K$, based on the j th component's right-censored data, and to plug these estimates in the system's reliability function to obtain an estimate of the system survivor function. This approach was implemented in Doss *et al.* [5]. We describe this more efficient estimator below.

For the i th system, denote by T_{ij} the lifetime of component j . Define $Z_{ij} = \min\{T_{ij}, S_i, \tau_i\}$ and $\delta_{ij} = I\{T_{ij} \leq \min(S_i^*, \tau_i)\}$. The right-censoring variable

for T_{ij} involves S_{ij}^* , which is the system lifetime of the original system but with the j th component perpetually functioning, so that the distribution of the resulting censoring variable is dependent on the structure function ϕ . More specifically, S_{ij}^* has survivor function given by

$$\bar{F}_j^*(t) = \Pr\{S_{ij}^* > t\} = h_\phi(\bar{F}_1(t), \dots, \bar{F}_{j-1}(t), 1, \dots, \bar{F}_{j+1}(t), \dots, \bar{F}_K(t)) \quad (13)$$

Next, for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, K$, define the stochastic processes

$$N_{ij}(s) = I\{Z_{ij} \leq s; \delta_{ij} = 1\}$$

$$\text{and } Y_{ij}(s) = I\{Z_{ij} \geq s\} \quad (14)$$

Aggregate these processes according to $N_j(s) = \sum_{i=1}^n N_{ij}(s)$ and $Y_j(s) = \sum_{i=1}^n Y_{ij}(s)$. The PLE of $\bar{F}_j(t)$ is then given by

$$\hat{\bar{F}}_j(t) = \prod_{s=0}^t \left[1 - \frac{dN_j(s)}{Y_j(s)} \right], \quad j = 1, 2, \dots, K \quad (15)$$

where $Z_{(n)j} = \max_{1 \leq i \leq n} Z_{ij}$. From PLE theory, if τ_i has survivor function $\bar{G}(t)$, then, for large n , $(\hat{\bar{F}}_1(t), \dots, \hat{\bar{F}}_K(t))$ will be approximately multivariate normal with mean $(\bar{F}_1(t), \dots, \bar{F}_K(t))$ and asymptotic covariance matrix of $\frac{1}{n} \Sigma_\phi(t)$ with

$$\Sigma_\phi(t) = \text{Dg} \left(\bar{F}_j(t)^2 \int_0^t \frac{-d\bar{F}_j(u)}{\bar{F}_j(u)^2 \bar{F}_j^*(u-)} \bar{G}(u-), \right. \\ \left. j = 1, 2, \dots, K \right) \quad (16)$$

where $\text{Dg}(\mathbf{a})$ denotes the diagonal matrix induced by $\mathbf{a}$.

From the component lifetime survivor function estimators in equation (15) and by equation (5), we form the component-based PLE of $\bar{F}_\phi(t)$ via

$$\tilde{\bar{F}}_\phi(t) = h_\phi(\hat{\bar{F}}_1(t), \hat{\bar{F}}_2(t), \dots, \hat{\bar{F}}_K(t)) \\ \times I\left\{t \leq \max_{i,j} Z_{ij}\right\} \quad (17)$$

Let

$$\mathbf{A}_\phi(p_1, \dots, p_K) = \left(\frac{\partial h_\phi(\mathbf{p})}{\partial p_j}, j = 1, 2, \dots, K \right) \quad (18)$$

which is the vector consisting of the reliability importance of the components. By the δ method, for

large n , it follows that $\tilde{\tilde{F}}_\phi(t)$ will be approximately normal with mean $\bar{F}_\phi(t)$ and asymptotic variance of

$$\text{Var}\left\{\tilde{\tilde{F}}_\phi(t)\right\} = \frac{1}{n} \mathbf{A}_\phi(\bar{F}_1(t), \dots, \bar{F}_K(t)) \Sigma_\phi(t) \mathbf{A}_\phi(\bar{F}_1(t), \dots, \bar{F}_K(t))' \quad (19)$$

As shown in Doss *et al.* [5], the estimator in equation (17) of the system lifetime survivor function is more efficient, in the sense of smaller variance, than the ordinary PLE $\hat{\tilde{F}}_\phi(t)$ given in equation (10). On the other hand, it should be recognized that $\tilde{\tilde{F}}_\phi(t)$ is a more complicated estimator than $\hat{\tilde{F}}_\phi(t)$. Furthermore, to construct $\tilde{\tilde{F}}_\phi(t)$, the exact failure times of the components that failed need to be known, which may not be feasible in practice, whereas the estimator $\hat{\tilde{F}}_\phi(t)$ only requires the right-censored system lifetimes.

Having obtained estimators of the system life survivor function in the form of $\hat{\tilde{F}}_\phi$ or $\tilde{\tilde{F}}_\phi$, one could use them to obtain estimators of other relevant parameters of the system lifetime, such as the mean, standard deviation, median, or quantiles. For instance,

possible estimators of the mean system lifetime are

$$\begin{aligned} \hat{\mu}_\phi &= - \int_0^\infty t d\hat{\tilde{F}}_\phi(t) \\ \text{and } \tilde{\mu}_\phi &= - \int_0^\infty t d\tilde{\tilde{F}}_\phi(t) \end{aligned} \quad (20)$$

Observe that since both $\hat{\tilde{F}}_\phi$ and $\tilde{\tilde{F}}_\phi$ are step functions the above integrals are just finite sums.

We illustrate these concepts with an example, using the three-component series-parallel system. For each of the three components we generate 20 random variates from the exponential distribution with mean 1. System survival times and the failure indicator random variables, $(\mathbf{Z}, \delta)$, are generated using this information and the structure function $\phi_{sp}(x_1, x_2, x_3) = \min\{x_1, \max\{x_2, x_3\}\}$. We also allow for random censoring *via* an exponential distribution with mean 1. Table 1 shows the raw data for the system times and the hand calculations for the PLE of $\tilde{F}_{sp}$, denoted by $\hat{\tilde{F}}_{sp}$. We repeat these calculations for each of the three components in the series-parallel system by using the ordered pairs $(\mathbf{Z}_j, \delta_j)$, $j = 1, 2, 3$. We show the raw data for the component times and the PLEs

Table 1 The system lifetimes and failure indicator are shown in the first two columns. Jumps in the $N(\cdot)$ counting process and changes in the at-risk counter, $Y(\cdot)$, are also given. Calculations for the PLE based on system lifetimes, $\hat{\tilde{F}}_{sp}$, are displayed

$Z_{(i)}$	$\delta_{(i)}$	$\Delta N(Z_{(i)})$	$Y(Z_{(i)})$	$\frac{\Delta N(Z_{(i)})}{Y(Z_{(i)})}$	$1 - \frac{\Delta N(Z_{(i)})}{Y(Z_{(i)})}$	$\hat{\tilde{F}}_{sp}(Z_{(i)})$
0.0054	0	0	20	0	1	1
0.0056	0	0	19	0	1	1
0.0468	0	0	18	0	1	1
0.0583	1	1	17	0.0588	0.9412	0.9412
0.0621	1	1	16	0.0625	0.9375	0.8824
0.0697	0	0	15	0	1	0.8824
0.0776	0	0	14	0	1	0.8824
0.0782	1	1	13	0.0769	0.9231	0.8145
0.1366	0	0	12	0	1	0.8145
0.1368	1	1	11	0.0909	0.9091	0.7405
0.2813	0	0	10	0	1	0.7405
0.2911	1	1	9	0.1111	0.8889	0.6582
0.5322	0	0	8	0	1	0.6582
0.5449	1	1	7	0.1429	0.8571	0.5641
0.5532	1	1	6	0.1667	0.8333	0.4701
0.5965	0	0	5	0	1	0.4701
0.6592	1	1	4	0.2500	0.7500	0.3526
0.6750	1	1	3	0.3333	0.6667	0.2351
0.6766	1	1	2	0.5000	0.5000	0.1176
0.7109	1	1	1	1.0000	0.0000	0.0000

Table 2 The first column displays the component failure times, and the censored system times as identified by the three following columns. PLEs for the survivor function for each of the components are shown. Finally, the component-based PLE for the series–parallel system is given in the last column

$Z_{(i)j}$	$\delta_{(i)1}$	$\delta_{(i)2}$	$\delta_{(i)3}$	$\hat{\bar{F}}_1(Z_{(i)j})$	$\hat{\bar{F}}_2(Z_{(i)j})$	$\hat{\bar{F}}_3(Z_{(i)j})$	$\hat{\bar{F}}_{sp}(Z_{(i)j})$
0.0054	0	0	0	1	1	1	1
0.0056	0	0	0	1	1	1	1
0.0088	0	1	0	1	0.9444	1	1
0.0468	0	0	0	1	0.9444	1	1
0.0583	1	0	0	0.9412	0.9444	1	0.9412
0.0621	1	0	0	0.8824	0.9444	1	0.8824
0.0697	0	0	0	0.8824	0.9444	1	0.8824
0.0736	0	0	1	0.8824	0.9444	0.9286	0.8789
0.0776	0	0	0	0.8824	0.9444	0.9286	0.8789
0.0780	0	0	1	0.8824	0.9444	0.8572	0.8754
0.0782	1	0	0	0.8145	0.9444	0.8572	0.8080
0.1262	0	0	1	0.8145	0.9444	0.7793	0.8045
0.1366	0	0	0	0.8145	0.9444	0.7793	0.8045
0.1368	0	1	0	0.8145	0.8500	0.7793	0.7875
0.1409	0	0	1	0.8145	0.8500	0.6927	0.7770
0.2549	0	0	1	0.8145	0.8500	0.6061	0.7664
0.2813	0	0	0	0.8145	0.8500	0.6061	0.7664
0.2911	0	1	0	0.8145	0.7438	0.6061	0.7323
0.3226	0	0	1	0.8145	0.7438	0.5051	0.7112
0.3924	0	1	0	0.8145	0.6375	0.5051	0.6684
0.5322	0	0	0	0.8145	0.6375	0.5051	0.6684
0.5449	0	1	0	0.8145	0.5100	0.5051	0.6170
0.5532	1	0	0	0.6787	0.5100	0.5051	0.5141
0.5965	0	0	0	0.6787	0.5100	0.5051	0.5141
0.6592	0	1	0	0.6787	0.2550	0.5051	0.4285
0.6750	1	0	0	0.4525	0.2550	0.5051	0.2857
0.6766	1	0	0	0.2263	0.2550	0.5051	0.1429
0.7109	1	0	0	0.0000	0.2550	0.5051	0.0000

for $\bar{F}_j$ for each of the components in Table 2. We then obtain the component-based PLE of $\bar{F}_{sp}$, denoted by $\hat{\bar{F}}_{sp}$, by evaluating the reliability function, $h_{sp}(p_1, p_2, p_3) = p_1[1 - (1 - p_2)(1 - p_3)]$ at $p_j = \hat{\bar{F}}_j(t)$, for $j = 1, 2, 3$ and $t \leq \max_{i,j} Z_{ij}$. Otherwise, we define $\hat{\bar{F}}(t) = 0$ when $t > \max_{i,j} Z_{ij}$. Note that statistical software, such as R, have built-in functions for calculating the PLE of $\bar{F}_{sp}$ and $\bar{F}_j$, for $j = 1, 2, 3$. The component-based PLE of $\bar{F}_{sp}$ is also shown in Table 2.

When the failure times for each of the components in the series–parallel system follows an exponential distribution with mean 1, the true system survivor function is given by

$$\bar{F}_{sp}(t) = \exp(-t)\{1 - [1 - \exp(-t)]^2\} \quad (21)$$

The PLE of $\bar{F}_{sp}$, the component-based PLE of $\bar{F}_{sp}$, and the true survivor function are shown in Figure 1.

The component-based PLE mimics the behavior of the system-based PLE closely. Both are comparable to the true survivor function.

Semiparametric Models

The models and inference methods considered in the preceding sections are restrictive, since they do not take into account the impact of covariates such as environmental conditions, operating characteristics, demographic variables, etc. This is especially crucial in the assessment of the risk or quality of systems, since the incorporation of relevant covariates may lead to more precise and accurate knowledge. In this section, we discuss models that incorporate covariates, and provide some inference methods for these models.

Consider a system whose performance is affected by a vector of covariates, say, a $q \times 1$ vector

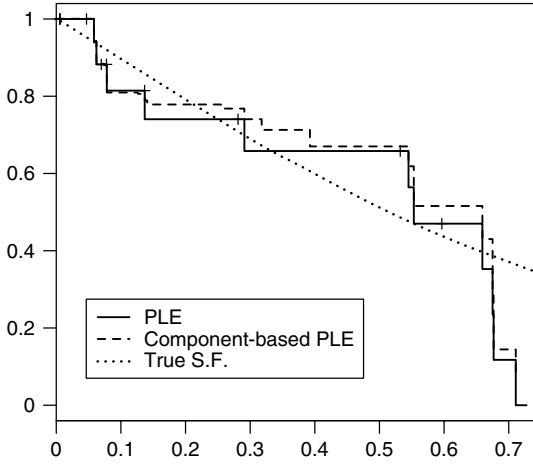


Figure 1 The PLE of $\bar{F}_{sp}$ denoted by $\hat{\bar{F}}_{sp}$ (solid line) and the component-based PLE of $\bar{F}_{sp}$ denoted by $\tilde{\bar{F}}_{sp}$ (dashed line) are shown using 20 randomly generated component and system failure times for the series-parallel system. The true survivor function is also shown (dotted line)

$\mathbf{V} = (V_1, V_2, \dots, V_q)'$. Such a covariate vector could be time-dependent; however, for simplicity we restrict ourselves to time-independent covariates. The simplest type of model directly relates the system lifetime S to the covariate vector $\mathbf{V}$. The two major models for doing so are accelerated failure time (AFT) models (also referred to as *log-linear models*) and hazard-based models.

In the AFT model, one postulates that

$$\log(S) = \beta\mathbf{V} + \sigma\epsilon \quad (22)$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_q)$ is a $1 \times q$ vector of regression coefficients, σ is a dispersion parameter, and ϵ is an error component whose distribution is typically assumed to belong to some parametric family of distributions. The appealing trait of these AFT models is the easy interpretability of the regression coefficients. Technically, however, this AFT model is still a parametric model, in contrast to the next class of models, which are semiparametric.

To introduce hazard-based models, we recall relevant functions. For a continuous positive random variable T , with density function $f(t)$ and survivor function $\bar{F}(t)$, the hazard rate function and the hazard function are defined, respectively, via

$$\lambda(t) = \frac{f(t)}{\bar{F}(t)} \quad \text{and} \quad \Lambda(t) = \int_0^t \lambda(u) du \quad (23)$$

These functions are related to f and $\bar{F}$ according to $f(t) = \lambda(t) \exp\{-\Lambda(t)\}$ and $\bar{F}(t) = \exp\{-\Lambda(t)\}$. The intuitive interpretation of $\lambda(t)(\Delta t)$, where Δt is a small positive number, is that this is the approximate conditional probability that $T \leq t + \Delta t$, given that $T \geq t$, that is, the probability of failing in a small interval starting from t given that the unit is alive just after t .

The Cox proportional hazards (PH) model (cf., Cox [6]) (see **Lifetime Models and Risk Assessment; Hazard and Hazard Ratio**) relates the system lifetime to the covariate vector $\mathbf{V}$ according to

$$\lambda_{s|\mathbf{V}}(s|\mathbf{v}) = \lambda_0(s) \exp\{\beta\mathbf{v}\} \quad (24)$$

where $\lambda_0(\cdot)$ is some baseline hazard rate function which is nonparametrically specified. This is a semiparametric model, since there is an infinite-dimensional parameter $\lambda_0(\cdot)$ and a finite-dimensional parameter β . It is well known (cf., Kalbfleisch and Prentice [7]) that the AFT model and the Cox PH model coincide if and only if the baseline hazard rate function $\lambda_0(t)$ belongs to the Weibull class, that is, $\lambda_0(t) = (\alpha\gamma)(\gamma t)^{\alpha-1}$. In this case, the distribution of the error term in the AFT model is the extreme-value distribution. Since the focus of this article is on nonparametric and semiparametric models, we henceforth concentrate on the Cox PH-type model. It should be noted that if $\bar{F}_0(t)$ is the survivor function associated with the baseline hazard rate function $\lambda_0(\cdot)$, then the conditional survivor function of the system lifetime under the Cox PH model, given $\mathbf{V} = \mathbf{v}$, is

$$\bar{F}_\phi(t|\mathbf{v}) = [\bar{F}_0(t)]^{\exp(\beta\mathbf{v})} \quad (25)$$

Perhaps, it may be difficult to justify the proportionality postulate for the system lifetime hazard function in the Cox model, since the system lifetime is determined by the component lifetimes and the structure function of the system. Instead, it may be easier to provide justification for such a proportionality assumption when dealing with the lifetimes at the component level. We could therefore have an alternative model wherein the Cox PH model is imposed at the component level. This will then induce a model for the system lifetime. However, the resulting system lifetime model need not anymore be of the Cox PH-type. This was the approach explored in Hollander and Peña [8]. To elaborate on this alternative model, for the j th ($j = 1, 2, \dots, K$) component lifetime T_j ,

we postulate that the conditional hazard rate function of T_j , given $\mathbf{V} = \mathbf{v}$, is

$$\lambda_j(t|\mathbf{v}) = \lambda_{0j}(t) \exp\{\beta_j \mathbf{v}\} \quad (26)$$

where $\lambda_{0j}(\cdot)$'s are baseline hazard rate functions, and β_j 's are $1 \times q$ regression coefficient vectors. This will entail that the conditional survivor function for the j th component is

$$\bar{F}_j(t|\mathbf{v}) = [\bar{F}_{0j}(t)]^{\exp\{\beta_j \mathbf{v}\}} \quad (27)$$

where $\bar{F}_{0j}$ is the survivor function associated with λ_{0j} . Consequently, the conditional survivor function for the system lifetime is given by

$$\bar{F}_\phi(t|\mathbf{v}) = h_\phi(\bar{F}_1(t|\mathbf{v}), \bar{F}_2(t|\mathbf{v}), \dots, \bar{F}_K(t|\mathbf{v})) \quad (28)$$

As a simple example, if the system is series, then the conditional system survivor function is $\bar{F}_{\text{ser}}(t|\mathbf{v}) = \prod_{j=1}^K [\bar{F}_{0j}(t)]^{\exp\{\beta_j \mathbf{v}\}}$, whose associated conditional hazard rate function is

$$\lambda_{\text{ser}}(t|\mathbf{v}) = \sum_{j=1}^K \lambda_{0j}(t) \exp\{\beta_j \mathbf{v}\} \quad (29)$$

which is not anymore of the Cox PH-type model, but rather is an additive Cox PH-type model. If one could assume, however, that $\beta_1 = \beta_2 = \dots = \beta_K = \beta$, then equation (29) reduces to a Cox PH-type model with a baseline hazard rate function of $\lambda_0(t) = \sum_{j=1}^K \lambda_{0j}(t)$.

Semiparametric Inference

Consider the situation in which n systems with the same structure function ϕ are monitored, and with the i th system associated with a covariate vector $\mathbf{v}_i$. As in an earlier section, the goal is to infer about the system lifetime survivor function and also the dependence of this survivor function on the covariates. We first consider the Cox-type model given in equation (24). We assume that system j is monitored over the period $[0, \tau_i]$, and so the observables are $\{(Z_i, \delta_i, \mathbf{v}_i), i = 1, 2, \dots, n\}$, where $Z_i = \min(S_i, \tau_i)$ and $\delta_i = I\{S_i \leq \tau_i\}$. On the basis of this data, the goal is to estimate both the baseline survivor function $\bar{F}_0(t)$ and the regression coefficient vector β , which will then lead to an estimate of the conditional system lifetime survivor function. Estimators of these parameters were obtained in Cox [6] and the properties of these

estimators were rigorously developed in Andersen and Gill [9]. We describe these estimators and some of their properties.

As developed by Cox in his seminal paper [6], when $\lambda_0(\cdot)$ is nonparametrically specified, the estimator of β is the maximizer of the so-called partial likelihood function. To describe this partial likelihood function in modern stochastic process notation, for $i = 1, 2, \dots, n$, define the processes

$$N_i(t) = I\{Z_i \leq t; \delta_i = 1\}$$

$$\text{and } Y_i(t) = I\{Z_i \geq t\} \quad (30)$$

The partial likelihood function is given by

$$L_P(\beta) = \prod_{i=1}^n \prod_{t=0}^{\infty} \left[\frac{\exp\{\beta \mathbf{v}_i\}}{\sum_{l=1}^n Y_l(t) \exp\{\beta \mathbf{v}_l\}} \right]^{dN_i(t)} \quad (31)$$

The $\hat{\beta}$ which maximizes this function is called the *partial likelihood maximum-likelihood estimator* (PLMLE) of β , and it could be obtained using iterative numerical methods. See, for instance, Fleming and Harrington [10], Andersen *et al.* [11], and Therneau and Grambsch [12] for procedures for computing $\hat{\beta}$. When n is large and under regularity conditions, it has been established in Andersen and Gill [9] that $\hat{\beta}$ is approximately normally distributed with mean β and a covariance matrix, which is approximated by

$$\widehat{\text{Cov}}(\hat{\beta}) = \left(-\frac{\partial^2 \log L_P(\beta)}{\partial \beta \partial \beta'} \Big|_{\beta=\hat{\beta}} \right)^{-1} \quad (32)$$

These results could be used to determine the importance of covariates by performing hypothesis tests and constructing confidence intervals for components of β .

To obtain an estimator of the baseline survivor function, we first get an estimator of the baseline hazard function Λ_0 . The estimator for this function, called the *Aalen–Breslow–Nelson estimator*, is given by

$$\hat{\Lambda}_0(t) = \sum_{i=1}^n \int_0^t \frac{dN_i(u)}{\sum_{l=1}^n Y_l(u) \exp\{\hat{\beta} \mathbf{v}_l\}}, \quad t \geq 0 \quad (33)$$

By virtue of the product–integral representation of a survivor function in terms of its hazard function

given by

$$\bar{F}_0(t) = \prod_{u=0}^t [1 - d\Lambda_0(u)] \quad (34)$$

a substitution-type estimator of $\bar{F}_0$ is obtained via

$$\hat{\bar{F}}_0(t) = \prod_{u=0}^t \left[1 - \frac{dN_i(u)}{\sum_{l=1}^n Y_l(u) \exp\{\hat{\beta}\mathbf{v}_l\}} \right] \quad (35)$$

Consistency properties and the joint asymptotic normality of $(\hat{\beta}, \hat{\Lambda}_0(\cdot))$, and consequently of $(\hat{\beta}, \hat{\bar{F}}_0(\cdot))$, were rigorously established in Andersen and Gill [9] using martingale methods.

Finally, an estimator of the conditional system survivor function is obtained through substitution to yield

$$\hat{\bar{F}}_\phi(t|\mathbf{v}) = \left[\hat{\bar{F}}_0(t) \right]^{\exp\{\hat{\beta}\mathbf{v}\}} \quad (36)$$

The importance of having such an estimator of $\bar{F}_\phi(t|\mathbf{v})$ is that one will be able to assess the probability of the system completing a task within a specified period of time for a given set of values of the covariates. This eventually could provide a more informed knowledge about the risk associated with employing the system and the level of quality of the system.

The estimator of the conditional system lifetime survivor function utilizes only the possibly right-censored system lifetimes. This is bound to be less efficient if the failure times of components that have failed can be determined, as in the case when there are no covariates. Analogously to the estimator $\hat{\bar{F}}_\phi(t)$ in equation (17) for the no-covariate case, we may exploit the relationship among the component conditional survivor functions $\bar{F}_j(t|\mathbf{v})$, $j = 1, 2, \dots, K$, and $\bar{F}_\phi(t|\mathbf{v})$ through the system reliability function as provided in equation (28). In particular, we suppose that the component failure times are related to the covariate vector according to a Cox PH model as specified in equation (26). By employing the analogous estimation procedures for $(\beta, \bar{F}_0)$ as described above, we can obtain the PLMLE and Aalen–Breslow–Nelson estimators, for each $j = 1, 2, \dots, K$, of $(\beta_j, \bar{F}_j(t|\mathbf{v}))$, denoted

by $(\hat{\beta}_j, \hat{\bar{F}}_j(t|\mathbf{v}))$. To obtain estimates for the j th component parameters, we use the observable data given by $\{(Z_{ij}, \delta_{ij}, \mathbf{v}_i), i = 1, 2, \dots, n\}$, where $Z_{ij} = \min(T_{ij}, S_i, \tau_i)$ and $\delta_{ij} = I\{T_{ij} \leq \min(S_{ij}^*, \tau_i)\}$, with S_{ij}^* as defined before. The estimator of the system conditional survivor function becomes

$$\begin{aligned} \hat{\bar{F}}_\phi(t|\mathbf{v}) &= h_\phi \left(\hat{\bar{F}}_1(t|\mathbf{v}), \hat{\bar{F}}_2(t|\mathbf{v}), \dots, \hat{\bar{F}}_K(t|\mathbf{v}) \right) \\ &\times I \left\{ t \leq \max_{i,j} Z_{ij} \right\} \end{aligned} \quad (37)$$

If the component failure times can be ascertained and the component Cox PH models hold, then this estimator is bound to be more efficient than the estimator $\hat{\bar{F}}_\phi(t|\mathbf{v})$. However, more research regarding the properties of the estimator in equation (37) is still warranted. Also, we point out that an efficiency comparison of the two estimators may not be entirely appropriate, since the Cox PH model for the system lifetime need not arise from the component Cox PH models which was the basis of the estimator in equation (37).

Dynamic Models

The probabilistic models considered in earlier sections are static in nature, in the sense that there is no proviso for the impact of the failed components on still functioning components. For instance, when components fail in a nonseries system, there might incur more load or stresses on the remaining functioning components to the extent that their residual failure rates may increase. Models that incorporate the evolving nature of coherent systems are referred to as dynamic models. Because of space limitations, we do not dwell on these models in this article, but simply restrict ourselves to alerting the reader to some recent pertinent work dealing with dynamic models. For instance, Hollander and Peña [13] discussed the need for such dynamic models when dealing with coherent structures and demonstrated the ensuing theoretical difficulties and complications arising from dynamic models. A general class of dynamic models for recurrent events, which is applicable to the modeling of coherent systems, was proposed in Peña and Hollander [14]. Some inference procedures for this class of models were described in the recent papers of Kvam and Peña [15], Peña *et al.* [16], and Stocker and Peña

[17]. Nevertheless, further research is called for to ascertain properties of dynamically modeled systems and to develop appropriate statistical inference procedures for such models.

Acknowledgment

Peña and Taylor acknowledge the research support provided by US NIH Grant GM56182 and Peña also acknowledges the research support of US NIH COBRE Grant RR17696.

References

- [1] Barlow, R. & Proschan, F. (1981). *Statistical Theory of Reliability and Life Testing: Probability Models*, Silver Spring, MD.
- [2] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.
- [3] Efron, B. (1967). The two-sample problem with censored data, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Prentice-Hall, New York, pp. 831–853.
- [4] Breslow, N. & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship, *Annals of Statistics* **2**, 437–453.
- [5] Doss, H., Freitag, S. & Proschan, F. (1989). Estimating jointly system and component reliabilities using a mutual censorship approach, *Annals of Statistics* **17**, 764–782.
- [6] Cox, D. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society* **34**, 187–220.
- [7] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, Wiley Series in Probability and Statistics, 2nd Edition, Wiley-Interscience [John Wiley & Sons], Hoboken.
- [8] Hollander, M. & Peña, E.A. (1996). Reliability models and inference for systems operating in different environments, *Naval Research Logistics* **43**, 1079–1108.
- [9] Andersen, P. & Gill, R. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics* **10**, 1100–1120.
- [10] Fleming, T. & Harrington, D. (1991). *Counting Processes and Survival Analysis*, John Wiley & Sons, New York.
- [11] Andersen, P., Borgan, O., Gill, R. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.
- [12] Therneau, T. & Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer, New York.
- [13] Hollander, M. & Peña, E.A. (1995). Dynamic reliability models with conditional proportional hazards, *Lifetime Data Analysis* **1**, 377–401.
- [14] Peña, E. & Hollander, M. (2004). Models for recurrent events in reliability and survival analysis, in *Mathematical Reliability: An Expository Perspective*, R. Soyer, T. Mazzuchi & N. Singpurwalla, eds, Kluwer Academic Publishers, Chapter 6, pp. 105–123.
- [15] Kvam, P. & Peña, E. (2005). Estimating load-sharing properties in a dynamic reliability system, *Journal of the American Statistical Association* **100**, 262–272.
- [16] Peña, E., Slate, E. & Gonzalez, J.R. (2007). Semiparametric inference for a general class of models for recurrent events, *Journal of Statistical Planning and Inference* **137**, 1727–1747.
- [17] Stocker, R. & Peña, E. (2007). A general class of parametric models for recurrent event data, *Technometrics* **49**, 210–220.

EDSEL A. PEÑA AND LAURA L. TAYLOR

Nonlife Insurance Markets

The Nature of Nonlife Insurance

Consumers and businesses prefer to avoid risk and uncertainty that have significant impact on their future income, wealth, or business prospects. They are in general risk averse, meaning they are willing to pay to avoid their individual risks. Risk pooling and the statistical law of large numbers make this possible and are the fundamentals of the insurance industry. By pooling a large number of risks where the causes of loss are significantly independent of each other, the insurance companies are able to predict the claims level to a reasonable degree of certainty. Pooling risks may be seen as portfolio diversification, where unsystematic random risks are reduced or eliminated by risk diversification and only systematic portfolio risks remain as the main risks to cover. Hence, insurance companies are better able to manage and absorb risks than individual consumers and businesses. In effect, consumers and businesses are transferring an uncertain payment in exchange for a much more certain payment, the insurance premium. Insurance exhibits economies of scale and it allows the economy as a whole to be more enterprising and to undertake longer term capital expenditure decisions which otherwise might not be made, or might be delayed. A well functioning nonlife insurance industry is one of the efficient fundamentals of local as well as global economies.

Market Characteristics

Regulations and Legislations

Market regulations and legislations controlled by federal legal authorities are crucial market constraints in nonlife insurance. The aim of the regulations is to protect both consumers and suppliers. Most commonly the directives relate to capital solvency, compulsory insurances, protection of consumers, and competition regulations, including pricing. In general we may differ between regulated and deregulated markets. The global trend is to deregulate to increase the market competition and hence stimulate better

and more efficient insurance solutions and customer service. The deregulation has particularly led to freedom in insurance pricing as a lever to increase competition. While premium tariffs have standard federal structures in highly regulated markets, price cooperation is not permitted between suppliers in fully price-deregulated markets.

Unlike the deregulation of competition and pricing, the global trend for consumer protection and insurance solvency is rather to strengthen the impact of these directives. The idea is to ensure efficient competition within society friendly and financially reliable market conditions. The solvency regulation in a nonlife insurance market is particularly important and affects the capital requirements of the insurance companies. Actuarial claims loss reserving is e.g., an important part of the solvency (*see Solvency*) regulatory. Strict solvency rules ensure the companies to have enough money to pay their future insurance liabilities and debts, and therefore protect both customers and shareholders.

The deregulation of nonlife insurance markets over the last two decades has increased mergers and acquisitions and thus the consolidation in most markets. The more cross-border regulatory harmonization takes place, the more cross-border consolidation and competition has appeared. Solvency objective, cost and revenue efficiency, and scale economy are major reasons behind mergers and acquisitions. A study by Cummins and Xie [1] concludes that acquiring firms in the US property-liability (P-L) industry achieve more revenue efficiency growth than nonacquirers, and that target firms experience greater cost and allocative efficiency than nontargets. Their main conclusions are that operating synergy exists between acquirers and target firms in the US P-L industry, and that the restructuring of the industry in the 1990s and early 2000s has produced efficiency gains.

Profitability and Underwriting Cycles

The profitability of the nonlife insurance industry has – like other industries – varied over time. However, the profitability has empirically tended to swing between profitable and unprofitable time periods alternating over time cycles. This observation is commonly known as the *underwriting* or *insurance* cycle and has traditionally been an important feature of nonlife insurance markets. Three explanations of the underwriting cycle have some degree of empirical

support; see [2]. These are cash flow underwriting, cost of equity and claims shock theory. *Cash flow underwriting* refers to a negative correlation between return on investments and insurance premium rates. *Cost of equity* refers to a downward pressure on insurance prices when stock markets rise above normal levels and hence generates an increase in insurance companies' available capital and a fall in their cost of capital. The *claims shock theory* states that insurance pricing tends to underestimate the potential for future claims during periods when there are no large individual claims or clusters of claims, while insurance prices rise sharply when such claims occur. This explanation is based on traditional economic shock model theories and assumes that the insurance market has a short memory.

Market Rivalry and Entry Dynamics

A market is a place where supply meets demand. In general, nonlife insurance markets are incomplete or inefficient, that is, there exist no markets with free entry and exit, nontransaction costs, perfect information, and pure rational customers. In effect, price is *not* the only competitive lever in incomplete markets when customers choose their insurers. Thus the insurers should take this incompleteness into account in their pricing models; see section titled "Pricing and Underwriting". Incomplete insurance market models have been widely studied within the theory of insurance economics. Competitive insurance market equilibria, optimal risk sharing and (Pareto) optimal insurance contracts, given incomplete market assumptions, are holistic quantitative model-based research areas which give valuable insights into the nature of insurance markets; see [3] for further reading.

The market incompleteness varies between different nonlife markets and affects the market competition to a large extent. Mature insurance markets with homogenous products are, in general, less incomplete than growing markets with less homogenous products. In growing markets the market players are concerned with expanding their shares of the market growth. Underlying reasons for growth may be general economic growth in the market and increased wealth, and therefore a need for consumers or businesses to reduce their individual risks to keep up their investments. Another reason may be new insurance needs generated by specific environmental changes in

the economy related to e.g., new legislation laws, new consumer articles or expanding business. According to [4], the players in any market have to position themselves in relation to five competitive forces: the rivalry among the existing competitors, the entry of new competitors, the bargaining power of buyers, the bargaining power of suppliers, and the threat of substitutes. The first three forces are in general the most important forces in the nonlife insurance markets and influence the profitability significantly. In mature insurance markets the competitive rivalry between the existing or incumbent insurers is more intensely related to capturing customers from each other compared to growing markets where the insurers compete more in attracting new customers into the market. Rivalry in price is common, but insurers also compete in product and service quality and variety, distribution and sales efficiency, brand identity, and other variables that are important for customers. It is crucial for insurers to optimize their business setup and insurance offers in relation to these competitive levers (*see Asset-Liability Management for Non-life Insurers*).

Barriers to insurance suppliers' entry into the market and barriers preventing customers from switching between suppliers are important competitive drivers and may arise from various sources. First, regulations that impose costs or conditions on entrants that are not imposed on existing suppliers in the market are a typical entry barrier. Regulation practice that delays supplier entrance may also diminish the competitive treat in the market. Second, switching barriers between the suppliers is particularly important for market competition. In nonlife insurance business there are most often 1-year contracts, even if the customers in general may switch from one insurer to another if the need for insurance ceases during the period, because e.g., the insured object is sold.

Nonlife insurance products are often complex and/or are purchased infrequently and hence difficult to compare. Hence, even if customers may switch their insurance company any time they want, the implication is still that customers often face high search costs of comparing insurance offers from different suppliers in the market. In general this has generated a demand in many markets for intermediaries who can help consumers and businesses to find their best insurance solutions and offers. In some markets, brokers or other intermediaries handle as much as 40% of the insurance market, where commercial

insurance lines in general have a higher broker share than the simpler and more commodity-based personal lines.

Other Market Characteristics

Moral hazard, adverse risk selection and uninsurable risks are three specific factors to take into account for insurance suppliers in a competitive market. *Moral hazard* (see **Moral Hazard**) occurs when insurance buyers are less careful in protecting their insured assets and/or cause financial loss to others because they know that an insurance company will bear the costs rather than themselves. Moral hazard may be reduced by use of deductibles and pricing incentives like no-claims discount or bonus-malus systems (see **Bonus–Malus Systems**) in motor insurance. [5] is a classical theoretical reference to moral hazard and its link to asymmetric information in insurance.

Adverse risk selection occurs when an insurance portfolio is made up of risks that are worse than average because the insurer has not set premiums according to individual risk levels and exposure. The premiums are rather set according to an average risk level in the portfolio or market that generates good value for high risks and bad value for low risks. Hence, low risks will leave the portfolio to get a better price offer at other places while high risks stay, and a mismatch between premium and risk level occurs in a negative spiral. In effect, imperfect information equilibrium has emerged between the insurer and the customers [6]; is a classical theoretical illustration of this equilibrium issue. The degree of adverse selection depends on the price sensitivity in the market and may be reduced by pricing the risks based on relevant and accurate risk information.

Uninsurable risks are risks that are impossible or at least difficult to calculate and that have some problems of adverse selection and/or moral hazard. Uninsurable risks are – by definition – unattractive to insurers, and sometimes more relevant for capital markets to acquire if the risks are large (like natural catastrophes, in some insurance markets). Insurable risks should have accidental and not intentional losses, and they should have economical feasible premiums, meaning that chance of loss must not be too high. Unemployment and a drop in property prices are examples of uninsurable risks.

Competitive Market Strategies

Diversified Strategies

Insurance suppliers have to relate closely to the market they operate in. As in all industries the strategic goal for a supplier in a nonlife insurance market is to gain value for shareholders and customers. To do so, the suppliers must select and build competitive domains with attractive characteristics. This is obviously more important in deregulated and highly competitive insurance markets than in regulated and less competitive markets. Anyhow, the general strategic challenge is to simultaneously balance long-term profitability and long-term market shares, that is, to balance loss and cost ratios and premium growth in terms of key performance indicators (see **Premium Calculation and Insurance Pricing**).

According to classical [4], long run above-average performance and sustainable competitive advantage should, in general, be generated from a differentiation strategy and/or a low cost leadership strategy. The competitive scope of both strategies could be broad range or narrow (niche) range to market segments. To outperform competitive rivals the chosen diversified strategy should better cope with the general market forces of existing competitor rivalry, new competitor entries and bargaining power of buyers. In nonlife insurance business there are three operative competitive levers, which are known as *winning core value skills* of an insurance company. These are pricing and underwriting, distribution and sales, and risk and capital management. Use of modern and up-to-date information technology is obviously also highly considerable, but only a tool to generate competitive advantages within each of the three operative levers. The following text briefly outlines the three levers from a *risk and market assessment* point of view.

Pricing and Underwriting

Insurance risk selection is a fundamental competitive driver in insurance markets, and risk pricing and underwriting the key levers to manage it. For all insurers it is important to avoid adverse risk selection by diversifying their risk pooling. Statistical information as well as experienced-based knowledge of the risk exposures and outcomes is therefore the basis on which to predict and price the risks. All suppliers in mature price-deregulated markets collect

risk information into data warehouse environments to make the information analytical and accessible. Existing suppliers with significant data information have an advantage compared with new market competitors without such information, even if new suppliers usually collect price information from existing ones to overcome this disadvantage when they enter the market. In personal insurance lines and small-to medium-sized commercial lines the state-of-the-art statistical method to price insurance risks is *generalized linear regression models*. Less data driven and more *a priori* engineering methods are more common when pricing risks in large commercial and industrial lines. The main reason behind this deviation in pricing methods is the difference in claims risk variability between the business lines. Private and small to medium commercial lines are, in general, characterized by a high number of standardized and independent types of risk exposure with typically a limited range of claim amounts and high claims frequency. Large commercial and industrial lines are, on the other hand, characterized by the opposite factors. In fact, nonlife insurance markets are very much affected by the specific combination of claims loss variability and average claims level in each market.

True market winners do not only price their insurance risks quite accurately, but also manage to identify and predict risk segments where the market in general makes profit and losses. They perform a combination of risk/cost pricing and market-based pricing, which makes it possible to optimize their insurance profitability based on tactical pricing and sales positions in the market. Unlike complete price sensitive markets – like e.g., efficient financial derivative markets – there exists no unique market price (which all market suppliers offer) for each unique individual risk in insurance markets. Thus, a differentiated market-based pricing strategy utilizes the market incompleteness where customers in addition to price preferences also have product, service and/or brand preferences. In effect, the degree of price sensitivity in market/customer segments in relation to product value proposition is important to take into account when using the pricing and underwriting lever as a strategic tool to optimize long-term insurance profitability. The insurance purchasing rate and portfolio retention rate may thus be included as stochastic estimation variables into a net present value based pricing model approach; see [7] for

more details. In summary, all market dynamic factors like growth, mobility and customer rotation are important elements that influence insurance companies significantly in their pricing and underwriting practice.

Distribution and Sales

Distribution-channel strategies in nonlife insurance markets are often conceived simply as a means to reach broad or narrow markets (based on the diversified strategic scope) and are seen as the mechanism to push products and services into the market. However, often the underlying risk selection effect varies significantly between the distribution channels. That is, the channel itself may be seen as an insurance risk selection dimension that may influence the pricing and underwriting substantially. It is, for example, well known that customers who buy their insurances through banking services typically and on average (all other factors equal) have lower claims risk than other customers because they, in average, have a better credit rate record (which is a significant, but controversial risk tariff factor in many markets). Equally, personal line homeowner and car insurances linked to real estate agencies and car dealers/manufacturers are not only examples of cost-efficient distribution strategies, but also show utilization of a positive risk selection effect in these channels.

There is a conglomeration of distribution strategies and setups in nonlife insurance business. The most common ones are independent agencies, brokers, direct service, bancassurance, retailers/white labels, real estate agencies, cars dealers/manufacturers, and other alliance partnerships. The mix of distribution models in a market depends very much on the characteristics of the specific market like growth, mobility, rivalry, and entry dynamics as earlier discussed. In addition the consumer behavior and insurance buying preferences in the market – like e.g., the penetration of telephone and internet as insurance sales and service communication tools – apparently also influence the choice of distribution strategy. New market entrants who build their business from scratch are often in a better infrastructural position to adopt new distribution opportunities and hence stress and scare existing and more traditional incumbent suppliers (see **Risk and the Media**).

Risk and Capital Management

Few businesses are as inherently risky in relation to their capital base as the nonlife insurance business. Reserve risk, premium risk, underwriting risk, operational risk, interest rate risk, equity price risk, credit risk, and investment risk are main risk categories which influence the key question: How much capital does the insurer need at the beginning of the year in order to be able to cover future financial liabilities with a specific percent as security? Enterprise risk management (ERM) has become an important centralized management approach to control the entire environmental risks that an insurance company faces. The key point is to have a simultaneous approach to all relevant financial risks at a business corporate level. By using dynamic financial analysis (DFA) techniques the insurers are able to model financial risk correlations and simulate statistical outcome effects to control the capital solvency security. The reinsurance program of a company is a key tool to appropriately adjust the security level, given all the risks that the company faces. Within the DFA approach, asset–liability management (ALM) is a way to allocate capital and return on equity demands down to each insurance line of business in the company.

Winning market performers govern ERM, DFA, and ALM within an optimized reinsurance setup. They use it as a competitive advantage to exploit hard markets faster and more fully than competitors with weaker capital cushion and worse solvency rating from rating agencies. In hard markets with struggling and negative earnings they do not need to increase their premium levels and/or downsize their expenses and businesses as much as their competitors. Risk and capital focused insurers are rather in a position to gain market shares through improved competitive price positions, improved internal cost, and service efficiency, and smart business investments.

Market Outlook and Trends

The nonlife insurance industry has been radically changed during the last decade. New insurance risks linked to climate change, air pollution, terrorism,

bird flu, nanotechnology, tinnitus, and electromagnetic fields have increased the general risk exposure of nonlife insurance business. Add to the general mixture the effects of globalization, deregulations, increased competition, changing demographics and pressures on margins, and it is quite obvious that the risk exposure will increase even more in the years to come. Technological innovation and commercialism; bancassurance and increased cross-financial services; increased cross-border mergers, acquisitions, and consolidations are also among the challenges which will continue to face the insurance industry in local as well as global markets; see [8] for further reading. In total, the general capability and management power of risk assessment will be a key lever to handle these challenges in an efficient and proper way.

References

- [1] Cummins, D.J. & Xie, X. (2006). Mergers and acquisitions in the US property-liability insurance industry: productivity and efficiency effects, Paper presented at *The HEC Montreal Conference Dynamic of Insurance Markets: Structure Conduct and Performance in the 21st Century*, Montreal, 4–5 May 2006.
- [2] Parsons, C. (2004). *Report on the Economics and Regulation of Insurance*, Cass Business School, London.
- [3] Dionne, G. (2000). *Handbook of Insurance*, Springer.
- [4] Porter, M. (1985). *Competitive Advantages: Creating and Sustaining Superior Performance*, Free Press.
- [5] Holmström, B. (1979). Moral hazard and observability, *Bell Journal of Economics* **10**, 74–91.
- [6] Rotchild, M. & Stiglitz, J. (1976). Equilibrium in competitive insurance: an essay on the economics of imperfect information, *Quarterly Journal of Economics* **90**, 629–650.
- [7] Holtan, J. (2007). Pragmatic insurance option pricing, *Scandinavian Actuarial Journal* **1**, 53–70.
- [8] Cummins, D.J. & Venard, B. (2007). *The Handbook of International Insurance: Between Global Dynamics and Local Contingencies*, Springer.

Related Articles

Nonlife Loss Reserving

Risk Classification in Nonlife Insurance

JON HOLTAN

Nonlife Loss Reserving

The amount that an insurer must pay the insured as a result of a claim is known equivalently as claim amount or loss amount. Similarly, the payments that make up this amount are known indistinctly as claim payments or loss payments. Thus loss reserving is also known as *claims reserving*. In insurance terminology, a reserve is a fund, set aside from all other company's operations, to acknowledge (in the accounting sense) the impact of uncertainty in the financial position of the firm owing to its insurance operations. For example, specific reserves may be established to attribute a financial cost to pending claims, fluctuations in technical results, unearned premiums, and the value of investments. Among these, claims reserves are created with the purpose of paying all future benefits on already existing obligations, whether the specific claims are known or not. In other words, these reserves are a provision for all claims corresponding to events that have occurred but have not yet been completely settled [1].

Given the uncertainty involved in the financial value of such payments, reserving is clearly a problem of estimation or more precisely, forecasting. Together with ratemaking, the determination of claims reserves is one of the main responsibilities of nonlife actuaries, who, by certifying the adequacy of insurer's liabilities, help market participants to assess the following:

- the financial health of an insurer and its net worth;
- the *solvency* (see **Solvency**) of the company, after other considerations, such as liquidity, market value of assets, etc. are established;
- the insurer's profits during a specific period;
- the amount of incurred claims (paid claims plus reserves), as a tool for ratemaking (see **Insurance Pricing/Nonlife**).

To judge the adequacy of reserves, actuaries must base their calculations on both statutory and internal criteria, always keeping in mind that, as in all forecasting procedures, reserving depends crucially not only on the specific method employed, but also on the assumptions regarding all variables that will have an influence on the possible outcomes of the insurance business.

Given the substantial impact on companies' values as listed above, it is not surprising that different market participants will judge the relevance of both methods and assumptions under quite different perspectives. Insurance supervisors, having a macroeconomic perspective of the insurance industry, try to avoid companies becoming unable to meet their commitments, therefore they encourage the use of methods and assumptions that could produce relatively high values, to guarantee that a substantial portion of the insurers' resources is always available to absorb the fluctuations inherent to the business. In contrast, managers will be interested in keeping reserve levels to a minimum, in consideration of their company's risk profile, assuring policyholders (and stockholders) that obligations (and profit levels) will be met. Finally, reserving and ratemaking actuaries will be concerned with the accuracy of the forecast, producing results that will not only keep the insurer in business, but also that will generate data for setting rates for future underwriting periods.

Case Reserves and Gross IBNR

Once a claim is reported to the insurer, the company will regularly assess, on an individual basis, the amount of the insurer's liability until its final settlement. The total amount of these known and reported claims estimates is called the *case reserve* and is the first provision to be included in the total reserve requirement [2].

Nevertheless, case reserves are subject to future adjustments, and even claims that were considered settled, must sometimes be reopened for additional payments or to take recoveries into account. Furthermore, in many cases, incurred claims are still unknown to the insurer, because they have not been reported, i.e. they are incurred but not reported (IBNR) or they were possibly reported but have not yet been recorded. All these provisions for known and unknown claims, in addition to case reserves, are included in the gross IBNR reserve. The methods surveyed in the following sections should be considered applicable to the gross IBNR reserves.

A word of warning: even after employing methods that are sanctioned by the supervisors and judged appropriate by actuaries, financial analysts, accountants, and the management, a company could still build inadequate reserves. The final test relies on the accuracy of the forecast.

General Aspects of Loss Reserving Methods

The method to use in each situation depends on many factors: endogenous (those that can be controlled by the company) and exogenous (those that are determined by forces external to the company). Among the endogenous factors, the main ones are: specific risk for which the reserves are being calculated, the type and amount of data, size and nature of the insurance company, computing resources, and technical expertise of the person doing the estimation. The main exogenous factors are the macroeconomic and business environments, and the underwriting cycle [1].

The determination of claims reserves is generally a two-step process. The first involves the compilation of a suitable set of historical data, possibly classified according to the lines of business, as well as by the accident or underwriting year. The data will usually consist of one or more of the following: number of claims reported, the number settled, and the amounts paid. Sometimes, there will be a measure of exposure such as the number of units covered or total premiums earned. The underwriting experience will also be recorded as it evolves over time for groups of risks, preferably homogeneous. The second step consists of the selection from all projection methods available, which range from the very simple to the very sophisticated stochastic models.

Nonlife claims reserving methods are usually classified as stochastic (sometimes called *statistical*) and *nonstochastic* (also called *deterministic*), depending on whether or not they allow for random variation [1, 3]. The nature and complexity of the methods have been evolving over time along with the development of statistical methods, as well as computing power and software availability. Deterministic methods are usually based on various constants or ratios that are assumed to be known without error and are also assumed to remain valid in the future. Stochastic methods (*see Stochastic Control for Insurance Companies*), on the other hand, include explicitly the random variation that can be present in the components of the model, or they use a random error term that is meant to include all the unknown effects. The inclusion of a random component makes them amenable to formal statistical estimation and inference. Furthermore, stochastic methods have the advantage over deterministic

ones that it may be possible to generate probability distributions, and consequently probability intervals for unsettled claims. Using these methods, we can compute, in addition to the estimated values of the reserves, some traditional risk measures, such as value at risk (VaR) (*see Value at Risk (VaR) and Risk Measures*), Tail VaR, and others [3, 4].

We normally estimate reserves using data with some degree of aggregation or grouping. The calendar period when the claim occurs or is reported is called the *period of origin* or *accident period*, and all claims occurring within this period are analyzed as a group. These claims are then followed as they are paid over several development periods, until they are all ultimately paid out. We refer to accident and development years, but the arguments apply for any other reference periods or time units.

We use what is standard notation in claims reserving. Let X_{it} = amount of claims paid in the t th development year corresponding to accident year i , for $i = 1, \dots, k, t = 1, \dots, s$ where s = maximum number of development years it takes to completely pay out the claims corresponding to a given exposure or accident year. The X_{it} are known as *incremental claims values*. In addition to incremental claims, we usually define a corresponding set of cumulative claims $C_{ij} = \sum_{t=1}^j X_{it}, i = 1, \dots, k, t = 1, \dots, s$ [2, 3]. An important quantity is “ultimate claims”, which for accident year i is defined as $N_i = C_{is} = \sum_{t=1}^s X_{it}$ i.e., “ultimate claims” is the total amount of claims that will eventually be paid out, over all s development years for those claims corresponding to accident year i [2].

Deterministic Models

In the simplest deterministic case, claims are assumed to come from a homogeneous process that is stable over time, and in consequence, some kind of average of observed claims would be representative of what we should expect in the future. Possibly, the simplest deterministic claims reserving method, based on this type of assumption, is the loss ratio method. Loss ratio can be defined in various forms, but here we consider it as the ratio of the amount of “ultimate claims” to total premiums for a specific line of business. The loss ratio thus computed will necessarily be based on past experience of the company [1, 2]. The loss ratio estimate of the reserves on a given date is

obtained as Premium income $\times$ loss ratio – losses paid to date.

It has the advantage that it is very easy to apply, yet, in its simplest form, it ignores any structure that may be present in the data, such as trends, and there is no guarantee that the historical loss ratio will continue to be true in the future. Assessment of current trends and analysis can provide additional information that can be used to improve on the estimates. For an extensive description of loss reserving concepts and methods, see [1–3, 5].

However, there usually exists some additional information that should be taken into account in the estimation process. It is usually assumed that at least we know the accident and development year for each claim, and so we use this information to group the claims. In this respect, the next level of sophistication in the deterministic models would be to separate the claims belonging to different years of origin. That is, we separate the claims according to the period in which they were reported, or accident year, and see how their experience has behaved historically.

Specifically, for each accident year, we look at the total or ultimate claims paid and how they evolved over time, i.e., the pattern of payments established over the intervening years. Given the value of ultimate claims, C_{is} , for a known number m of accident years, $i = 1, \dots, m$, the pattern of accumulated claim payments can be written as a percentage of these ultimate claims. Using our notation, we compute $P_{it} = C_{it}/C_{is}$, $i = 1, \dots, m$, $t = 1, \dots, s$.

If we can show, for subsequent accident years, that the pattern of these percentages is stable, then they provide a means for arriving at an estimate of ultimate claims for future years. There are different ways in which the P_{it} can be applied to predict future claims. The most straightforward is to obtain

an average proportion for each development year $\bar{P}_t = \sum_{i=1}^m P_{it}$, $t = 1, \dots, s$. Hence, for any accident year i for which cumulative claims are known only up to development year $r < s$, ultimate claims may be projected as $\hat{N}_i^G = C_{ir}/\bar{P}_r$, and the corresponding estimated reserves as $\hat{R}_i = \hat{N}_i^G - C_{ir}$. This is called the *grossing-up* method [1]. Total reserves are then obtained by adding up estimated reserves over all the accident years where ultimate claims had not been reached.

The chain-ladder method is probably the mostly used claims reserving method. In its original form, it is deterministic, although there are some stochastic models that reproduce the same reserve estimates. It is also based on the assumption that the development pattern is stable, so that a constant proportion of ultimate claims is paid in each development year, irrespective of the year of origin. Usually, it is assumed that $k = s$ and that the available data are in the form of a triangle [1, 2]. In the notation given above, if the k th accident year is the most recent one, and there is information only on its first development year, then there will be information on two development years for claims in accident year $k - 1$, and so on, up to the first observed accident year for which there will be k development years observed and ultimate claims are completed in those k years. The triangle will be as in Table 1.

The data matrix is split into a set of known or observed variables (the upper left-hand part) and a set of variables whose values are to be predicted (the lower right-hand side). Thus we know the values X_{it} for $i + t \leq k + 1$, and assume all $X_{it} > 0$. The triangle of known values is called the *runoff triangle*. For ease of presentation, we have defined X_{it} to be incremental claims. The figures in the cells of the runoff triangle may represent

Table 1 Runoff triangle presentation of loss data

Year of origin	Development year						
	1	2	...	t	...	$k - 1$	k
1	X_{11}	X_{12}	...	X_{1t}		$X_{1,k-1}$	X_{1k}
2	X_{21}	X_{22}	...	X_{2t}		$X_{2,k-1}$	—
3	X_{31}	X_{32}	...	X_{3t}		—	—
⋮							
$k - 1$	$X_{k-1,1}$	$X_{k-1,2}$		—		—	—
k	X_{k1}	—		—		—	—

Table 2 Cumulative loss payments for a line of general insurance and reserves under the chain-ladder method^(a)

Accident year	Cumulative loss payments (thousands of dollars)					Reserves
	Development year					Chain ladder
	1	2	3	4	5	
1988	2000	6000	9000	11 200	<i>14 000</i>	
1989	2600	6840	10 920	<i>15 600</i>	19 500	3900
1990	2380	8960	<i>14 400</i>	19 373	24 217	9817
1991	3120	<i>10 800</i>	17 003	22 875	28 594	17 794
1992	3800	12 265	19 309	25 979	32 473	28 673
Total reserve						60 184
Development factors		3.23	1.57	1.35	1.25	

^(a) Reproduced from [2] with permission from Actex Publications Inc., © 1993

different quantities: claim numbers, total payments (which can be viewed as weighted claim numbers), average payments etc. In addition, the figures can be cumulative or in incremental form. Usually, one form can be transformed into the other.

The chain-ladder method is based on the use of a set of link ratios F_{it} defined as $F_{it} = C_{it}/C_{i,t-1}$ that can be combined or weighted in various ways to obtain estimators for ultimate claims. Hence, the method is also known as the *link-ratio* method. In their most common formulation, they are used to compute development factors denoted by $\{f_j; j = 2, \dots, k\}$, defined as

$$f_j = \frac{\sum_{i=1}^{k-i+1} C_{i,j-1} F_{ij}}{\sum_{i=1}^{k-i+1} C_{i,j-1}} = \frac{\sum_{i=1}^{k-i+1} C_{ij}}{\sum_{i=1}^{k-i+1} C_{i,j-1}}, \quad j = 2, \dots, k \quad (1)$$

These factors are estimates of the growth (or decrease) in cumulative claims, between development years $j-1$ and j . Then, the chain-ladder technique estimates ultimate claims for accident year i as

$$\hat{N}_i^{CL} = C_{i,k-i+1} \cdot f_{k-i+2} \cdot f_{k-i+3} \cdot \dots \cdot f_k, \quad \text{for } i = 2, \dots, k \quad (2)$$

and outstanding claims (reserves) for the i th accident year as $\hat{R}_i = \hat{N}_i^{CL} - C_{i,i-k+1}$ or $\hat{R}_i = C_{i,k-i+1} \cdot (f_{k-i+2} \cdot f_{k-i+3} \cdot \dots \cdot f_k - 1)$ [1, 2]. Total reserves are again obtained by adding up estimated reserves over those accident years where ultimate claims had not been reached, i.e., for $i = 2, \dots, k$.

Example 1 The upper portion of Table 2 shows the development triangle of cumulative claims for a certain line of general insurance [2]. Application of equation (1) yields the development factors, f_j , shown in the last row of the table. These are applied successively for each accident year, as indicated in equation (2), to the last cumulative amount paid, $C_{i,i-k+1}$, which is in the main diagonal (in italics), to obtain $\hat{N}_i^{CL}$. This is the estimate of the amount that will ultimately be paid out for each accident year. These values are shown in the column corresponding to development year 5. They are then used to compute the required reserve for each accident year, $\hat{R}_i$. The total required reserve is \$60 184.

Variants of the method have been proposed, such as the Bornhuetter–Ferguson method, which uses external information to obtain initial estimates of ultimate claims for each accident year. This can be done by using loss ratios or alternative techniques. The estimates of ultimate claims are then combined with the development factors of the chain-ladder technique to estimate the required reserves [1, 2].

Deterministic claims reserving methods are relatively straightforward to apply and can easily be implemented in a spreadsheet. However, they have some drawbacks. They all assume a stable development pattern and, in their simplest form, the grossing-up and chain-ladder methods do not take possible trends into account. They can be modified to take into account these elements, but at the expense of their simplicity. The main disadvantage of all deterministic models, however, is that it is not possible to obtain variance estimates for the reserves.

Stochastic Methods

The increase in computing power, developments in statistical software, and the need to incorporate additional elements in the modeling and estimation of claims reserves, as well as the need to obtain measures of their variability, have gradually led to the development and implementation of stochastic methods. These methods add one or more stochastic elements to the claims reserving process. In essence, claims amounts or counts are considered as random variables, and models are specified to include as many components as necessary to take into account all the factors that influence the claims process, endogenous or exogenous. In addition, stochastic models allow the validity of the model assumptions to be tested statistically, and produce estimates not only of the expected value, but also of their variance [1, 3, 4, 6–9].

Like deterministic models, stochastic models usually assume a stable runoff pattern, at least initially. However, even if the past settlement pattern is reasonably stable, there may be considerable uncertainty regarding whether it will continue in the future. Another source of uncertainty is claims inflation. The provision to be held will also be affected by assumed investment earnings. In practice, it is seldom possible to do any better than suggest a fairly wide range, of not-unreasonable provisions, preferably in the form of probability intervals, based on different assumptions as to future claim inflation, investment earnings, etc., and possibly on different statistical methodologies. For an extensive description of stochastic loss reserving concepts and methods, see [1, 3, 4].

Some methods have been proposed that reproduce results from the chain ladder, while obtaining measures of their variability [7–10]. Others are intended to take advantage of formal statistical modeling and testing framework [6, 11–13]. One of the more widely used and general formulations is the following. Let the claims in the t th development year of accident year i , be represented by the random variable X_{it} (where $X_{it} > 0$), $i, t = 1, \dots, k$. We assume the data are in the form of a runoff triangle so that we have known observed values for $i + t \leq k + 1$.

Define $Y_{it} = \log(X_{it})$, and assume, in addition, that,

$$Y_{it} = \mu + \alpha_i + \beta_t + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2) \quad (3)$$

$i = 1, \dots, k, t = 1, \dots, k$ and $i + t \leq k + 1$, i.e., an unbalanced two-way analysis of variance (ANOVA) model. This was originally used in claims reserving by Kremer [11]; see also [1, 3, 12–14].

Thus X_{it} follows a lognormal distribution, and

$$f(y_{it} | \mu, \alpha_i, \beta_t, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{1}{2\sigma^2} \times (y_{it} - \mu - \alpha_i - \beta_t)^2 \right] \quad (4)$$

Let $T_U = (k + 1)k/2 =$ number of cells with known claims information in the runoff triangle; and $T_L = (k - 1)k/2 =$ number of cells in the lower triangle, whose claims are unknown that we want to estimate, to obtain the reserves. If $\underline{y} = \{y_{it}; i, t = 1, \dots, k, i + t \leq k + 1\}$ is a T_U -dimension vector that contains all the observed values of Y_{it} , and $\underline{\theta} = (\mu, \alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k)'$ is the $(2k + 1)$ vector of parameters, then the Likelihood function can be written as

$$L(\underline{\theta}, \sigma | \underline{y}) = (2\pi)^{-T_U/2} \sigma^{-T_U} \exp \left[-\frac{1}{2\sigma^2} \sum_i^k \sum_t^k \times (y_{it} - \mu - \alpha_i - \beta_t)^2 \right] \quad (5)$$

where the double sum in the exponent is for $i + t \leq k + 1$, i.e., the upper portion of the triangle. From equation (5) we can obtain maximum-likelihood estimators of the parameters in $\underline{\theta}$ and, from these, generate estimates of the outstanding claims for the lower triangle along with their variances [4, 12–15]. Additional effects can be incorporated in the models at a cost of complicating them. Nevertheless, much richer information is obtained from the complete distribution of future claims [4, 14, 15].

Example 2 The upper portion of Table 3 shows the development triangle of incremental claims, X_{it} , for the same data as in Table 2. A special case of the model in equation (3) was also fitted to this data. For statistical and computational efficiency, μ and β_1 are set equal to zero [11, 14], but this does not affect the estimation of the reserves. Thus, maximum-likelihood estimators for the parameters were obtained for the model: $Y_{it} = \text{Log}(X_{it}) = \alpha_i + \beta_t, i = 1, \dots, 5, t = 2, \dots, 5$. They are given in Table 4.

Table 3 Incremental loss payments for a line of general insurance and reserves under the ANOVA linear regression model

Accident Year	Incremental loss payments (thousands of dollars)					Regression	
	Development year					Reserves	Standard error
	1	2	3	4	5		
1988	2000	4000	3000	2200	2800	–	–
1989	2600	4240	4080	4680	3956	3956	857
1990	2380	6580	5440	4408	4573	8981	1337
1991	3120	7680	5945	5232	5427	16 605	2140
1992	3800	8248	6800	5984	6207	27 239	3839
Total						56 782	5363

Table 4 Maximum-likelihood estimates of the parameters in model (3)

α_1	α_2	α_3	α_4	α_5	β_2	β_3	β_4	β_5
0.539	0.702	0.693	0.795	0.930	0.775	0.582	0.454	0.491

The estimated incremental claims for the lower triangle in Table 3 (shaded) are computed as $\hat{X}_{it} = \exp(\hat{\alpha}_i + \hat{\beta}_t)$ and then summed up by rows to obtain the reserves by accident year. These are given in the second to last column of the table. Total reserves are now \$56 782, somewhat smaller than those of the chain-ladder method. However, now we can obtain standard errors of the estimates using bootstrap or other methods [15]. The last column of Table 3 shows bootstrap standard errors. These may be used to obtain confidence intervals if the Normality assumption can be justified.

Further sophistication of the claims model allows the use of alternative distributions to the lognormal that is used in equation (4). One option is to model separately claim counts and average claims, i.e., let $X_{it} = N_{it} \cdot M_{it}$ where N_{it} is the number of closed claims and M_{it} is the corresponding average claim, both for the t th development year and corresponding to year of origin i . The estimates from the separate models can be combined to generate the distribution of future claims [14, 16, 17]. There is an advantage of using average claims, since more information is explicitly used in the modeling process and so the actuary can perform separate analyses for settled claims and for reported claims.

Statistical generalized linear models (GLMs) [18] can also be used in the claims reserving process, allowing us to use distributions other than the lognormal. This has been applied in claims reserving in [19, 20] by means of bootstrap methods to obtain the predictive distribution of outstanding future claims.

In fact, it can be shown that many of the existing models can be viewed as special cases of GLMs [21]. Other methods, such as the Kalman filter [22], have also been used to include additional structure in the models, as well as to provide the complete predictive distribution of future claims.

An alternative approach to claims reserving is to use Bayesian statistical methods. In many cases Bayesian methods can provide analytic closed forms for the predictive distribution of the variables involved, e.g., outstanding claims, and predictive inference is then carried out directly from this distribution and any of its characteristics and properties, such as quantiles, can be used for this purpose. However, if the predictive distribution is not of a known type, or if it does not have a closed form, then it is possible to derive approximations by means of Monte Carlo (MC) simulation methods. One alternative is the application of direct MC, where the random values are generated directly from their known distribution [14]. In recent years, the application of MC methods in loss reserving has developed extensively, in particular with the introduction of Markov chain Monte Carlo (MCMC) techniques [16, 17, 19, 23].

Bayesian methods and simulation techniques were also applied to the data of the two previous examples to generate the complete predictive distribution of future claims (in the lower triangle), Figure 1. This distribution has a mean equal to \$61 073, which should be compared with the reserves estimated by the other two methods.

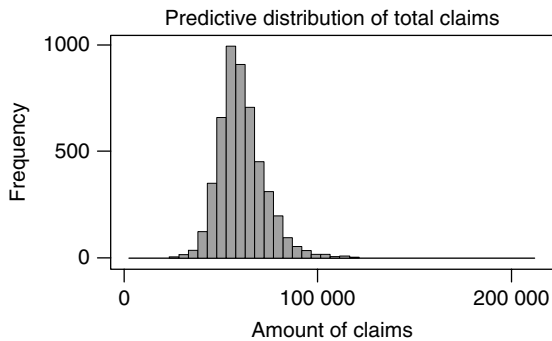


Figure 1 Predictive distribution of total claims obtained by Bayesian methods and simulation

References

- [1] Taylor, J.M., The Faculty and Institute of Actuaries (1997). *Claims Reserving Manual*.
- [2] Brown, R.L. (1993). *Introduction to Ratemaking and Loss Reserving for Property and Casualty Insurance*, ACTEX Publications.
- [3] Taylor, G.C. (2000). *Claim Reserving. An Actuarial Perspective*, Elsevier Science Publishers, New York.
- [4] England, P. & Verrall, R.J. (2002). Stochastic claims reserving in general insurance (with discussion), *British Actuarial Journal* **8**, 443–544.
- [5] Hossack, I.B., Pollard, J.H. & Zenwirth, B. (1999). *Introductory Statistics with Applications in General Insurance*, 2nd Edition, Cambridge University Press, Cambridge.
- [6] Verrall, R. (1991). The chain ladder and maximum likelihood, *Journal of the Institute of Actuaries* **118**(III), 489–499.
- [7] Verrall, R. (2000). An investigation into stochastic claims reserving models and the chain-ladder technique insurance, *Mathematics and Economics* **26**, 91–99.
- [8] Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve structure, *ASTIN Bulletin* **23**, 213–225.
- [9] Mack, T. (1994). Which stochastic model is underlying the chain ladder method? *Insurance Mathematics and Economics* **15**, 133–138.
- [10] Renshaw, A.E. & Verrall, R. (1994). A stochastic model underlying the chain-ladder technique, *Proceedings XXV ASTIN Colloquium*, Cannes.
- [11] Kremer, E. (1982). IBNR claims and the two-way model of ANOVA, *Scandinavian Actuarial Journal*, 47–55.
- [12] Doray, L.G. (1996). UMVUE of the IBNR reserve in a lognormal linear regression model, *Insurance Mathematics and Economics* **18**, 43–57.
- [13] Verrall, R. (1991). On the estimation of reserves from loglinear models, *Insurance Mathematics and Economics* **10**, 75–80.
- [14] de Alba, E. (2002). Bayesian estimation of outstanding claims reserves, *North American Actuarial Journal* **6**(4), 1–20.
- [15] England, P. & Verrall, R. (1999). Analytic and bootstrap estimates of prediction errors in claims reserving, *Insurance Mathematics and Economics* **25**, 281–293.
- [16] Ntzoufras, I. & Dellaportas, P. (2002). Bayesian Modelling of outstanding liabilities incorporating claim count uncertainty, *North American Actuarial Journal* **6**(1), 113–136.
- [17] Scollnik, D.P.M. (2001). Actuarial modeling with MCMC and BUGS, *North American Actuarial Journal* **5**(2), 96–125.
- [18] Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E. & Thandi, N. (2004). *A Practitioner's Guide to Generalized Linear Models*, CAS 2004 Discussion Paper Program.
- [19] Verrall, R. (2004). A Bayesian generalized linear model for the Bornhuetter-Ferguson method of claims reserving, *North American Actuarial Journal* **8**(3), 67–89.
- [20] Pinheiro, P.J.R., Andrade e Silva, J.M. & Centeno, M.L. (2003). Bootstrap methodology in claim reserving, *The Journal of Risk and Insurance* **70**, 701–714.
- [21] Kaas, R., Goovaerts, M., Dhaene, J. & Denuit, M. (2001). *Modern Actuarial Risk Theory*, Kluwer Academic Publishers, Boston.
- [22] de Jong, P. (2006). Forecasting runoff triangles, *North American Actuarial Journal* **10**(2), 28–38.
- [23] de Alba, E. (2006). Claims reserving when there are negative values in the runoff triangle: Bayesian analysis using the three-parameter log-normal distribution, *North American Actuarial Journal* **10**(3), 1–15.

Related Articles

Copulas and Other Measures of Dependency

ENRIQUE DE ALBA AND DIEGO HERNÁNDEZ

Nonparametric Calibration of Derivatives Models

In financial markets, a derivative is a security whose price depends on the price of one (or more) underlying security (or securities). The most prominent example may be an S&P 500 index option, whose price depends on the S&P 500 index. While the price of the underlying security is subject to supply and demand, the price of the derivative is largely driven by arbitrage: trades that take advantage of relative price cheapness or richness against the underlying. As a result of arbitrage, the price of the derivative becomes essentially predictable if the price of the underlying is given. Nonetheless, it is not necessarily a trivial business to gauge the relative price cheapness or richness of derivatives, particularly for exotic derivatives. To consistently price and effectively hedge various exotic and ill-liquid derivatives, we use mathematical models.

A derivatives model consists of postulated price dynamics of the underlying security (or securities). The postulations are based on our understanding of price behavior of the underlying security in the past. With a model, we can price various derivatives and calculate the price sensitivities with respect to various risk factors. The premise, however, is that the model can price correctly a selected set of derivatives that dominate liquidity. These derivatives, typically consisting of standard European call and put options, are called *benchmark derivatives*. Calibration, in general, means a procedure to pin down model coefficients by reproducing the market prices of benchmark derivatives (and perhaps other inputs). The prices of the benchmark derivatives are believed to carry information on the dynamics of the underlying security for the future.

Model users usually have options between a parametric and a nonparametric calibration. Under the parametric calibration, model coefficients take the form of parametric functions, which arise from empirical findings, of time and/or state variables. Under the nonparametric calibration, on the other hand, model coefficients are expanded into a set of basis functions, e.g., piece-wise constant functions. The parameters or the expansions are determined

through a trial-and-error procedure that eventually reproduces the input prices.

It is a trade-off in choosing between these two types of calibrations with regard to efficiency, robustness, and accuracy. Because the number of parameters to solve for is small, parametric calibrations can often be executed instantly, but fitting accuracy may be fair or even poor. Nonparametric calibrations, however, can achieve higher fitting accuracy at the price of solving for a bigger number of expansion coefficients. The choice for either approach is made on the basis of practical considerations. When fitting accuracy is important, which is often the case for a liquid option market (e.g., S&P 500 index options and London Interbank Offer Rates (LIBOR) derivatives), preference is given to nonparametric calibration.

In this article, we describe nonparametric calibration technologies for equity option model and for the LIBOR market model. The former model has a single state variable, while the latter model has many. With these established technologies, we show that nonparametric calibrations in finance are far from plain applications of standard optimization routines, instead, they require delicate mathematical modeling and careful analysis.

Reconstructing Local Volatility Surface for Equity Derivatives Models

Under the deterministic volatility model for equity options, a stock price, S_t , is assumed to follow a continuous one-factor diffusion process

$$\frac{dS_t}{S_t} = (r_t - q_t) dt + \sigma(S_t, t) dW_t \quad (1)$$

where W_t is a Brownian motion under the pricing measure, r_t the interest rate, q_t the dividend yield, and $\sigma(S, t) : \mathbb{R}^+ \times [0, \tau) \rightarrow \mathbb{R}$ is a smooth and bounded deterministic function (so a unique solution to equation (1) exists). For an option with payoff $f(S_T, T)$ (to occur at time $T \leq \tau$), its time- t value, $V(S_t, t)$, satisfies the celebrated Black–Scholes–Merton equation [1]:

$$\begin{aligned} \frac{\partial V}{\partial t} + (r_t - q_t)S \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2(S, t) \frac{\partial^2 V}{\partial S^2} - r_t V &= 0, \\ V(S, T) &= f(S, T) \end{aligned} \quad (2)$$

The current option value, $V(S_0, 0)$, can be solved numerically, e.g., through a lattice method [2].

Calibration of this model is an inverse problem: given a set of market prices of options, we try to determine the volatility function, $\sigma(S, t)$. According to Dupire [3], $\sigma(S, t)$ could be calculated explicitly if we had market prices of call options for continuums of strike and maturity. This is, however, not the case. In reality, we only have the prices of around-the-money call/put options for a few maturities. The conventional approach to find a reasonably smooth $\sigma(S, t)$ is to adopt and then minimize an objective function (that regulates the smoothness of $\sigma(S, t)$) under the price constraints. This approach, however, is considered too slow for production use.

One favored approach [4] that works quite well is to use a two-dimensional spline functional to represent the local volatility function. We take $p \leq m$ spline knots, $\{(s_i, t_i)\}_{i=1}^p$, in the region $[0, \infty) \times [0, \tau]$. Given the values of the local volatility surface at those knots, $\{\sigma_i = \sigma(s_i, t_i)\}_{i=1}^p$, we uniquely define the local volatility surface through cubic spline interpolation (using certain spline end conditions). An option is then priced using the local volatility surface. Calibration is achieved by solving the following problem:

$$\begin{aligned} \min_{\{\sigma_i\}} \sum_{j=1}^m w_j [V_j(S_0, 0; \{\sigma_i\}) - \bar{V}_j]^2 \\ \text{subject to } l \leq \sigma_i \leq u, \quad \forall i \end{aligned} \quad (3)$$

where $\{V_j(S_0, 0; \{\sigma_i\})\}$ are option values calculated using a lattice method, $\{\bar{V}_j\}_{j=1}^m$ are market prices of the options, and $\{w_j\}$ are positive weights, which allow users to place importance on each input price according to liquidity or other considerations. Problem (3) can be solved efficiently by standard minimization algorithms.

This approach achieves a smooth solution through cubic spline interpolations, and thus avoids using a smoothness regularization. Note that the deterministic volatility model enjoys certain degree of popularity because of its intuitiveness, although there is a continuing debate in the industry over the performance of the model.

Reconstructing Local Volatility Surface for LIBOR Market Model

LIBOR market model is considered the benchmark model for interest-rate derivatives. Let $f_j(t)$ be the arbitrage-free forward lending rate (for simple compounding) seen at time t for the period (T_j, T_{j+1}) , the LIBOR market model takes the form

$$\begin{aligned} df_j(t) = f_j(t) \gamma_j(t) \cdot (d\mathbf{W}_t - \sigma(t, T_{j+1}) dt), \\ 1 \leq j \leq N \end{aligned} \quad (4)$$

where W_t is a multidimensional Brownian motion under the pricing measure, $\gamma_j(t)$ is the volatility of $f_j(t)$, and $\sigma(t, T_{j+1})$, the volatility of the T_{j+1} -maturity zero-coupon bond, is an explicit function of $\{\gamma_j(t)\}$ (see [5]). Note that $\gamma_j(t) = 0$ for $t \geq T_j$, as at $t = T_j$ the forward rate $f_j(t)$ is set (or fixed) and thus becomes “dead”. Calibration here means to specify $\{\gamma_j(t)\}$ by reproducing, typically, input prices of at-the-money (ATM) swaptions and forward-rate correlations.

A swaption is an option to enter into a swap, which is a financial contract to exchange interest payments, a fixed-rate stream for a floating-rate stream. The floating rate is indexed to, typically, a par yield in the LIBOR market. The swaption for a future period, say (T_m, T_n) , can alternatively be priced as a series of contingent cash flows, $\Delta T_j (R_{m,n}(T_m) - K)^+$, to occur at T_j , $j = m + 1, \dots, n$, where $\Delta T_j = T_{j+1} - T_j$ is the j^{th} accrual interval, K is the prespecified fixed rate for the underlying swap, and $R_{m,n}(t)$ is the time- t prevailing swap rate (or par yield) for the period (T_m, T_n) , given by

$$R_{m,n}(t) = \sum_{j=m}^{n-1} w_j(t) f_j(t) \quad (5)$$

with

$$w_j(t) = \frac{\Delta T_j P(t, T_{j+1})}{A_{m,n}(t)} \quad (6)$$

Here, $P(t, T_{j+1})$ is the discount factor for cash flows occurring at T_{j+1} , and

$$A_{m,n}(t) := \sum_{j=m}^{n-1} \Delta T_j P(t, T_{j+1}) \quad (7)$$

is the value of an annuity. The payoff of the option at time $t = T_m$ is thus

$$V(T_m; T_m, T_n, K) = A_{m,n}(T_m) \cdot \times (R_{m,n}(T_m) - K)^+ \quad (8)$$

In the LIBOR market, swaptions are priced using Black's formula:

$$V(t; T_m, T_n, K) = A_{m,n}(t)[R_{m,n}(t)N(d_+) - KN(d_-)] \quad (9)$$

where $N(\cdot)$ is the standard normal accumulative function,

$$d_+ = \frac{\ln \frac{R_{m,n}(t)}{K} + \frac{1}{2} \bar{\sigma}_{m,n}^2 T_m}{\bar{\sigma}_{m,n} \sqrt{T_m - t}} \quad (10)$$

$$d_- = d_+ - \bar{\sigma}_{m,n} \sqrt{T_m - t} \quad (11)$$

with

$$\bar{\sigma}_{m,n}^2 = \frac{1}{T_m - t} \sum_{j=m}^{n-1} \sum_{k=m}^{n-1} w_j(t) w_k(t) \times \int_t^{T_m} \gamma_j(s) \cdot \gamma_k(s) ds \quad (12)$$

In the market, the swaption price $V(t; T_m, T_n, K)$ can be observed, and is quoted using $\bar{\sigma}_{m,n}$, the so-called implied (Black's) volatility of the swaption. When $n = m + 1$, $R_{m,m+1}(t) = f_m(t)$, i.e., the swap rate reduces to a forward rate, and the swaption reduces to an option on $f_m(t)$, called a *caplet*, which is also actively traded. Owing to the one-to-one correspondence between the to the dollar price and its implied volatility, the problem of calibrating the LIBOR model can be equivalently stated as follows: given

1. implied volatilities, $\{\bar{\sigma}_{m,n}\}$, for a set of ATM swaptions (including caplets); and
2. correlation matrices for forward rates, $\{\mathbf{C}^i = (C_{jk}^i)\}$,

determine $\{\gamma_j(t)\}$ so that both the implied volatilities and correlation matrices are matched.

Under the nonparametric approach, we look for piece-wise constant volatilities of the form

$$\gamma_j(t) = \gamma_j^i = s_{i,j}(a_{j,1}^i, a_{j,2}^i, \dots, a_{j,n}^i)^T =: s_{i,j} \mathbf{a}_j^i, \quad \text{for } t \in [T_{i-1}, T_i), \quad 1 \leq i \leq j \leq N \quad (13)$$

with

$$s_{i,j} = \|\gamma_j^i\|_2 \quad (14)$$

and

$$\|\mathbf{a}_j^i\|_2 = 1 \quad (15)$$

In terms of $\{s_{i,j}\}$, equation (12) becomes

$$\bar{\sigma}_{m,n}^2 = \frac{1}{m} \sum_{i=1}^m \sum_{j,k=m,n-1} s_{i,j} s_{i,k} (w_j w_k C_{j,k}^i) \quad (16)$$

while correlation matching implies that there should be

$$(\mathbf{a}_j^i)^T \mathbf{a}_k^i = C_{jk}^i, \quad i \leq j, k \leq N \quad (17)$$

As already seen, the equations for $\{\sigma_{i,j}\}$ and $\{\mathbf{a}_j^i\}$ are decoupled, and the solution of $\{a_j^i\}$ can be obtained readily from a Cholesky matrix decomposition of $\mathbf{C}^i$. The problem for $\{s_{i,j}\}$, however, is underdetermined and thus ill-posed.

To make the problem for $\{s_{i,j}\}$ well posed, we adopt a smoothness regularization

$$\min \|\nabla \mathbf{s}\|^2 + \epsilon \|\mathbf{s} - \mathbf{s}_0\|^2 \quad \text{for some } \epsilon > 0 \quad (18)$$

where $\|\cdot\|$ is the usual two norm and ∇ is the discrete gradient operator such that

$$\nabla s_{i,j} = \begin{pmatrix} s_{i,j} - s_{i-1,j} \\ s_{i,j} - s_{i,j-1} \end{pmatrix} \quad (19)$$

Therefore, to determine $\{s_{i,j}\}$, we solve equation (18) under the constraints equation (16). Note that the second term of the regularization function allows the use of *a priori* volatility surface, and it also ensures the positive definiteness of the objective function. Under matrix notations,

$$S = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{pmatrix} \quad \text{with} \quad s_i = \begin{pmatrix} s_{i,i} \\ s_{i,i+1} \\ \vdots \\ s_{i,N} \end{pmatrix} \quad (20)$$

we can recast equation (18) and equation (16) into

$$\begin{aligned} \min_S & (S - \tilde{S}_0)^T A (S - \tilde{S}_0) \\ \text{s.t.} & \quad S^T G_{m,n} S = \tilde{\sigma}_{m,n}^2, \quad (m, n) \in \Lambda \end{aligned} \quad (21)$$

for some matrices A and $\{G_{m,n}\}$ with $A^T = A > 0$, $G_{m,n} = G_{m,n}^T \geq 0$, and a constant vector $\tilde{S}_0$. Here, Λ stands for the set of input swaptions. Note that equation (21) differs from a conventional quadratic programming problem in its quadratic instead of linear constraints. Because of that, conventional quadratic programming methodologies do not work for (21). Fortunately, such an unconventional problem can be

solved by the following strategy. We first superimpose a convex and monotonically increasing function to the objective function:

$$\begin{aligned} \min_S & U \left((S - \tilde{S}_0)^T A (S - \tilde{S}_0) \right) \\ \text{s.t.} & \quad S^T G_{m,n} S = \tilde{\sigma}_{m,n}^2, \quad (m, n) \in \Lambda \end{aligned} \quad (22)$$

where we can take, for instance, $U(x) = x^2$. Then, we solve equation (22) along the approach of the method of Lagrange multipliers. As a consequence of its additional convexity, the objective function now dominates the constraints, and equation (22) features a unique solution. Such a strategy takes full advantage of the quadratic functionality in both objective function and constraints. In fact, on the basis of the Lagrange multiplier method, we can develop a Hessian-based descending algorithm, in which each descending search is achieved by solving a symmetric eigenvalue problem, and thus renders an algorithm fast enough for production use. We refer to [6] for the details of the solution procedure and analysis. Figure 1 shows the local volatility surface constructed using the sterling pound data of February 3, 1995.

References

- [1] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- [2] Hull, J. (2003). *Options, Futures and Other Derivatives*, 5th Edition, Prentice Hall, pp. 392–427.
- [3] Dupire, B. (1994). Pricing with a smile, *Risk* **7**, 32–39.
- [4] Coleman, T., Li, Y. & Verma, A. (1999). Reconstructing the unknown local volatility function, *Journal of Computational Finance* **2**(3), 77–102.
- [5] Brace, A., Gatarek, D. & Musiela, M. (1997). The market model of interest rate dynamics, *Mathematical Finance* **7**(2), 127–155.
- [6] Wu, L. (2003). Fast at-the-money calibration of LIBOR market model through Lagrange multipliers, *Journal of Computational Finance* **6**(2), 39–77.

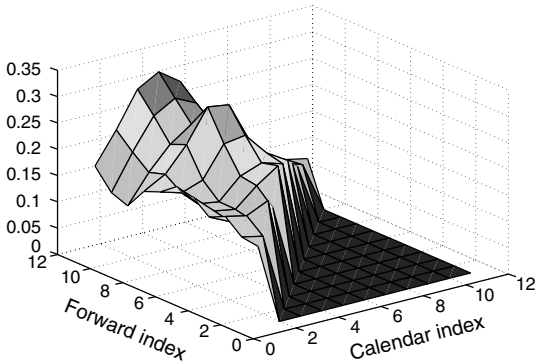


Figure 1 Volatility surface of a three-factor model

LIXIN WU

Nuclear Explosion Monitoring

Risk Analysis in Explosion Monitoring

Because the ramifications of misidentifying an explosion are so serious, seismologists and statisticians who monitor for nuclear explosions follow a near mantra that begins with the commitment to “miss no explosions”. Declaring an explosion to be an earthquake is a profoundly serious mistake by itself, but assigning a cost to this error is untenable from a nuclear explosion monitoring point of view. The cost of misidentifying an explosion could be political, financial, environmental – or most serious, the loss of life – and may not be realized for decades. The cost of declaring an earthquake to be an explosion is an operational burden that is also difficult to quantify because the reasons for making the error can vary greatly, dramatically changing the subsequent data collection and analysis. This focus on correct identification makes risk quantification in nuclear explosion monitoring unique. Risk quantification in this setting does not explicitly declare costs for decisions although costs are implicit in the Bayesian classification methods. The commitment to never misidentify historical explosions drives the prior-to-data probability that an event is an explosion to be high, regardless of the geophysical location or time frame of an event. The question of performance then centers on false alarms and the associated operational burden.

In the seismic context, the approach to identifying an unknown event begins as an examination of the seismograms of the event recorded at many locations. A seismogram is the stream of ground motion data observed by a sensor. Regional stations relatively close to the event and long-range or teleseismic stations produce seismograms similar to those shown in Figure 1, in this case from a regional station. The next step is to extract a set of seismogram features that have previously been determined as useful for discriminating between earthquakes and explosions. They are nearly always related to characteristics of the early arriving body wave or P phases and the later arriving shear or S phases, labeled in Figure 1. It is clear from this straightforward example that characteristics of the relative amplitudes or frequencies of the P or S phases might be sufficient

for discriminating between these particular events. However, the distinctions can become quite obscure for many events and additional discriminating features are needed (see, for example [1–3]).

Likelihood-Based Characterizations of Risk

The discriminating features are either continuous or categorical variables that are extracted from a collection of seismograms. Examples include event hypocenter (latitude, longitude, and depth) and the ratio of two different magnitude measures. Seismologists need the ability to examine summary statistics from the discriminants separately because there are features that can be used to unequivocally rule out the event as an explosion. For example, locating the event in the ocean or in an unlikely testing region with high confidence provides strong evidence against the explosion hypothesis. Estimated depths of greater than 20 km also provide strong evidence against the event being an explosion. There is the strong possibility that only a small number of discriminants will be observed. Therefore, a critical component of the system is a measure that allows an analyst to evaluate the discriminants one at a time in the general process of ruling out explosions.

To summarize the analysis, the data (discriminants) used to make a decision as to whether an explosion can be ruled out come as a set of potentially incompletely observed discriminants, say $D_1, D_2, \dots, D_n$. Conceptually, we assume that these are random variables with a given joint distribution that can either be developed from a model or estimated from empirical data. The decision to concentrate on ruling out explosions has led the explosion monitoring community to focus on the null hypothesis H_0 that the event is an explosion. Technical evidence is then sought to disprove this hypothesis. Suppose that, in any given case, we observe a value d_i for a particular discriminant. The p value is defined in the usual way for this single discriminant, that is, as $P_i = \text{Prob}\{D_i > d_i | \text{explosion}\}$, assuming that large values of the discriminant lead to small p values, and therefore, to rejection of H_0 , the explosion hypothesis. The seismologist looks at the collection of p values, regarding them singly as evidence against the explosion hypothesis when they are small.

Although it may be sufficient in many cases to use the collection of p values, $P_i, i = 1, \dots, n$, to

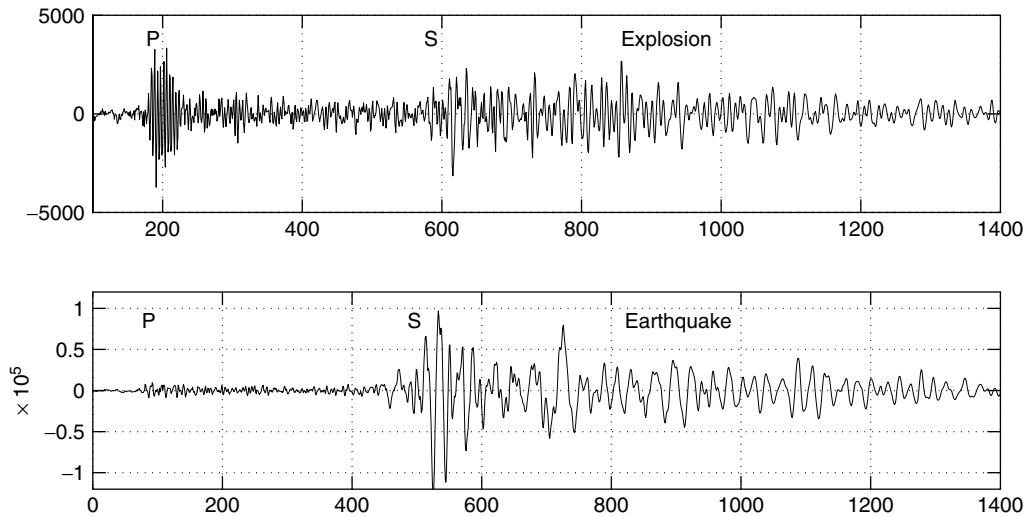


Figure 1 A nuclear explosion from the Nevada test site compared to a nearby earthquake as recorded from regional distances. These two are easily distinguished and do not represent the difficulties that arise when discriminating in general

make a decision, the overall risk may be reduced by combining the individual values to obtain an overall discriminant score and associated p value. A linear discriminant analysis (*see Credit Scoring via Altman Z-Score*) using raw waveform measurements as discriminants has been the historical approach to handling this problem (see, for example [4–6]).

In [7], a linear–quadratic discrimination compromise using regularization is proposed for seismic discrimination. In [8], arcsine transformed p values as inputs to discriminant analysis is proposed (the arcsine transformed p values (Y_i) are called *standardized discriminants*). Standardized discriminants are approximately normally distributed, and are unitless and comparable from event to event and across geographical regions. The theoretical properties of p values (standardized discriminants) as discrimination features is an open research question. The advantages they provide for uniformity in scaling and treating categorical inputs seem persuasive. The event classification matrix (ECM) analysis proposed in [8] includes reporting individual discriminant p values, multivariate discrimination scores, and posterior classification probabilities for each event under analysis.

Because the historical data set cannot be an exhaustive one, an event may be dissimilar enough from historical events that no decision can be made, and further investigation is warranted. As such, the capability to make a *no identification* decision

is integral to the process. Conceptually, we may illustrate the possible decisions as in Figure 2 for two observed discriminants d_1 and d_2 leading to transformed p values Y_1 and Y_2 . Each panel shows the joint distributions of the two transformed p values along with their projections on the two dimensional Y axis. Typical observations leading to classification into the categories (a) explosion, (b) earthquake, (c) indeterminate, and (d) unidentified are shown in Figure 2. The distinction between indeterminate and unidentified is that indeterminate is on the boundary between known distributions and we are unable to make a decision, whereas unidentified is inconsistent with both of the event categories.

Discriminants Used in Explosion Monitoring

The first four discriminants given below are well established for monitoring from teleseismic distances (greater than 2000 km). The physical basis for these discriminants is discussed at length in the literature (see [1, 9–11]). The details for a statistical formulation of these discriminants is in [8]. Recently, the necessity for monitoring at regional distances (less than 2000 km) has stimulated renewed interest in using data from regional arrays. The last feature is an example of a useful regional discriminant that

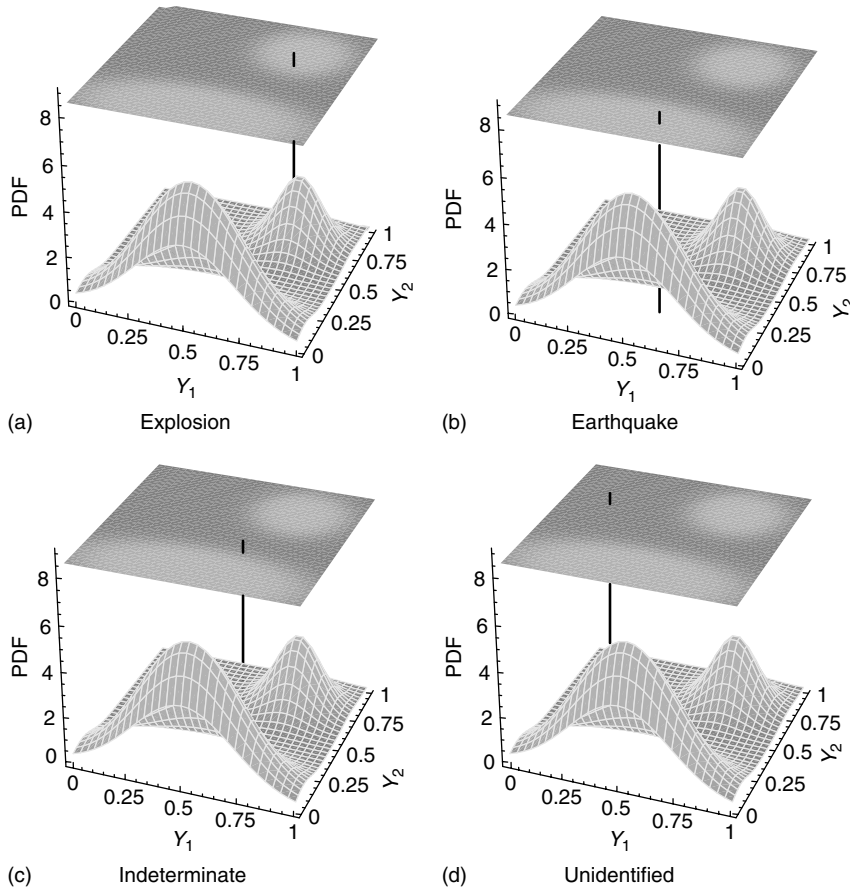


Figure 2 Illustration of likelihood functions showing observed events that would be classified into the four categories (a) explosion, (b) earthquake, (c) indeterminate, and (d) unidentified

can lead to improved methods of classification from regional data (for example, see [2, 3, 10]).

Location, Depth

The combined costs and limitations of mining and drilling technology make deep underground explosions very unlikely. As a part of standard processing, all earthquakes and potential explosions are located with a confidence region using a nonlinear least-squares model. P-wave arrival times at each of the seismometers are related nonlinearly with a travel time model to a common latitude, longitude, and depth (hypocenter). If the event is located with high confidence in an area where testing is unlikely, one may discount the explosion hypothesis. If z is the true depth and z_0 is the critical depth beyond which

an explosive source can be buried, the null hypothesis consistent with explosion characteristics is $H_0 : z \leq z_0$. The likelihood ratio test for $z = z_0$ yields a monotone function of a t -statistic. The p value is simply the tail probability of a central t -distribution calculated from the observed value of the test statistic. A small p value leads to an inference inconsistent with H_0 ; a large one leads to the inference consistent with H_0 .

Reflected pP Depth Phase

The P wave travels directly from a teleseismic disturbance to a seismic station. In contrast, the pP wave is a reflected P wave that travels from the teleseismic disturbance to the earth's surface before being reflected to the station. For a single seismic

station, the arrival time delay between the P wave and pP wave is a function of event depth. Explosions and shallow earthquakes (SEQs) produce very short time delays across a network of stations and the pP waves are therefore very difficult to detect. The pP wave from a deep earthquake (DEQ) is more readily detected across a network of stations because of longer delay times that exhibit a systematic relationship with distance to the event. Observed pP waves with a systematic pattern of delays *versus* distance is counted as evidence against the null hypothesis that the event is shallow (e.g., an explosion). The statistical model providing the depth phase discriminant is a compound one that combines (a) the number of observed pP waves with (b) a measure of the systematic behavior of delays.

Surface-Wave Energy

The ratio of body wave energy to surface-wave energy is based on the physics that earthquakes at depths less than roughly 50 km excite relatively more surface-wave energy than a single-shot explosion. In log units, the null hypothesis consistent with explosions is that the surface wave magnitude (M_S)

is less than the body wave magnitude (m_b), that is, $H_0: M_S - m_b \leq \Delta_0$. A small p value leads to the inference inconsistent with H_0 ; a large one leads to the inference consistent with H_0 .

Several different estimators for surface-wave energy exist, ranging from the 20-s waves originally proposed by Marshall and Basham [11] to the recent 7-s waves investigated by Bonner *et al.* [12], which are available for regional events like those shown in Figure 1. Figure 3 shows 40 SEQs and 126 explosions in the western United States presented in [12]. Note the good separation between the earthquake and explosion populations.

Polarity of First Motion

The direction of the first prominent cycle of the P-wave arrival or *polarity* is a useful discriminant because the energy of an underground single-shot explosion produces upward P-wave energy at a seismometer, regardless of its location. In contrast, an earthquake can produce both upward and downward P-wave arrivals across a well-dispersed global network of stations. If the polarity is upward at all stations, then the discriminant is not used because

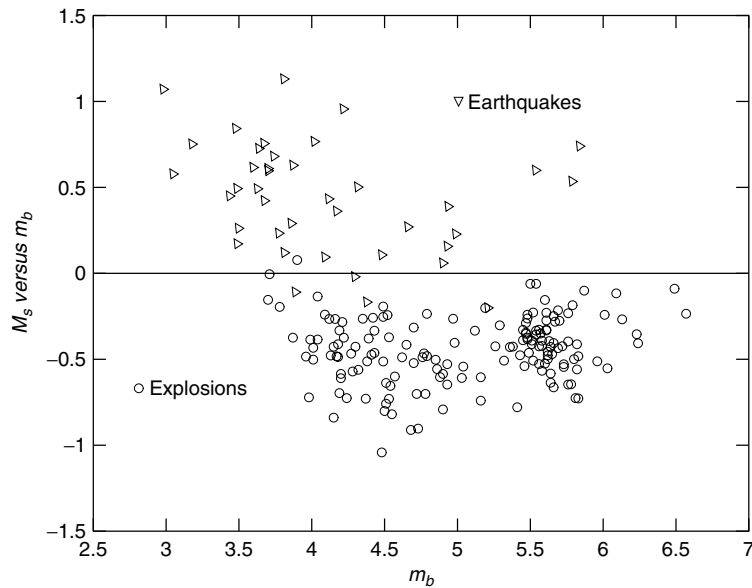


Figure 3 Plot of M_S versus m_b discriminant values for 40 shallow Western United States earthquakes and 126 explosions. The scores are based on the likelihood-ratio criterion, derived assuming regressions with equal slopes and unequal intercepts. Equal prior probabilities were assumed for the two populations [Reproduced from [12]. © Seismological Society of America, 2003.]

station configuration may play a role in observing only upward polarity. The model for the polarity discriminant also includes the probability of correctly identifying the direction of first motion, which is often quite difficult.

Relative Energy in Regional P and S Frequency Bands

The two events in Figure 1 suggest that the amplitudes and short-term spectra of the P and S phases behave differently for explosions and earthquakes. In particular, the ratio of P to S power integrated over the frequency band 4–8 Hz turns out to be larger for explosions than for earthquakes. Hence the logarithms of the ratio of P to S spectra can be used as a discriminant.

An Example: The Event Classification Matrix (ECM)

The hypothesis-testing framework results in discriminant p values that are unitless and comparable from event to event and across geographical regions. Also, because a p value is derived from an observed discriminant, it may be viewed as a discriminant. ECM analysis uses this dual interpretation of p values.

Risk management begins with the prior probabilities that an event is an earthquake or explosion. Combining the priors and the likelihoods of standardized discriminants (Y_i) with Bayesian classification methods (see **Statistics for Environmental Toxicity**) yields an estimate of the posterior probabilities that an event is an earthquake or an explosion. Examining how posterior classification probabilities are modified by the data provides useful input to forensic analysis.

The individual discriminant p values, and a combined analysis with standardized discriminants that in part gives posterior probabilities comprise ECM analysis. Tables 1 and 2 demonstrate how the ECM may be used to communicate the individual discriminant p values and subsequent Bayesian decision. Table 1 shows ECM results from a SEQ with body wave magnitude 5.32 that occurred near the border between Uzbekistan and Turkmenistan. Table 2 shows results from a DEQ with body wave magnitude of 5.39 in the Andes mountains region of Argentina.

The hypocenter depth discriminant for the shallow event returns a p value of 0.96, a clear indication that the event is shallow, and thus consistent with both SEQs and explosions. Conversely, the surface-wave energy discriminant strongly rejects the explosion hypothesis with a p value of 0. Note that neither the reflective phase nor first motion discriminants are available for this event, yet a decision can

Table 1 ECM results for shallow earthquake (SEQ) in south central Eurasia

Discriminant	Explosion hypothesis	p value
Hypocenter depth	$H_0 : \text{depth} \leq z'_0$	0.96
Reflective phases	$H_0 : \text{no observed } pP$	Not observed
Surface-wave energy	$H_0 : \tilde{M}_S - \tilde{m}_b \leq \Delta_0$	0.00
Polarity of first motion	$H_0 : \text{single point explosion}$	Not observed
Posterior probability: $P(EX) = 0.083 : P(SEQ) = 0.914 : P(DEQ) = 0.001$		

EX: explosion

Table 2 ECM results for deep earthquake (DEQ) in Argentina

Discriminant	Explosion hypothesis	p value
Hypocenter depth	$H_0 : \text{depth} \leq z'_0$	0.00
Reflective phases	$H_0 : \text{no observed } pP$	0.19
Surface-wave energy	$H_0 : \tilde{M}_S - \tilde{m}_b \leq \Delta_0$	0.23
Polarity of first motion	$H_0 : \text{single point explosion}$	0.00
Posterior probability: $P(EX) = 0.0 : P(SEQ) = 0.064 : P(DEQ) = 0.934$		

EX: explosion

still be made with high confidence. Assigning prior probabilities of 0.333 for the three populations yields an estimated posterior probability of 0.914 that the event is a SEQ.

All four discriminants were observed for the second event and show that the p -value interpretation must be taken in the context of a discrimination feature rather than simply as a hypothesis test result. The hypocenter depth discriminant strongly indicates a deep event, while the first motion discriminant strongly indicates an earthquake. The surface-wave energy and reflective phase discriminants offer some contradictory evidence but are not strong enough to result in an indeterminate solution. The posterior classification rule clearly identifies (again, correctly so) the event as a DEQ with posterior probability 0.934.

In terms of forensic analysis, an important feature of the ECM is the ability to observe the behavior of each discriminant as well as the overall posterior classification results. If there is any ambiguity in the result, the individual discriminants can be used to determine where to begin the forensic investigation.

Discussion

We have summarized some of the rather unique features of the explosion monitoring problem that have led to using the ECM as an aid for making risk-adverse decisions. First, the use of features one at a time is motivated by the knowledge that several of them provide unequivocal evidence that can rule out explosive origins without proceeding to a consideration of other features. In cases where this is not possible, discriminant analysis can be used to combine the features into a summary giving both an overall explosion p value and the posterior probability that the event was an explosion. The investigator still retains the option for interpreting the results in a political context.

The core problem in explosion monitoring is screening the many events that occur everyday. The fact that overall p values are computed for a very large number of events suggests that their empirical distributions can be utilized to control the false discovery rate (FDR) (see **Microarray Analysis; Digital Governance, Hotspot Detection, and Homeland Security**). In the context of seismic identification, the FDR is defined as the ratio of

false alarms to detections (see [13]). A false alarm is mistakenly identifying an event as an explosion, whereas detection is correctly identifying the event as an explosion. Hence, the above approach to simultaneous testing for multiple events also benefits from summaries in terms of p values.

Acknowledgment

The authors acknowledge the support of Ms. Leslie A. Casey and the National Nuclear Security Administration Office of Nonproliferation Research and Development for funding this work. This work was completed by Pacific Northwest National Laboratory under the auspices of the US Department of Energy, under contract DE-AC05-RLO1830. We gratefully acknowledge the Editor, Dr. Edward Melnick for comments that improved the clarity of the final version of the paper.

References

- [1] Blandford, R.R. (1982). Seismic event discrimination, *Bulletin of the Seismological Society of America* **72**, 69–87.
- [2] Bennett, T.H. & Murphy, J.R. (1986). Analysis of seismic discrimination capabilities using regional data from western U.S. events, *Bulletin of the Seismological Society of America* **76**, 1069–1086.
- [3] Taylor, S.R., Denny, M.R., Vergino, E.S. & Glazer, R.E. (1989). Regional discrimination between NTS explosions and Western U.S. earthquakes, *Bulletin of the Seismological Society of America* **79**, 1142–1176.
- [4] McLachlan, G.J. (1992). *Discriminant Analysis and Pattern Recognition*, John Wiley & Sons, New York.
- [5] Booker, A. & Mitronovas, W. (1964). An application of statistical discrimination to classify seismic events, *Bulletin of the Seismological Society of America* **54**, 961–977.
- [6] Shumway, R.H. (1996). Statistical approaches to seismic discrimination, in *Monitoring a Comprehensive Test Ban Treaty*, A.M. Dainty & E.S. Husebye, eds, Kluwer, Dordrecht, pp. 791–803.
- [7] Anderson, D.N. & Taylor, S.R. (2002). Application of regularized discrimination analysis to regional seismic event identification, *Bulletin of the Seismological Society of America* **92**, 2391–2399.
- [8] Anderson, D.N., Fagan, D.K., Tinker, M.A., Kraft, G.D. & Hutchenson, K.D. (2007). A mathematical statistics formulation of the teleseismic explosion identification problem with multiple discriminants, *Bulletin of the Seismological Society of America* **97**(5), 1730–1741.
- [9] Evernden, J.F. (1975). Further studies on seismic discrimination, *Bulletin of the Seismological Society of America* **65**, 359–391.

-
- [10] Pomeroy, P.W., Best, W.J. & McEvilly, T.V. (1982). Test ban treaty verification with regional data: a review, *Bulletin of the Seismological Society of America* **72**, S89–S129.
- [11] Marshall, P.D. & Basham, P.W. (1972). Discrimination between earthquakes and underground explosions employing an improved M_s scale, *Geophysical Journal of the Royal Astronomical Society* **29**, 431–458.
- [12] Bonner, J.L., Harkrider, D.G., Herrin, E.T., Russell, S.A., Tibuleac, I.M. & Shumway, R.H. (2003). Evaluation of short-period near-regional M_s scales for the Nevada test site, *Bulletin of the Seismological Society of America* **93**, 1–19.
- [13] Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

DEBORAH K. FAGAN, DALE N. ANDERSON
AND ROBERT H. SHUMWAY

Numerical Schemes for Stochastic Differential Equation Models

Background

Stochastic differential equation (SDE) models are now widely used to describe quantities whose current value is known, but whose future value is uncertain. Typically, nonlinear SDEs cannot be solved analytically, and hence approximate solutions must be obtained from numerical simulations. The design and analysis of numerical methods for SDEs is quite subtle, but the field has matured to the extent where many schemes are available and their approximation properties are well understood [1–3].

Black–Scholes Asset Price Model

The simplest nontrivial stochastic differential equation model for asset price movement takes the form

$$dS(t) = \mu S(t) dt + \sigma S(t) dW(t) \quad (1)$$

Here, $S(t)$ represents the asset price at time t , and we assume that the time-zero asset price, $S(0)$, is known. The solution $S(t)$ is a stochastic process – in other words, $S(t)$ is a random variable at each time point, t . This model involves a Brownian motion, $W(t)$, and two constant parameters.

- μ : the parameter μ is called the *drift* or *expected growth* of the asset, because $\mathbb{E}[S(t)] = S(0)e^{\mu t}$.
- σ : the parameter σ is called the *volatility*. It governs the second moment of the asset price in the sense that $\mathbb{E}[S(t)^2] = S(0)^2 e^{(2\mu + \sigma^2)t}$.

In the case where $\sigma = 0$, there is no randomness in equation (1) and the model simply describes compound interest for a bank deposit under a constant interest rate μ . Hence, we may view equation (1) as modeling a situation where the “rate of return” is uncertain. This simple model has proved extremely useful for describing a wide range of financial assets, and forms the basis of the classic Black–Scholes option valuation theory [4–6]. It is tractable in the

sense that an explicit solution is available:

$$S(t) = S(0)e^{\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W(t)} \quad (2)$$

Using the defining properties of Brownian motion, we can simulate solution paths by discretizing a time interval $[0, T]$ into subintervals of length $\Delta t = T/N$ and letting $t_i = i\Delta t$. Then the asset price $S(t_i)$ evolves into the asset price $S(t_{i+1})$ according to

$$S(t_{i+1}) = S(t_i)e^{\left(\mu - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\sqrt{\Delta t}Z_i} \quad (3)$$

where the $\{Z_i\}$ are independent random variables with the standard normal distribution.

In MATLAB [7], we could compute and plot a discretized solution path as follows:

```
>> T = 1; N = 100; Dt = T/N; mu = 0.06;
sigma = 0.25; Szero = 1;
>> Spath = Szero*cumprod(exp((mu-sigma
^2)*Dt+sigma*sqrt(Dt)*randn(N,1)));
>> plot(Spath)
```

Figure 1 shows fifty such paths – in each case, the discrete points $(t_i, S(t_i))$ are joined to give a piecewise linear curve. It follows from equation (2) that $S(t)$ is lognormally distributed. In particular, at the final time $t = T$, the asset price has density given by

$$f(x) = \frac{\exp\left(\frac{-\left(\log\left(\frac{x}{S_0}\right) - \left(\mu - \frac{\sigma^2}{2}\right)T\right)^2}{2\sigma^2 T}\right)}{x\sigma\sqrt{2\pi T}}, \quad \text{for } x > 0 \quad (4)$$

with $f(x) = 0$ for $x \leq 0$. As confirmation, Figure 1 is the final asset prices $S(T)$ for 10^5 discrete asset paths, with the density $f(x)$ superimposed as a dashed line.

Simulating General Stochastic Differential Equation Models

The simple model (1) has been modified and extended in many directions as researchers have attempted to produce more realistic descriptions of particular

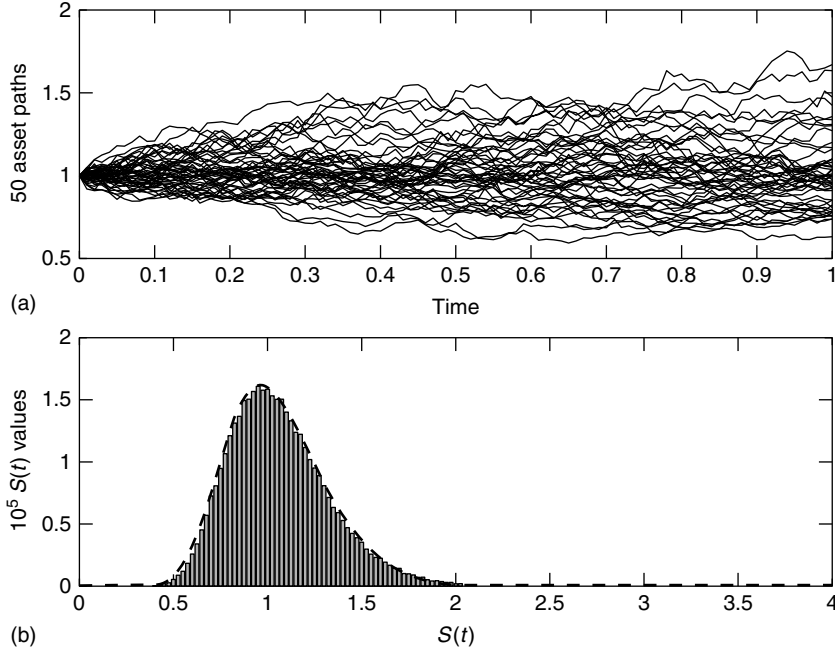


Figure 1 Asset price paths and a histogram at expiry time

financial objects. For example, the *mean-reverting square root process*

$$dS(t) = \lambda (\mu - S(t)) dt + \sigma \sqrt{S(t)} dW(t) \quad (5)$$

where λ , μ , and σ are positive constants, was originally proposed by Cox *et al.* [8] to describe interest rates, but now features in many other models. In this case, the noise coefficient scales with the square root of the asset value, when $S(t)$ is large, this is often more realistic than using the linear scaling in equation (1). For this SDE we have

$$\mathbb{E}[S(t) - \mu] = e^{-\lambda t} (S(0) - \mu) \quad (6)$$

and

$$\begin{aligned} \mathbb{E} \left[S(t)^2 - \mu^2 - \frac{\sigma^2 \mu}{2\lambda} \right] &= \left(2\mu + \frac{\sigma^2}{\lambda} \right) \\ &\times (S(0) - \mu) e^{-\lambda t} \\ &+ \left(S(0)^2 + \left(\mu + \frac{\sigma^2}{2\lambda} \right) \right. \\ &\times \left. (\mu - 2S(0)) \right) e^{-2\lambda t} \end{aligned} \quad (7)$$

so that

$$\lim_{t \rightarrow \infty} \mathbb{E}[S(t)] = \mu \quad (8)$$

and

$$\lim_{t \rightarrow \infty} \mathbb{E}[S(t)^2] = \mu^2 + \frac{\sigma^2 \mu}{2\lambda} \quad (9)$$

Hence, μ the long-term mean, λ the rate at which the mean converges, and the noise coefficient σ affect the variance. It can be shown that for $S(0) \geq 0$, nonnegativity of $S(t)$ is preserved; that is, $S(t) \geq 0$ for all $t \geq 0$, with probability one. Further, the solution may attain the value zero if and only if $\sigma^2 > 2\lambda\mu$; see [9, Section 9.2] or [10, Section 7.1.5] for more details. Heston's stochastic volatility asset price model [11] takes the form

$$\begin{aligned} dX(t) &= \lambda_1 (\mu_1 - X(t)) dt \\ &+ \sigma_1 X(t) \sqrt{V(t)} dW_1(t) \end{aligned} \quad (10)$$

$$\begin{aligned} dV(t) &= \lambda_2 (\mu_2 - V(t)) dt \\ &+ \sigma_2 \sqrt{V(t)} dW_2(t) \end{aligned} \quad (11)$$

In this case, there are two components. The squared volatility $V(t)$ in equation (11) satisfies a mean-reverting square root process. For the asset price $X(t)$ in equation (10), the random quantity $\sqrt{V(t)}$ then takes on the volatility role that was played by the constant σ in equation (1), and the drift includes mean

reversion. In this model, the two Brownian motions $W_1(t)$ and $W_2(t)$ may be correlated.

Generally, if we move away from the simple linear SDE equation (1), then analytical solutions cease to be available and numerical simulations must be performed. Suppose we have a general Ito SDE

$$dY(t) = f(Y(t)) dt + g(Y(t)) dW(t) \quad (12)$$

where $Y(t)$ has m components, and the Brownian motion $W(t)$ has d independent components, so $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^m \rightarrow \mathbb{R}^{m \times d}$, with $S(0) \in \mathbb{R}^m$ given. Then the Euler–Maruyama method computes approximations $Y_n \approx Y(t_n)$, with $t_n = n\Delta t$, of the form

$$Y_{n+1} = Y_n + f(Y_n)\Delta t + g(Y_n)\Delta W_n \quad (13)$$

where $\Delta W_n = W(t_{n+1}) - W(t_n)$ is the Brownian increment. In practice, each component of ΔW_n can be simulated by scaling a standard normal sample by $\sqrt{\Delta t}$; in MATLAB, this corresponds to `sqrtdt*randn`.

To illustrate the idea, Figure 2 shows an Euler–Maruyama approximation. Here we consider the linear SDE equation (1) with $\mu = 0.1$, $\sigma = 0.4$, and $S(0) = 1$, for $0 \leq t \leq 1$. The solid line shows the exact values of the solution (2) over a numerically generated Brownian path, at a spacing of 2^{-8} . (Discrete values have been joined in a piecewise linear fashion for clarity.) The asterisks show the values obtained by applying the Euler–Maruyama method (equation 13) with $\Delta t = 2^{-6}$. Here, the Brownian increments ΔW_n correspond to those for the exact

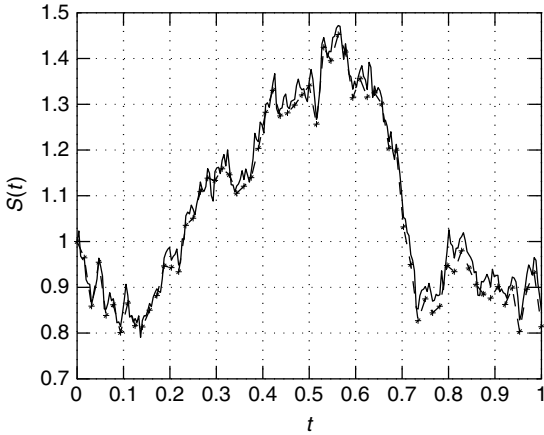


Figure 2 Exact SDE path (solid) and numerical approximation from Euler–Maruyama method (asterisks)

path. We see that Euler–Maruyama closely follows the exact path, but there is some visible error. At the final time $t = 1$, the discrepancy is 0.0359.

Traditional numerical analysis for timestepping methods regards the stepsize Δt as a small parameter and looks at the error in the numerical approximation in the limit $\Delta t \rightarrow 0$. In our context, at any time t the difference $Y(t_n) - Y_n$ between the exact and numerical solutions is a random variable, and consequently there are many, nonequivalent ways to measure the error.

One approach is to measure the “error of the means”, leading to so-called *weak convergence theory*. We say that a method has *weak order of convergence* equal to γ if there exists a constant C such that for all functions p in some class

$$|\mathbb{E}p(Y_n) - \mathbb{E}p(Y(\tau))| \leq C\Delta t^\gamma \quad (14)$$

at any fixed $\tau = n\Delta t \in [0, T]$ and Δt sufficiently small. (More precisely, the order γ is the largest such number.) Typically, the functions p allowed in equation (14) must satisfy smoothness and polynomial growth conditions. For appropriate functions f and g in equation (12), it can be shown that the Euler–Maruyama method (equation 13) has weak order of convergence $\gamma = 1$. In the case where $S(0)$ is nonrandom and $g \equiv 0$, the SDE collapses to a deterministic ordinary differential equation (ODE), whereby equation (13) becomes the Euler method, which is well known to be a first-order method in the classical ODE sense. This first-order accuracy, therefore, carries through to the SDE setting when weak convergence is considered.

By contrast, instead of the “error of the means” we could consider the “mean of the errors,” leading us into the realm of *strong convergence theory*. A method is said to have *strong order of convergence* equal to γ if there exists a constant C such that

$$\mathbb{E}|Y_n - Y(\tau)| \leq C\Delta t^\gamma \quad (15)$$

for any fixed $\tau = n\Delta t \in [0, T]$ and Δt sufficiently small. If f and g satisfy appropriate conditions, it can be shown that Euler–Maruyama has strong order of convergence $\gamma = \frac{1}{2}$. Note that this marks a departure from the deterministic setting – the presence of the stochastic term degrades the order from 1 to $\frac{1}{2}$.

It is natural to ask: which is more important, weak or strong convergence? The answer, of course, depends upon the context in which simulations are being performed. In many cases, the amount of interest is an expected value (for example, the discounted

average payoff of an option, under a risk-neutral measure), so that weak convergence is a natural concept. In other cases, it is of interest to observe particular solution paths, or to generate time series to test a filtering algorithm, making strong convergence highly relevant. The distinction has recently been blurred by the remarkable result in reference [12] that Monte Carlo estimation of SDE functionals can be made more efficient with a multilevel approach – computing solution paths over a range of Δt values, with different numbers of paths at each level. In this case, even though a weak (expected value) quantity is being computed, both the analysis and the resulting algorithm rely implicitly on the strong convergence properties of the numerical method.

Although the Euler–Maruyama method (equation 13) is still widely used, many methods with higher orders of weak and/or strong convergence have been derived. Generally, unless there is some special structure in the problem that can be exploited, achieving higher order is computationally expensive, and hence orders above 2 are very rarely used. Just as Taylor series form the basis for deriving and analyzing convergence properties of numerical methods for ODEs, the generalized concept of Ito–Taylor series forms the basis on which numerical methods for SDEs are studied. The text [2] covers the appropriate theory and gives many examples of methods. Other references in this area include [3], which deals with the state of the art in numerical SDE solving, and [13, 14] which give accessible, practical, introductions for those wishing to avoid measure theory and probability.

The SDE framework has been extended in various directions in attempts to describe real processes more accurately. For example, *jumps* [15, 16] are often introduced to account for abrupt, unpredictable “one-off” changes and, more generally, models can switch instantaneously between different SDEs to account for global changes in a physical process (for example, the market may lurch from confident to nervous, requiring a change in the volatility parameter). In many cases, it is straightforward to perform numerical simulations of these more sophisticated SDE-based models, although analyzing the approximation power of the numerical methods (e.g., proving convergence results) can be a lot more tricky [17].

References

- [1] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*, Springer, Berlin.
- [2] Kloeden, P.E. & Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin, Third Printing.
- [3] Milstein, G.N. & Tretyakov, M.V. (2004). *Stochastic Numerics for Mathematical Physics*, Springer, Berlin.
- [4] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- [5] Higham, D.J. (2004). *An Introduction to Financial Option Valuation*, Cambridge University Press, Cambridge.
- [6] Hull, J.C. (2000). *Options, Futures, and Other Derivatives*, 4th Edition, Prentice Hall, Upper Saddle River.
- [7] Higham, D.J. & Higham, N.J. (2005). *MATLAB Guide*, 2nd Edition, SIAM, Philadelphia.
- [8] Cox, J.C., Ingersoll, J.E. & Ross, S.A. (1985). A theory of the term structure of interest rates, *Econometrica* **53**, 385–407.
- [9] Mao, X. (1997). *Stochastic Differential Equations and Applications*, Horwood, Chichester.
- [10] Kwok, Y.K. (1998). *Mathematical Models of Financial Derivatives*, Springer-Verlag, Berlin.
- [11] Heston, S.I. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**, 327–343.
- [12] Giles, M.B. (2006). *Multi-level Monte Carlo Path Simulation*, Tech. Rep. NA-06/03, University of Oxford Computing Laboratory.
- [13] Higham, D.J. (2001). An algorithmic introduction to numerical simulation of stochastic differential equations, *SIAM Review* **43**, 525–546.
- [14] Higham, D.J. & Kloeden, P.E. (2002). MAPLE and MATLAB for stochastic differential equations in finance, in *Programming Languages and Systems in Computational Economics and Finance*, S.S. Nielsen, ed, Kluwer Academic Publishers, Boston, pp. 233–269.
- [15] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, Chapman & Hall/CRC, Florida.
- [16] Cyganowski, S., Grüne, L. & Kloeden, P.E. (2002). MAPLE for jump-diffusion stochastic differential equations in finance, in *Programming Languages and Systems in Computational Economics and Finance*, S.S. Nielsen, ed, Kluwer Academic Publishers, Boston, pp. 441–460.
- [17] Mao, X. & Yuan, C. (2006). *Stochastic Differential Equations with Markovian Switching*, Imperial College Press.

Related Articles

Continuous-Time Asset Allocation
Lévy Processes in Asset Pricing
Mathematical Models of Credit Risk

DESMOND J. HIGHAM AND PETER E. KLOEDEN

Occupational Cohort Studies

Occupational cohort studies provide a powerful tool to investigate the risk of disease associated with a specified exposure. By a cohort, we mean a group of individuals that are followed over time for the purpose of identifying specific outcomes of interest. A cohort is prospective if the cohort is defined prior to the observed follow-up period and retrospective if the cohort is defined at a previous point in time, prior to the period of follow-up (*see Cohort Studies*). Since for many chronic diseases, an increased risk becomes apparent long after initial exposure, prospective cohorts are often impractical. For a retrospective cohort study, the follow-up period has already occurred and thus it only remains to determine previous exposures and identify health outcomes. Using the information on employment records, exposure can usually be estimated better in an occupational cohort than in the general population.

During the 1970s and 1980s, the retrospective cohort study became the gold standard for evaluating the relationship of exposure in occupational groups to chronic disease outcomes. Although estimates of exposure can usually be obtained from employment records, retrospective measures of disease are more difficult to obtain. Since measures of morbidity are often not maintained by employers and available information may vary in quality, a common measure of health outcome in a retrospective cohort study of an occupational group is mortality. The focus of this chapter will be on issues related to the design, conduct, analysis, and interpretation of retrospective occupational cohort studies where the primary outcome is cause-specific mortality.

Definition of the Cohort

In defining the cohort, we distinguish two types – a fixed cohort and a dynamic cohort. The fixed cohort selects all individuals who are working at a previous point in time. The cohort may be further restricted by requiring that individuals work in particular departments or plants, or that they work a minimum duration of time by the specified date. A dynamic cohort permits workers to enter over a

time interval. If the interval is large, the study may contain workers with considerably different hire dates and lengths of employment and thus will provide a more representative sample of the historical workforce. However, workers entering in the later years may contribute little information to the relationship of exposure and chronic disease, since they are usually younger (with lower baseline risk of chronic disease) and have less cumulative exposure. It is standard practice in a dynamic cohort to require a minimum duration of employment to be included in the cohort. Knowledge of the etiology of the disease of interest may influence the selection of the minimal duration of required employment. However, if very short durations of employment are used to determine eligibility, there may be a large increase in individuals who have insufficient exposure to elevate their risk of disease, thus unnecessarily increasing the resources necessary to complete the study. For many studies, a requirement of at least a 1-month or even a 1-year minimal employment seems to be reasonable.

Cohort Validation

The validation of a cohort is an important aspect of the conduct of an occupational mortality study. A retrospective cohort study depends on whether the available records are adequate to form the cohort and develop a job chronology for cohort members. Some plants periodically purge records because of space and cost considerations. Such purging is unlikely to be random, since employers often maintain files by category, i.e., active employees, retirees, and deaths. The records most likely to be purged are ones that are least likely to be needed and records of employees that died or terminated employment a long time ago are often targeted for elimination. Seniority lists, interviews with long-term employees, and aggressiveness in finding old records that may be in storage should all be explored. Examination of frequency distributions by year of hire, year of termination, duration of employment, time since first employment, and first letter of last name often identifies patterns that are consistent with missing records and can be used as a basis for further inquiry with the company or union. In a cohort study of aluminum reduction plant workers, the company had the work histories of cohort members available on microfilm [1]. However, when the microfilm was compared with a cohort constructed from seniority lists, it was estimated that 28%

of the records in one plant had not been microfilmed because of a processing error. The transition by many companies to computerized records does not eliminate the need for cohort validation, since errors may occur in the process of computerization of original hard copies. Furthermore, companies that have begun to computerize may still have earlier records that are not computerized or employees where only the more recent work histories have been computerized.

Vital Status and Cause of Death Determination

When conducting a mortality study, the vital status of cohort members must be determined at the specified date defining the end of the follow-up period. The initial information in regard to the vital status of the worker is usually the employer. Workers still actively employed or receiving pension benefits are assumed to be alive. If the employee died while employed or a death benefit was paid to survivors, the company may also have information on the date of death. Resolution of the vital status of the remainder of the cohort requires more effort. The available sources useful in determining vital status have varied over the years, in part as a reflection of confidentiality laws and their interpretation. Current sources useful for vital status determination include the Pension Benefit Information Company, Social Security Administration, Health Care Financing Administration, Motor Vehicles Bureaus, the National Death Index, and the Veterans Administration. Subjects whose dates of death are after 1979 can be sent to the National Death Index to obtain death certificate numbers, which can then be used to request the death certificates from the states. These sources vary in their ability to identify vital status, and several investigators have made comparisons among them [2–5]. The process of determining vital status can take considerable time and should be initiated early in the conduct of a study. Large percentages of individuals with missing vital status or unknown cause of death could introduce bias. Fortunately, with persistence, 95% vital status determination and 95% death certificates for known deaths is usually attainable.

After death certificates are obtained for individuals whose date of death occurred in the follow-up period, the cause of death should be coded by a qualified nosologist according to strict rules guided by

the International Classification of Disease (ICD). The rules are periodically changed, and long term studies may have different ICD revisions being in effect during the follow-up period. This can create a problem when conducting the analysis, since the definition of the disease is changing over time. Some investigators code the deaths certificates in the revision in which the death occurred, while others code them in a single revision and use comparability ratios to adjust the expected number of deaths for time intervals when other revisions were in effect. Comparability ratios are usually available in the transition year that a change in the coding occurs, and provide an estimate of the expected decrease or increase of death for selected disease categories owing to the change in ICD revision.

Exposure

The employment record containing the sequences and dates of jobs and departments provides some indication of exposure. Incorporation of this information into a format that can be used to test for a relationship of exposure can be a considerable data processing task. In a cohort of 23 326 aluminum reduction plant workers [1], there were 82 521 numerically distinct coded job categories in the 14 plants. One of the standard approaches is to transform these jobs into a job-exposure-matrix (JEM). Such a matrix provides an estimate of the average daily exposure for each job for specific plants and calendar time periods. Unfortunately, exposure measurements are not likely to be available for early time periods in many retrospective studies, and it may be necessary to extrapolate from exposure measurements in the present time period, taking into consideration general plant records and changes in technology. In a large cohort of synthetic vitreous fiber production workers, Smith *et al.* [6] describe five steps in the construction of the exposure matrix for each plant: (a) development of a technical history of operations including periods of stability, points of change in exposure, and potential confounding exposures, (b) development of a department-job dictionary, (c) collection of company exposure measurements and development of quantitative measures, (d) integration of estimates with the technical history to produce exposure extrapolation tables, and (e) development of an exposure matrix for each plant. The final step is to link average exposures with

the job and department for plant and calendar time, and use this in the analysis. Preferably, quantitative estimates of exposure can be made, but for some cohorts it may only be possible to classify into categories (e.g., exposed *versus* not exposed or none, low, medium, and high). Typical summary measures of exposure include time since first exposure, cumulative exposure, average intensity of exposure, and length of time in a high exposure category. Although analysis based on a primary exposure index is preferred, it is still useful to supplement exposure analysis with analysis based on length of employment in jobs and/or departments considered exposed or highly exposed. There has been much discussion of the accuracy of the JEM approach in estimating true exposure (e.g. [7–9]). Some of the concern relates to the variability of exposures of individuals with the same job title. This variability is probably higher in population-based studies that apply the JEM to a range of industries (e.g. [10, 11]) than in studies where the JEM is constructed separately for each plant. An alternative to the JEM is to have a panel of experts construct work histories for individuals [12]. There continues to be discussion of the relative merits of the two approaches. Recent examples of the application of the JEM approach to occupational cohorts include its use in the textile industry in Shanghai, China [13], and its use in a US cohort of 15 plants producing synthetic vitreous fibers [6].

Another criticism of retrospective methods of constructing exposure is the potential error resulting from extrapolation to time periods for which there are no exposure measurements. This is always a legitimate concern, but there is no totally satisfactory way to circumvent the uncertainty of such extrapolation. The investigator should do as much validation as possible. Although no direct measurements may exist for some jobs at early time periods, sometimes, indirect measures can be used in the extrapolation. These include airborne measurements taken during the same time period at other plants with similar processes, urinary measurements that are correlated with airborne concentrations, and measurements available on other exposures that are known to be correlated with the exposure of interest.

Mortality Indices

A variety of summary indices have been used in occupational mortality studies [14, 15]. The most

common, the standardized mortality ratio (SMR), is defined as the ratio of the observed to the expected number of deaths where the expected number of deaths is determined from a standard population. Usually, but not always, it is the convention to multiply the ratio by 100. Typically, the SMR will be adjusted for age, gender, race, and calendar time, and the standard comparison is countrywide or regional cause-specific death rates.

The SMR for the i th cause of death is given by

$$SMR_i = \frac{\sum_j d_{ij}}{\sum_j D_{ij} p/P} \times 100 \quad (1)$$

where d_{ij} is the observed number of deaths from cause i in stratum j , p is the person-years at risk in the occupational group, D_{ij} is the number of deaths from cause i in stratum j in the standard population, and P is the person-years at risk in the standard population.

A person-year is defined as the equivalent of one individual's experience for 1 year. The total person-years at risk is obtained by summing the years an individual is at risk of death, over all individuals in the study. Figure 1 shows six individuals in a dynamic cohort and their corresponding person-years at risk. The cohort definition required at least 1 year employment between January 1, 1980 and December 31, 1994, and follow-up is continued through December 31, 2000. Assuming that the estimate of the expected value is based on a large population, the observed number of deaths can be assumed to be a Poisson distribution with known parameter. We can then place confidence intervals on the SMR, using the method given in Bailor and Ederer [16]. Despite its widespread use, there are theoretical and practical limitations of the SMR. When considering the usefulness of an index in comparing the mortality patterns of two populations, Silcock [17] proposes the following two desirable properties:

1. If the ratio of all the age-specific rates for the two populations is in the interval $[a, b]$ then the ratio of the overall indices for the two populations is in $[a, b]$.
2. If the overall indices of the two populations differ then the inequality should be because of the fact that some of the age-specific rates are unequal in the two populations.

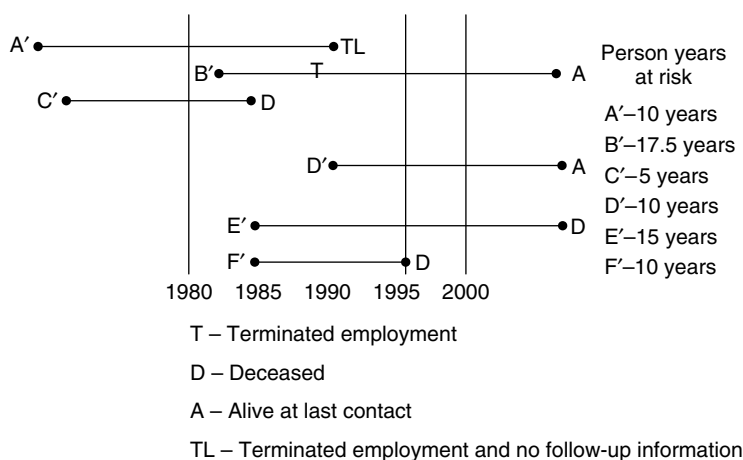


Figure 1 Person-years at risk for members of a hypothetical cohort. T: terminated employment, D: deceased, A: alive at last contact, and TL: terminated employment and no follow-up information

Silcock demonstrates the SMR has neither of these properties. Subsequent investigators have highlighted these and additional limitations of the SMR as a mortality index (e.g. [15, 18]). Although, in general, the SMR lacks some desirable mathematical properties, under certain assumptions, it is an appropriate index for comparing different exposure groups.

Gail [19] notes that if we assume the observed deaths obey a Poisson distribution indexed by age strata and population (e.g., exposure group), then the SMRs among several populations can be compared using a multiplicative model. Tests of hypothesis of equality of the SMRs can be conducted as long as there is no interaction with age. Gail provides a formal test for interaction as well as a test for equality of SMRs across populations, assuming no interaction exists. Although the same test can be conducted in the context of a regression model, Gail's approach is formulated in terms of simple χ^2 statistics and is conceptually more readily interpretable, particularly for nonstatisticians.

A decision that must be made when using a SMR is the selection of the control group. For cohorts in the United States, the most common external control group is the total US population but others can be used, including county and state populations. Arena *et al.* [20] compare the SMRs for selected causes of death for 31 000 nickel workers, using the total US population as a control group to the SMRs obtained when the local population was the control. Although mortality patterns were similar, the SMRs

tended to be lower when local population rates were used. In the conduct of the aluminum reduction plant workers study, only one of the Soderberg plants had a statistically significantly elevated SMR for lung cancer. However, this plant was located in a region known to have high rates of lung cancer. When comparisons were made with local county rates the SMR decreased from 146 to 83 and was no longer statistically significant. In general, although for most of the studies, county, state, and US rates lead to the same conclusions, there are exceptions, and possible dependence of the results on the choice of the standard population should always be considered.

The selection of the appropriate control is a difficult one in epidemiological studies, since it can result in both known and unknown sources of bias. One particular bias that has been given considerable attention in occupational cohort studies using an external control group is the healthy worker effect. The healthy worker effect is the terminology referring to a particular type of confounding that results because we are comparing a group of "healthier" workers to a general population that may include people too unhealthy to work [21–23]. Regardless of the reasons, it is frequently observed that occupational cohorts have all-cause SMRs less than 100 when compared with general populations. Unfortunately, a variety of factors affect the healthy worker effect, including age, length of follow-up, race, and specific cause of death. Methods of addressing the healthy worker effect include selection of an internal control

and testing of a dose–response relationship. This type of selection bias tends to attenuate in cohorts with longer follow-up and appears stronger for categories of cardiovascular disease than cancer. It is also possible that selection bias can occur after the workers enter the cohort, if they select themselves into job categories with less exposure or selectively terminate employment [24]. The term healthy worker survivor effect is used to describe a selection of workers from the exposed jobs of interest as a result of events occurring after initial employment. Arrighi and Hertz-Picciotto [25] provide an overview of the healthy worker survivor effect and suggest several approaches to control it. Robins [26–28] suggests some more analytically sophisticated methods of addressing this type of selection bias.

Another disadvantage of the SMR is that except for potential confounders such as age, race, gender, and calendar time, one cannot easily adjust for other potential confounding variables. Variables such as date of hire, birthplace, smoking history, and employment history at nonstudy plants are all variables that may be related to risk and confound the relationship of the exposure of interest to the cause-specific disease. Unfortunately, cause-specific death rates for general populations are not usually available for strata based on these other potential confounders, so they cannot be incorporated into the computation of an SMR. These variables are more readily controlled within the framework of a regression model.

Example 1 Rockette and Arena [1] investigated the mortality patterns of workers with five or more years of employment in 14 aluminum reduction plants. The investigation used data collected in a previous study by other investigators, so the completeness of the cohort was verified using company seniority lists. Two plants were identified as having an unacceptable number of missing records that could not be obtained (6 and 16%, respectively). This resulted in one plant being excluded from some of the analysis. The final cohort consisted of 21 829 members. Follow-up for vital status was conducted using lists of active employees and pensioners, the Social Security Administration, State Motor Vehicle Bureaus, and telephone follow-up. A complete vital status determination was made on 98.8% of the cohort, and death certificates were obtained on 97.6% of those presumed to be dead. Determination of the primary cause of death was done by a nosologist, according to the

seventh revision of the International Classification of Disease (ICDA). SMRs adjusted for age, race, and calendar year were used to compare cause-specific mortality of reduction plant workers to the total US male population. Since the follow-up period covered several ICDA revisions, comparability ratios were used to convert rates that were equivalent to the seventh revision. As expected, the “all causes” SMR was significantly low for both whites ($SMR = 88.0$) and nonwhites ($SMR = 73.2$), suggesting the presence of a healthy worker effect. Significantly elevated SMRs were found in selected working subgroups, for pancreatic cancer, lymphohematopoietic cancers, genitourinary cancers, and nonmalignant respiratory disease. SMRs were compared for different exposure groups after testing to determine if the assumptions of a multiplicative model were violated. There was a significant increasing risk from pancreatic cancer with increasing length of employment in the pot-rooms ($p < 0.01$).

Other Mortality Indices

Other indices have been used in occupational mortality studies. The comparative mortality figure (CMF) (sometimes referred to as the *standardized rate ratio*) meets both of Silcock’s requirements for a mortality index. The CMF is a method of direct adjustment where the observed deaths in the cohort are compared to an expected value obtained by applying the observed age-specific death rates in the cohort to a population with a standard age distribution. Although the CMF has more desirable analytical properties than the SMR, the estimated mortality rates tend to be highly variable for most causes of death because of the low frequency of occurrence. It is seldom used in occupational cohort studies.

Sometimes, there is no well-defined cohort, or the resources are not available to conduct a cohort study. In these situations, one can conduct a proportionate mortality study using deaths known to the employer. The proportionate mortality ratio (PMR) is used as the summary index. Steenland and Beaumont [29] used a PMR study to investigate cause-specific mortality in members of the Granite Cutters Union. Deaths were identified from union death benefit records and a statistically significant PMR of 418 was obtained for nonmalignant respiratory disease. Checkoway *et al.* [30] have noted that PMR studies

are closely related to a special type of case–control study.

The PMR compares the proportion of deaths from a specific cause to the proportion of expected deaths in a standard population. Similar to an SMR, it is summed over strata that are typically defined by age, gender, race, and calendar time. The formula for the PMR is equivalent to the formula for an SMR where the person-years at risk have been replaced by total deaths. Specifically,

$$PMR_i = \frac{\sum_j d_{ij}}{\sum_j D_{ij}d/D} \times 100 = \frac{\sum_j d_{ij}/d}{\sum_j D_{ij}/D} \times 100 \quad (2)$$

where d_{ij} is the observed number of deaths from cause i in stratum j , d is the total number of observed deaths, D_{ij} is the number of deaths from cause i in stratum j in the standard population, and D is the total number of deaths in stratum j in the standard population.

A serious limitation of proportionate mortality studies is the failure to include many of the deaths that occurred after the worker terminated employment. A second limitation is due to an undesirable property of the PMR index. Specifically, if there is a deficit in one cause of death there is correspondingly an increase in other categories. This characteristic of the PMR has been called the *see-saw effect* [31].

The PMR can be approximated by the ratio of the cause-specific SMR and the SMR for all causes [31–34]. Since the healthy worker effect usually results in an all causes SMR less than 100, and since the healthy worker effect is usually stronger for cardiovascular disease, there is a tendency for the PMR to be elevated for malignancies. This tendency for an elevated PMR for cancers is made worse by the practice of testing multiple hypothesis in the cancer category. Rockette and Arena [35] show in a simulation study that the probability of at least one falsely elevated PMR in 24 cancer categories was 0.89, if there is a healthy worker effect of 15% in all cardiovascular diseases. This was due both to multiple comparisons and the see-saw effect. In contrast, the probability of a falsely elevated SMR in a malignancy category was 0.36, and was not affected by a healthy worker effect in cardiovascular disease.

Because of these limitations, the PMR is usually used only as a screening tool to determine if there is a need for further investigation rather than a method to obtain an estimate of risk. If there is a well-defined cohort, there is no need to use a PMR instead of an SMR.

Regression Models

Although SMR analyses have historically comprised a large part of the analysis done in retrospective cohort studies, use of regression models provide several advantages. Specifically, regression analysis permits incorporation of continuous as well as discrete variables, can readily incorporate confounding variables, and can be used to test for interaction (i.e., effect modification). Although often used in studies with internal controls, they can also be formulated in a manner that can be viewed as a generalization of an SMR analysis.

Proportional Hazards Model

A cohort study can be viewed as time to event analysis where we are concerned with relating a time-dependent exposure variable to the time at which a death occurs for a specific cause. A commonly used model to investigate “time until response” variables is the proportional hazards model [36]. The proportional hazards model assumes that

$$\ln \frac{\lambda(t; \vec{x})}{\lambda_0(t; \vec{x})} = \vec{\beta} \vec{x} \quad (3)$$

where t denotes time, $\vec{x}$ denotes a vector of covariates, $\lambda_0(t; \vec{x})$ is the baseline hazard function, and $\lambda(t; \vec{x})$ is the hazard function for a given vector of covariates.

The hazard function gives the conditional failure rate which we define to be the limit of the probability (as $\Delta t \rightarrow 0$) that an individual has an event in the small interval $(t, t + \Delta t)$ given that the individual has survived to time t .

The covariate vector $\vec{x}$ may contain components that relate to exposure as well as potential confounders. For cumulative exposure, a lag time is sometimes used, since recent exposures are unlikely

to cause chronic disease. Gilbert [37] used the proportional hazards model to investigate the relationship of radiation exposure to mortality from myeloid leukemia. Instead of treating confounding variables as covariates, Gilbert treats them as stratification variables, and cumulative exposure becomes the only covariate. This approach results in a more robust model, since the covariates are not required to satisfy the proportionality assumption.

Assuming only one covariate x (cumulative exposure), the derivative of the log likelihood function with respect to β is given by

$$L'(\beta) = \sum_{i=1}^n x_i(t_i) - \frac{\sum_{k \text{ in } R_i} x_k(t_i) e^{\beta x_k(t_i)}}{\sum_{k \text{ in } R_i} e^{\beta x_k(t_i)}} \quad (4)$$

where i indexes the deaths due to a particular cause, t_i is the calendar year of death for worker i , R_i denotes the risk set of workers at time t_i , and $x_k(t_i)$ is the cumulative exposure for the k th worker at time t_i . The risk set at time t_i is the set of workers who are alive and can have an event at time t_i . If confounding variables are addressed using stratification, then the risk set (R_i) is restricted to workers having the same set of confounding variables as the worker with a death at time t_i . Standard numerical techniques can be used to solve the equation $L'(\beta) = 0$ to obtain the maximum-likelihood estimate. Statistical packages such as SAS provide several methods to test the hypothesis $H_0 : \beta = 0$, including the likelihood-ratio test, Wald's statistic, and the score statistic. The score statistic provides a particularly intuitive answer. Asymptotically, each $x_i(t)$ is normally distributed with mean μ_i (estimated by the average exposure of members of R_i) and variance σ_i^2 (estimated by the variance of the exposure measurements of members of R_i). In estimating the variance, we use in the denominator the number of individuals in R_i . The test statistic

$$Z = \frac{\left(\sum_{i=1}^n x_i(t_i) - \sum_{i=1}^n \mu_i \right)}{\left[\sum_{i=1}^n \sigma_i^2 \right]^{1/2}} \quad (5)$$

is asymptotically $N(0,1)$. Thus, at each point where a death occurs, we are comparing the cumulative

exposure of the individual with a death at time t_i to the average cumulative exposure of individuals in the risk set.

Breslow *et al.* [38] propose a regression model that can also incorporate external rates. Let $\lambda_i^*(t)$ be the rate for the i th individual in the t th year and assume the relationship

$$\ln \frac{\lambda_i(t)}{\lambda_i^*(t)} = \beta x \quad (6)$$

where x_i is a vector of covariates (i.e., exposure variables and potential confounding variables) and $\lambda_i^*(t)$ is a rate determined by age and the calendar time period. This model is similar to the Cox proportional hazards model except that $\lambda_i^*(t)$ are known rates for the general population. For the special case where there is a single covariate with $x_i(t) = 1$, the maximum-likelihood estimate of β is the total observed deaths divided by the expected deaths. The expected deaths are obtained by applying the external rates to each individual's person-years. Thus, this approach provides a generalization to the regression setting of standard SMR analysis.

Breslow *et al.* [38] also suggest using age as the time variable rather than calendar time. This adjusts for age more precisely than if it were a stratification variable, and it also attenuates some of the problems that can occur if cumulative exposure and time on study are too highly correlated. When age is considered to be the primary time variable, the number of individuals in successive risk sets does not necessarily decrease over time.

Poisson Regression

Poisson regression models the number of deaths occurring in different strata (defined by demographic variables and exposure). They are particularly useful in accommodating grouped data. A summary of Poisson regression models can be found in Breslow and Day [39]. Following their notation, we let

d_{jk} be the observed deaths in the j th age-group and k th exposure group, n_{jk} be the person-years in the j th age-group and k th exposure group, and λ_{jk} be the unknown mortality rate in the j th age-group and k th exposure group.

We assume the d_{jk} are independent Poisson variables with corresponding $E(d_{jk}) = V(d_{jk}) = \lambda_{jk} n_{jk}$.

Assuming a multiplicative model, we obtain

$$\ln \lambda_{jk} = \alpha_j + \beta_k \quad (7)$$

where α_j is the effect associated with the j th age-group ($\alpha_j = \lambda_{j1}$) and β_k is the effect associated with the k th exposure group ($\beta_1 = 0$). Additional variables may be added to the model to incorporate the effect of potential confounders, or product terms can be included to test for effect modification. There is also an additive form of the Poisson regression model that sometimes provides a better fit of the data, but for some data sets may result in convergence problems when iteratively estimating the parameters.

Within the framework of existing software, the Poisson regression model is often easier to implement than the proportional hazards model if there are time-dependent covariates (e.g., exposure). This is because the time-varying aspects of the problem for Poisson regression are incorporated in the computations of person-years.

Hybrid Analysis

Initially, occupational cohort studies were seen as an alternative to case-control studies or PMR studies. Case-control studies are true retrospective studies in that they start with a group having the cause-specific death of interest and an appropriately selected control group, and then compare the previous exposures in the two groups. Case-control studies are subject to more biases than cohort studies some of which relate to the methods of selection of the cases and controls. Also, since a well-defined population of individuals who can have an event is not required, estimates of risk for the individual groups cannot be obtained.

However, an advantage of a case-control study is that it requires less resources than a cohort study and takes less time to complete. In addition, information can be obtained on potential confounders that are not feasible for a large cohort. An important improvement in the design of cohort studies was the incorporation of a case-control study within the cohort, resulting in a study with some of the advantages of each approach. Such a study does not have the same potential for bias in selecting cases and controls that are associated with a classical case-control study, and in addition, direct estimates of risk can be obtained. However, unlike a standard cohort study, adjustment can be made for additional confounders

by obtaining more detailed information on the set of cases and a sample of controls.

Furthermore, since primary outcomes of interest in occupational cohorts often include smoking related cancers (e.g., lung cancer) that could confound the relationship of exposure and disease, incorporating a case-control study within the framework of the cohort is considered by many to be an important aspect of the investigation of mortality patterns of occupational groups.

Nested Case-Control Studies

The concept of a nested case-control study appeared to have been first proposed by Mantel [40], and first applied in an occupational setting by Liddell *et al.* [41] to investigate lung cancer risk in a cohort of asbestos workers. In such a study, a historical prospective study is conducted in the manner previously described. Once the follow-up is completed, the deaths are identified for the cause of interest. A random sample of the controls is selected from the risk set for each case. Often controls are required to have similar birth years, race, and gender. In addition, controls are required to have obtained the same age as the case. When tabulating exposure for the controls, one considers only the exposures acquired up to the time of death of the case. There are additional constraints that are sometimes placed on the selection process. Sometimes, a case is not allowed to serve as its own control, or deaths from other causes that may be related to exposure are excluded. However, it is generally agreed that to avoid bias, controls can include individuals that later become a case [42, 43].

In summary, a nested case-control study has all of the advantages of a historical cohort study and, in addition, permits more detailed information to be collected on potential confounders on the disease cases and a set of randomly selected controls. Thus, the design is superior to conducting a standard cohort study, since it is now feasible (because of the smaller sample size) to obtain more detailed exposure information and/or information on other factors (i.e., smoking, diet, etc.), using interviews and questionnaires.

Case-Cohort Study

An alternative design, which also samples from a cohort study, is the case-cohort design [44]. The

overall concept is the same as the nested case-control study, except that random samples are obtained from the cohort, independent of the health outcome status of the cohort member. Armstrong *et al.* [45] use a case-cohort design to investigate the cancer mortality in a cohort of aluminum reduction plant workers. An advantage of the case-cohort design is that one does not need to wait for the completion of follow-up before starting to obtain the more detailed information required in the case-control component of the study. Various statistical issues that facilitate the proper application of nested case-control studies and case-cohort studies have been addressed in the literature (e.g. [46–49]). These include sample-size estimation, appropriate methods of sampling, utilization of matching, and a comparison of the relative advantages of the two designs.

Example 2 Marsh *et al.* [50] investigated the mortality experience of 32 110 workers employed for at least 1 year from 1945 to 1978 at any of 10 US fiberglass plants. After determining vital status and obtaining death certificates, cause-specific mortality was investigated by computing SMRs using US and local county rates as the control group. Since a primary outcome of interest in the cohort was respiratory system cancer (RSC), a nested matched case-control study was conducted, which enabled the investigators to adjust for the potential confounding effects of smoking. Cases were defined as all male members who died from RSC during 1970–1992, and for each case, one control was randomly selected from all male members at risk at the same age as the corresponding worker who died. Controls were also matched to cases by date of birth (within 1 month), but to avoid overmatching on exposure, controls were not matched with regard to plant. Smoking histories were obtained through structured telephone interviews with the respondent (some controls were alive) or a knowledgeable informant such as a surviving family member. Interviews were completed for 88.3% of the 716 cases, and 80.2% of the 713 controls. Estimated risk ratios were obtained by applying conditional logistic regression to the matched case-control sets. Workers in areas with fiber exposure had a RR of 1.79 ($p = 0.17$) and after adjustment for smoking, this decreased to 1.37 ($p = 0.50$). Investigation of duration of fiber exposure, cumulative exposure, and average intensity of exposure revealed no pattern of increasing risk for RSC mortality for any of the three

exposure measures with or without adjustment for smoking.

Summary

Retrospective cohort studies in occupational groups provide a powerful tool to identify relationships of exposure to chronic diseases. The studies require the existence of historical employment records and usually use cause-specific mortality as the primary outcome. The conduct of these studies requires considerable effort both in obtaining estimates of exposure and in ascertaining the vital status and cause of death of the large number of individuals that typically comprise the cohort. Methods of analysis continue to improve. Earlier analyses relied mostly on the SMR as a summary statistic. Interpretation of the findings of these studies were hampered by limitations of the index, a particular type of bias known as *the healthy worker effect*, and the difficulty of incorporating potential confounders that are not considered in standard population death rates. The development of appropriate regression models has addressed many of the limitations of these early analyses. Furthermore, the importance of evaluating the effect of potential confounders often results in nested case-control studies or case-cohort studies being included as part of the analysis.

References

- [1] Rockette, H.E. & Arena, V.C. (1983). Mortality studies of aluminum reduction plant workers: potroom and carbon department, *Journal of Occupational Medicine* **25**(7), 549–557.
- [2] Boyle, C.A. & Decoufle, P. (1990). National sources of vital status information: extent of coverage and possible selectivity in reporting, *American Journal of Epidemiology* **131**, 160–168.
- [3] Schall, L.C., Marsh, G.M. & Henderson, V.L. (1997). A two-stage protocol for verifying vital status in large historical cohort studies, *Journal of Occupational and Environmental Medicine* **39**, 1097–1102.
- [4] Rich-Edwards, J.W., Corsano, K.A. & Stampfer, M.J. (1994). Test of the national death index and equifax nationwide death search, *American Journal of Epidemiology* **140**, 1016–1019.
- [5] Wentworth, D.N., Neaton, J.D. & Rasmussen, W.L. (1983). An evaluation of the social security administration master beneficiary record file and the national death index in ascertainment of vital status, *American Journal of Public Health* **73**, 1270–1274.

- [6] Smith, T.J., Quinn, M.M., Marsh, G.M., Youk, A.O., Stone, R.A., Buchanich, J.M. & Gula, M.J. (2001). Historical cohort study of US man-made vitreous fiber production workers: VII. Overview of the exposure assessment, *Journal of Occupational and Environmental Medicine* **43**(9), 809–823.
- [7] Benke, G., Sim, M., Fritsch, L. & Aldred, G. (2000). Beyond the job exposure matrix (JEM): the task exposure matrix (TEM), *Annals of Occupational Hygiene* **44**(6), 475–482.
- [8] Bouyer, J. & Hemon, D. (1993). Studying the performance of a job exposure matrix, *International Journal of Epidemiology* **22**(6), 565–571.
- [9] Kauppinen, T., Toikkanen, J. & Pukkala, E. (1998). From cross-tabulations to multipurpose exposure information systems: a new job-exposure matrix, *American Journal of Industrial Medicine* **33**, 409–417.
- [10] Sheineck, G., Plato, N., Alfredsson, L. & Norell, S.E. (1989). Industry-related urothelial carcinogens: applications of a job-exposure matrix to census data, *American Journal of Industrial Medicine* **16**, 209–224.
- [11] Pannett, B., Coggon, D. & Acheson, E.D. (1985). A job-exposure matrix for use in population based studies in England and Wales, *British Journal of Industrial Medicine* **42**, 777–783.
- [12] Luce, D., Gerin, M., Berrino, F., Pisani, P. & Leclerc, A. (1993). Sources of discrepancies between a job exposure matrix and a case-by-case expert assessment for occupational exposure to formaldehyde and wood-dust, *Journal of Epidemiology* **22**(6), S113–S120.
- [13] Wernii, K.J., Astrakianakis, G., Camp, J.E., Ray, R.M., Chang, C.-K., Li, G.D., Thomas, D.B., Checkoway, H. & Seixas, N.S. (2006). Development of a job exposure matrix (JEM) for the textile industry in Shanghai, China, *Journal of Occupational and Environmental Hygiene* **3**, 521–529.
- [14] Marsh, G.M. (2007). Epidemiology of occupational diseases, in *Environmental and Occupational Medicine*, 4th Edition, W.N. Rom, ed, Walters Kluwer, Lippincott William & Wilkins, pp. 32–53.
- [15] Kilpatrick, S.J. (1959). Occupational mortality indices, *Population Studies* **13**, 183–192.
- [16] Bailor, J.C. & Ederer, F. (1964). Significance factors for the ratio of a Poisson variable to its expectation, *Biometrics* **20**, 639–643.
- [17] Silcock, H. (1959). The comparison of occupational mortality rates, *Population Studies* **13**, 183–192.
- [18] Gaffey, W.R. (1976). A critique of the standardized mortality ratio, *Journal of Occupational Medicine* **18**(3), 157–160.
- [19] Gail, M. (1978). The analysis of heterogeneity for indirect standardized mortality ratios, *Journal of Royal Statistical Society, Part 2* **141**, 224–234.
- [20] Arena, V.C., Sussman, N.B., Redmond, C.D., Costantino, J.P. & Trauth, J.M. (1998). Using alternative comparison populations to assess occupation-related mortality risk: results for the high nickel alloys workers cohort, *Journal of Occupational Epidemiology Medicine* **40**(10), 907–916.
- [21] McMichael, A.J. (1976). Standardized mortality ratios and the “healthy worker effect”: scratching beneath the surface, *Journal of Occupational Medicine* **18**(3), 165–168.
- [22] Fox, A.J. & Collier, P.F. (1976). Low mortality rates in industrial cohort studies due to selection for work and survival in the industry, *British Journal of Preventive and Social Medicine* **30**, 225–230.
- [23] Vinni, K. & Hakama, M. (1980). Healthy worker effect in the total Finnish population, *British Journal of Industrial Medicine* **37**, 180–184.
- [24] Mazumdar, S. & Redmond, C.K. (1979). Evaluating dose-response relationships using epidemiological data on occupational subgroups, in *Proceedings of a SIMS Conference on Energy and Health*, Alta, N.E. Breslow & A.S. Whittemore, eds, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, pp. 265–282.
- [25] Arrighi, H.M. & Hertz-Picciotto, I. (1994). The evolving concept of the healthy worker survivor effect, *Epidemiology* **5**, 189–196.
- [26] Robins, J.M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect, *Mathematical Modelling* **7**, 1393–1512.
- [27] Robins, J.M. (1987). Addendum to “a new approach to causal inference in mortality studies with a sustained exposure period – application to control of the health worker survivor effect”, *Computers and Mathematics with Applications* **14**, 923–945.
- [28] Robins, J.M. (1987). Errata for “a new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect”, *Computers and Mathematics with Applications* **14**, 946–953.
- [29] Steenland, K. & Beaumont, J. (1986). A proportionate mortality study of granite cutters, *American Journal of Industrial Medicine* **9**, 1984–1201.
- [30] Checkoway, H., Pearce, N.E. & Crawford-Brown, D.J. (1989). *Research Methods in Occupational Epidemiology*, Oxford University Press, New York.
- [31] DeCoufle, P., Thomas, T.L. & Pickle, L.W. (1980). Comparison of the proportionate mortality ratio and standardized mortality ratio risk measures, *American Journal of Epidemiology* **111**(3), 263–269.
- [32] Wong, O. & Decoufle, P. (1982). Methodological issues involving the standardized mortality ratio and mortality ratio in occupational studies, *Journal of Occupational Medicine* **24**, 299–304.
- [33] Kupper, L.L., McMichael, A.J., Symons, M.J. & Most, B.M. (1978). On the utility of proportional mortality analysis, *Journal of Chronic Diseases* **31**, 15–22.
- [34] Roman, E., Beral, V., Inskip, H. & McDowall, M. (1984). A comparison of standardized and proportional mortality ratios, *Statistics in Medicine* **3**, 7–14.

- [35] Rockette, H.E. & Arena, V.C. (1987). An evaluation of the proportionate mortality index in the presence of multiple comparisons, *Statistics in Medicine* **6**(1), 71–77.
- [36] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.
- [37] Gilbert, E.S. (1976). The assessment of risks from occupational exposures to ionizing radiation, *Environmental Health: Quantitative Methods*, Proceedings of a conference sponsored by SIAM Institute of Mathematics and Society and supported by the National Science Foundation, Alta, Utah, pp. 209–225.
- [38] Breslow, N.E., Lubin, J.H., Marek, P. & Langholz, B. (1983). Multiplicative models and cohort analysis, *Journal of American Statistical Association* **78**(381), 1–12.
- [39] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research: Volume II – the Design and Analysis of Cohort Studies*, IARC Scientific Publications No. 82, IARC, Lyon.
- [40] Mantel, N. (1973). Synthetic retrospective studies and related topics, *Biometrics* **29**, 479–486.
- [41] Liddell, F.D., McDonald, J.C. & Thomas, D.C. (1977). Methods of cohort analysis: appraisal by application to asbestos miners, *Journal of the Royal Statistical Society, Series A* **140**, 469–491.
- [42] Lubin, J.H. & Gail, M.H. (1984). Biased selection of controls for case-control analyses of cohort studies, *Biometrics* **40**, 63–75.
- [43] Robins, J.M., Gail, M.H. & Lubin, J.H. (1986). Biased selection of controls for case-control analysis of cohort studies, *Biometrics* **42**, 293–299.
- [44] Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.
- [45] Armstrong, B., Tremblay, C., Boris, D. & Gilles, T. (1994). Lung cancer mortality and polynuclear aromatic hydrocarbons: a case-cohort study of aluminum production workers in Arvida, Quebec, Canada, *American Journal of Epidemiology* **139**, 250–262.
- [46] Langholz, B. & Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies, *Statistical Science* **11**, 35–53.
- [47] Wacholder, S., Gail, M.H., Pee, D. & Brookmeyer, R. (1989). Alternative variance and efficiency calculations for the case-cohort design, *Biometrika* **76**, 117–123.
- [48] Wacholder, S., Gail, M. & Pee, D. (1991). Selecting an efficient design for assessing exposure-disease relationships in an assembled cohort, *Biometrics* **47**(1), 63–76.
- [49] Wacholder, S., McLaughlin, J.K., Silverman, D.T. & Mandel, J.S. (1992). Selection of controls in case-control studies, I: principles, *American Journal of Epidemiology* **135**, 1029–1041.
- [50] Marsh, G.M., Youk, A.O., Stone, R.A., Buckanich, J.M., Gula, M.J., Smith, T.J. & Quinn, M.M. (2001). Historical cohort study of US man-made vitreous fiber production workers I; 1992 fiberglass cohort follow-up initial findings, *Journal of Occupational and Environmental Medicine* **43**(9), 741–756.

Related Articles

Association Analysis

Causality/Causation

History of Epidemiologic Studies

HOWARD E. ROCKETTE

Occupational Risks

Risk is the probability of an adverse event occurring [1]. Occupational risk, specifically, deals with the probability of injury or illness occurring as a result of hazards within the workplace. In general, there are six types of occupational hazards: safety, chemical, biological, physical, ergonomic, and psychosocial (Table 1) [2]. Some examples of safety hazards include the operation of moving parts or vehicles, functioning in slippery or entangled conditions, or working with explosive materials. In contrast, chemical health hazards include exposures to dusts or fibers (e.g., silica or asbestos), solvents (e.g., benzene or trichloroethylene – see **Benzene**), heavy metals (e.g., lead or mercury), and pesticides (e.g., organochlorines or organophosphates). Biological hazards involve the exposure to pathogens that may be present in the workplace such as human immunodeficiency virus (HIV), hepatitis B, tubercle bacillus, and legionella. Physical hazards that may be encountered in occupational settings include radiation, noise, heat, and temperature. Repetitive motion, awkward posture, or lifting are all examples of ergonomic hazards that can occur in the workplace. Lastly, psychosocial hazards are hazards that can negatively affect the mental health of workers. Unusual or inconsistent shift work can fall into this class of hazards [3]. Table 1 provides examples of hazards that fall within each of these hazard categories along with an example of a potential corresponding adverse health effect.

Industrial hygienists and safety professionals are trained to recognize, evaluate, and control these workplace hazards to ultimately prevent injury and illness. Occupational health practitioners, such as physicians or nurses who specialize in occupational medicine, are trained to recognize the hazards in a patient's workplace that may have resulted in his/her medical diagnosis. Together, these professionals help reduce the probability of the occurrence of workplace illness and injury.

History of Occupational Health

The management and control of occupational risks has a rich history. Occupational risks were first noted in mid-sixteenth-century texts such as Georgius Agricola's, *De Re Metallica (On the Nature of Metals)*.

Bernardino Ramazzini published a text 150 years later, solely devoted to occupational health entitled, *De Morbis Artificum Diatriba (Diseases of Workers)*. Although occupational health has been recognized for centuries, much of modern day state and federal legislation regarding workplace safety and health was brought about in response to several unfortunate events [4]. For example, public awareness of occupational risks in the U.S. peaked after two major events occurred in the twentieth century, the Triangle Shirtwaist Fire and the Hawks Nest Incident (discussed below).

In 1911, a 10-story building in New York City caught fire and killed 146 young women of the Triangle Shirtwaist Company. Overcrowded and poorly supervised working conditions, locked exit doors, inadequate fire exits, and a lack of sprinklers were to blame for the deaths of the garment workers, many of whom died while trying to escape by jumping from the ninth floor windows [5]. This fire helped pave the way to more comprehensive legislation to regulate occupational risk both at the state and federal level. In fact, by 1920, almost all states had some form of limited workplace safety law in place [6]. Furthermore, 2 years after the fire, the US Department of Labor was created to address the protection of the rapidly growing labor force in the United States.

One of the first true tests of the Labor Department was dealing with the Hawk's Nest Incident, beginning around 1930. This project involved the construction of the Hawk's Nest Tunnel near Gauley Bridge, West Virginia as part of a hydroelectric project. The exact death toll is not known, but most accounts report that approximately 500 workers died and over 1500 workers became ill due to silica exposure [7, 8]. Broadcast as front page news and carried in such widely read news journals as *Time*, *Newsweek*, *The Nation*, and *Science*, the Gauley Bridge story, containing details regarding the hazards of silica dust exposures and appropriate methods of protection [9], spread to millions of Americans. Both the Triangle Shirtwaist Fire and the Hawk's Nest Incident created immediate public outrage. As a result, safety inspectors were hired, and many other serious safety and health problems confronting workers were soon identified.

As public awareness and concern over occupational risk continued to grow, the U.S. Congress began to give federal workplace safety and health legislation serious consideration, especially in the late 1960s. At the time, it was estimated that 14 000

Table 1 General examples of occupational hazards and potential adverse effects by type of occupational risk

Risk	Hazard	Potential adverse effect
Safety	Sharp equipment	Lacerations
	Moving machinery	Amputations
	Wet or icy environment	Slips
	Manifolds	Entanglement
	Scaffolding	Falls
	Reactive materials	Explosions
Chemical	Flammable materials	Fires
	Inhalation of excess dusts	Organ damage
	Enzyme exposure	Sensitization
	Inhalation of friable asbestos fibers	Cancer
	Pesticide exposure	Acute response
Biological	Dermal exposure to solvents	Dermatitis
	Needles	HIV
	Mold	Legionella
	Working around or with an infected population	Tubercle bacillus
Physical	Contact with bodily fluids	Hepatitis B
	Ionizing radiation	Cancer
	Noise	Hearing loss
Ergonomic	Repetitive motion	Carpel tunnel syndrome
	Awkward posture	Backaches
	Heavy lifting	Muscle strain
	Hot tools	Skin burns
Psychosocial	Rotating work shifts	Fatigue
	Hostile or harassing environment	Stress

workers died annually from work-related accidents and the financial losses from work-related deaths and accidents were estimated at \$1.5 billion in wages and \$8 billion in the gross national product, annually [10]. On December 29, 1970, President Richard Nixon signed into law the Occupational Safety and Health (OSH) Act, creating the Occupational Safety and Health Administration (OSHA), a regulatory agency, and the National Institute for Occupational Safety and Health (NIOSH), a research institute. The purpose of the act was “to assure so far as possible every working man and woman in the nation safe and healthful working conditions and to preserve our human resources” [11]. The passage of the OSH Act in 1970 was the single most important event in American history for workplace safety and health.

Regulating Occupational Risk

Under the OSH Act, part of OSHA’s responsibility is to regulate occupational risk. OSHA does this

by writing both recommendations and enforceable policies and procedures designed to limit and control occupational hazards. For example, it is against OSHA regulation to allow workers to perform jobs on work surfaces over 6 ft high without having guard rails, a safety net, or a personal fall arrest system in place [12]. The purpose of this regulation is to reduce the risk of worker injury or fatality from falling. Table 2 presents additional examples as to how OSHA works to regulate occupational risks.

OSHA also regulates occupational risk by setting occupational exposure limits (OELs) that they call permissible exposure limits (PELs). PELs represent maximum acceptable exposure levels an employee may be exposed to without incurring a significant risk of adverse health effects. The American Conference of Governmental Industrial Hygienists (ACGIH), a society consisting of mostly governmental and academic industrial hygienists, was the first professional organization to establish OELs in the United States. In 1941, the Threshold Limit Values for Substances

Table 2 Examples of how OSHA regulates different types of occupational risk

Occupational risk	Hazard	OSHA guideline/regulation
Safety	Flammable liquids	29 CFR 1910.106 “Flammable and combustible liquids”
Chemical	Benzene	29 CFR 1910.1028 “Benzene”
Biological	Bloodborne pathogens	29 CFR 1910.1030 “Bloodborne pathogens”
Physical	Noise	29 CFR 1910.95 “Occupational noise exposure” Guidelines such as “Ergonomics for the prevention of musculoskeletal disorders”
Ergonomic	Repetitive motion	Guidelines for “Retail grocery stores”

Committee of ACGIH was formed and charged with investigating, recommending, and annually reviewing OELs for chemicals in the workplace. In 1946, the committee adopted 148 OELs and referred to them as *maximum allowable concentrations*. These were later termed *threshold limit values* (TLVs), and the first publication of the documentation or evidence used to establish the TLVs was made available in 1962 [13]. A similar committee, the MAK Commission, was established in Germany in 1955, which also began publishing recommended OELs [14].

Many OELs and work practices designed to reduce occupational risks were adopted by state and federal governments. For example, by 1954, 19 states had incorporated the TLVs into their state health and safety programs [15, 16]. In addition, the US Department of Labor adopted the 1950 ACGIH TLVs as maximum exposure limits when they published the Safety and Health Standards for Contractors performing federal supply contracts under the Walsh–Healy Public Contracts Act, in 1952 [17]. Furthermore, under Section 6 (a) of the OSH Act, the Secretary of Labor was permitted to promulgate any national consensus standard or established federal standard as an occupational safety or health standard over a 2-year period if it resulted in improved safety or health. As such, OSHA adopted the safety and health standards under the Walsh–Healy Act that included the 1968 ACGIH list of TLVs. The TLVs became enforceable by law as OSHA PELs [18].

OSHA has promulgated many different types of PELs for numerous chemicals and physical agents present in the workplace. The most common PEL is an 8-h time weighted average (TWA) (8-h TWA), which represents a maximum exposure level averaged

over an 8-h work day to which workers may be repeatedly exposed assuming 40 h workweeks. For example, the 8-h TWA OSHA PEL for trichloroethylene is 100 ppm. OSHA also has PELs designed to protect against acute exposures. For these exposures, OSHA has promulgated a short-term exposure level (STEL) that identifies a maximum exposure level that should not be exceeded in a 15-min time period. An example of this is the benzene 5 ppm STEL. Furthermore, OSHA sometimes sets ceiling values that represent values that should not be exceeded under any circumstance. In the case of chlorine gas exposure, OSHA promulgated a ceiling level of 1 ppm.

Cancer Risk

In a landmark article, *Determining an Acceptable Level of Risk* by Travis and Hatemer-Frey [19], they identified that 70% of chemical carcinogens had a postregulatory added lifetime estimated risk greater than 1×10^{-6} and the other 30% had a postregulatory added lifetime estimated risk greater than 1×10^{-4} . From this, it was revealed that among regulated chemicals, some risks in the range of 1×10^{-6} to 1×10^{-3} were acceptable [19].

OSHA PELs are designed to minimize the risk of all types of adverse health effects that may be caused by occupational exposure. However, one of the most sensitive health endpoints that PELs regulate against is cancer. Allowable exposure to carcinogenic hazards varies from chemical to chemical [20]. Although OSHA has used actual data of work-related deaths as benchmarks to develop occupational acceptable risk levels, OSHA has not generally regulated below 1×10^{-3} for occupational carcinogens owing

to technical feasibility. The OSHA PELs regulate workplace hazards to a level where the risk is not necessarily insignificant, but to where the risk is considered “acceptable” in a workplace environment [21].

OSHA has, *de facto*, identified a level of acceptable risk associated with the workplace by defining “safe” as not the equivalent of risk-free, because in some work environments the activities may or may not involve some level of minimal risk [22]. Furthermore, some research has shown that workers are presumed to be willing to accept higher levels of risk, especially when there are significant gains such as salary or benefits [23, 24]. The Supreme Court decision, *Industrial Union Department, AFL-CIO v. American Petroleum Institute et al.* 448 U.S. 607 [1980] suggested that significant occupational risk be determined by comparing the risk in question with other common occupational risks. It was also determined by the Court that an occupational lifetime cancer risk of 1×10^{-3} is significant and, therefore, this decision was extremely influential in defining the acceptable occupation risk for OSHA [22].

Unlike occupational risks, where, as mentioned previously, the level of risk is considered in exchange for some beneficial gain to the worker, environmental exposures and hazards have been regulated U.S. by the US Environmental Protection Agency (EPA) to generally control theoretical risks to between 1×10^{-4} (*de manifestis*) and 1×10^{-6} (*de minimis*) [22]. While occupational and environmental risks have been regulated differently (that is, have different target risk criteria), OSHA has recognized that if the technology was available, certain workplace hazards should be regulated below 1×10^{-3} . Quite similarly, EPA has repeatedly rejected the notion that it can establish a universal, “brightline” acceptable risk that should never be exceeded under any circumstances [22].

Controlling Occupational Risks

To comply with OSHA regulations designed to minimize occupational risks, occupational health professionals, such as industrial hygienists must recommend the implementation of various control measures. The three types of controls utilized, listed in order of effectiveness and often referred to as the *hierarchy of controls* are: engineering, administrative, and personal protection, as illustrated in Table 3. In practice, occupational risks are often minimized using a combination of the three approaches.

Engineering controls are considered the best method of control and should be implemented first because they include methods that systematically eliminate or minimize the degree of exposure. These controls often involve changing the industrial process. For example, eliminating the use of the hazardous material or substituting the material with a less hazardous one are ideal engineering controls. Enclosing a hazardous process so that a worker does not have the potential to be exposed, and the use of local exhaust ventilation are other engineering controls measures. If implementing engineering controls is not technologically or economically feasible, an occupational health professional may have to rely on administrative controls; often considered the second best method.

Administrative controls usually involve a change in procedures. Reducing the time a worker is required to work with or around a hazardous substance or limiting the quantity of hazardous material a worker is required to use are examples of administrative controls. Another example is when one establishes procedures that dictate where and how a worker must sit or stand in his/her work area, in relation to the hazardous material.

Personal protective equipment (PPE) is the third control measure; and it should be used as the technique of last resort only when the implementation

Table 3 Examples of occupational risk controls

Engineering	Administrative	Personal protection
Process elimination	Job placement	Respirators
Material substitution	Worker rotation	Ear muff/plugs
Process automation	Job timing	Gloves/barrier creams
Process isolation	Change in housekeeping procedures	Protective clothing
Enclosure of process	Change in maintenance procedures	Immunizations

of engineering and administrative controls fails to meet minimum standards designed to protect workers. Personal protection in the form of such items as respirators, hard hats, safety shoes, and gloves, etc. can be very effective, although their use may often pose some incremental risk to the worker. Although there is a wide variety of effective PPE manufactured by numerous reputable manufacturers, the appropriate use of such equipment relies heavily on training and proper selection of equipment. Size, fit, and maintenance of PPE is essential to its efficacy, and therefore, proper employee training must be conducted before and periodically during the use of these projects. For these reasons, using PPE as a means of controlling occupational risks is the least reliable.

Conclusion

Occupational risks vary with the nature of the work environment. The economy in the United States has changed very rapidly in the past century. As the economy of a country changes, the structure of the work environment also changes. One specific example of this is how many American workers have experienced the shift in work from production factories to the service sector. With any type of shift in the work environment or work task, there will be a shift in the occupational hazards and risks as well. The most common types of occupation hazards in factories (e.g., severed digits due to unguarded machinery) are not the same as the most common types of occupational hazards that workers are experiencing in the service sector (e.g., exposure to pathogenic agents due to needle pricks). Acknowledging these dissimilarities and others such as those that may come from a change in the demographics of the workforce (e.g., higher percentage of women) or the development of new products resulting in new hazards (e.g., nanomaterials) is crucial to being able to appropriately recognize, evaluate, and control occupational risks.

References

- [1] Rodricks, J.V. (1992). *Calculated Risks*, Cambridge University Press, Cambridge.
- [2] Sathra, S.S. & Rampal, K.G. (1999). *Occupational Health: Risk Assessment and Management*, Blackwell Publishing.
- [3] Levy, B.S. & Wegman, D.H. (2000). Occupational health: an overview, in *Occupational Health: Recognizing and Preventing Work-Related Disease and Injury*, B.S. Levy & D.H. Wegman, eds, Lippincott Williams & Wilkins, Philadelphia.
- [4] Henshaw, J.L., Gaffney, S.H., Madl, A.K. & Paustenbach, D.J. (2007). The employer's responsibility to maintain a safe and healthful work environment: an historical review of societal expectations and industrial practices, *Employee Responsibilities and Rights Journal* 19(3), 173–192.
- [5] Linder, D. (2002). *The Triangle Shirtwaist Factory Fire Trial*, Retrieved 9/6/05, from <http://www.law.umkc.edu/faculty/projects/ftrials/triangle/trianglefire.html>.
- [6] MacLaury, J. (2005). *Government Regulation of Workers' Safety and Health, 1877-1917*, Retrieved 10/6/05, from <http://www.dol.gov/asp/programs/history/monoregsafeintrotoc.htm>.
- [7] *Silicosis Toll from Hawk's Nest may Never be Made Known, Declare Doctor and Lawyer*, The Charleston Daily Mail: 1, 1936.
- [8] U.S. House of Representatives (1936). *Hawks Nest Tunnel*, House of Representatives Subcommittee, vol. 80, no. 5 pp. 1–6.
- [9] Lucas, A. & Paxton, A. (2005). *About the Hawk's Nest Incident – Background for Muriel Rukeyser's "The Book of the Dead"*, Retrieved 10/6/05, from http://www.english.uiuc.edu/maps/poets/m_r/rukeyser/hawksnest.htm.
- [10] Bureau of National Affairs (1970). Background and legislative history, *The Job Safety and Health Act of 1970*, Washington, DC, Chapter 2, pp. 13–21.
- [11] United States 91st Congress (1970). *Occupational Safety and Health Act of 1970*, Public Law 91–596, Government Printing Office, Washington, DC.
- [12] OSHA (1995). *Fall Protection*, U.S. Department of Labor, 29 CFR 1926.501.
- [13] ACGIH (2004). *History of the American Conference of Governmental Industrial Hygienists (ACGIH)*, Retrieved 4/22/05, from <http://www.acgih.org/About/history.htm>.
- [14] Paustenbach, D. & Langner, R. (1986). Corporate occupational exposure limits: the current state of affairs, *American Industrial Hygiene Association Journal* 47(12), 809–818.
- [15] Pabst, A.C. (1954). Industrial hygiene in the petroleum industry, in *42 National Safety Congress: Current Safety Topics in the Petroleum Industry*, National Safety Council, Chicago.
- [16] Sorger, Ed (1958). *Safety Standards for Protection Against Occupationally Acquired Diseases*, Department of Labor and Industries – State of Washington, Olympia.
- [17] Tobin, M.J. (1952). *Safety and Health Standards for Contractors Performing Federal Supply Contracts Under the Walsh-Healey Public Contracts Act*, U.S. Department of Labor, Wage and Hour and Public Contracts Divisions, Washington, DC.
- [18] United States Congress (1971). *Legislative History of the Occupational Safety and Health Act of 1970*, Committee on Labor and Public Welfare, Washington, DC.

- [19] Travis, C.C. & Hattemer-Frey, H. (1988). Determining an acceptable level of risk, *Environmental Science and Technology* **22**(8), 873–876.
- [20] Bingham, E., Cohrssen, B. & Powell, C.H. (eds) (2001). *Patty's Toxicology*, John Wiley & Sons.
- [21] Rodricks, J.V., Brett, S.M. & Wrenn, G.C. (1987). Significant risk decisions in federal regulatory agencies, *Regulatory Toxicology and Pharmacology* **7**(3), 307–320.
- [22] National Research Council (2004). *Review of the Army's Technical Guides on Assessing and Managing Chemical Hazards to Deploy Personnel*, National Academy Press, Washington, DC.
- [23] Starr, C. (1969). Social benefit versus technological risk, *Science* **165**(899), 1232–1238.
- [24] Whipple, C. (1988). Acceptable risk, in *Carcinogen Risk Assessment*, C.C. Travis, ed, Plenum Press, New York, pp. 157–170.

Related Articles

Asbestos

Chernobyl Nuclear Disaster

Radioactive Chemicals/Radioactive Compounds

What are Hazardous Materials?

SHANNON H. GAFFNEY AND JENNIFER D.
ROBERTS

Odds and Odds Ratio

The odds are an alternative way of quantifying chance. Readers are likely to be familiar with the concept from sports betting. In gambling, it is customary to state the chance of an event as the ratio of the number of times that it is expected to happen over the number of times that it is not – the so-called odds in favor of the event. For example, for a horse that previously won 4 races and lost 6, the chance of winning a future race might be stated by an odds of 4 over 6 ($4/6 = 0.67$) for winning or 6 over 4 ($6/4 = 1.5$) in favor of losing. (Odds are a useful currency in betting. This is because, when investing monies in support of or against a bet where the jackpot is divided between the winners, the odds representing the ratio of these monies translate directly into a winner's return per unit investment.)

Importantly, the odds of an event do not represent proportions, and so the odds of it happening and the odds of it not happening do not add up to one. Rather, due to its definition, the odds in favor of an event are the inverse of the odds against the event. Thus, for the horse racing example, while we might estimate that our horse has a $100 - 40 = 60\%$ risk of losing based on knowing that it won 40% of previous races, knowing that the odds in favor of winning are 0.67 would lead us to conclude that the odds for losing are $1/0.67 = 1.5$.

Odds can be compared between two groups by calculating a ratio. The *odds ratio* measures how the odds of an event, typically of getting a disease, varies between two categories, typically a group that is exposed to a risk factor and one that is not. The specific definitions are most easily understood in terms of a 2×2 table of exposure by disease as shown in Table 1. Therein, N denotes the population size and a , b , c , and d the

absolute frequencies of the respective combinations of the levels of the risk factor (exposed or not exposed) and the disease factor (present or absent). The odds of the disease within a subpopulation is then defined by the ratio of the number of subjects that have the disease over the number of those that have not, i.e., $o(\text{disease present}|\text{exposed group}) = a/b$ and $o(\text{disease present}|\text{nonexposed group}) = c/d$. The odds ratio (OR), of disease comparing the exposed with the nonexposed subpopulation is then given by the ratio

$$\text{OR} = \frac{a/b}{c/d} = \frac{a/c}{b/d} \quad (1)$$

For example, an odds ratio of coronary heart disease, comparing obese people with nonobese people, of two would be interpreted as a doubling of the odds of heart disease when putting on weight. Similarly, an odds ratio of contracting measles, comparing vaccinated children with nonvaccinated children, of 0.5 would be interpreted as a halving of the odds of contracting measles.

An appealing property of the odds ratio is that its population value can be estimated from most observational study designs. Specifically, while indices based on risk proportions, for example relative risk (see **Relative Risk**), cannot be estimated directly from *case-control studies* (see **Case-Control Studies**), this is possible for odds ratios. Mathematically, the odds ratio of disease, comparing the exposed with the nonexposed subpopulation, equates to the odds ratio of exposure, comparing cases (subpopulation with disease present) with controls (disease absent), with the latter being estimable from case-control designs. Case-control studies are frequently used in practice, and this explains the widespread use of the odds ratios as a measure of the effect of exposure on disease. In addition, the odds ratio is often used to approximate the relative

Table 1 Two-way classification of exposure by disease

		Disease		Total
		Present	Absent	
Risk factor	Exposed	a	b	$a + b$
	Not exposed	c	d	$c + d$
	Total	$a + c$	$b + d$	$N = a + b + c + d$

risk ratio of disease with this approximation holding when the disease is relatively rare (say less than 5% prevalence).

History of Epidemiologic Studies
Logistic Regression

SABINE LANDAU

Related Articles

Epidemiology as Legal Evidence

Operational Risk Development

Operational risk (*see* **Simulation in Risk Management; Operational Risk Modeling; Near-Miss Management: A Participative Approach to Improving System Reliability; Compliance with Treatment Allocation**) has rapidly moved up the agenda of all banks, largely in response to the positioning of operational risk as a Pillar 1 risk under Basel II (*see* **Credit Risk Models; Solvency; Informational Value of Corporate Issuer Credit Ratings; Risk in Credit Granting and Lending Decisions; Credit Scoring**). This means that for banks implementing Basel II, the capital requirement of the bank will in part be assessed on the basis of a “model” of its operational risks.

There are three approaches allowed under Basel II for assessing the capital a bank must hold against its operational risk:

- the basic indicator approach – fixed % gross income
- the standardized approach – % of eight business lines gross income
- the advanced measurement approach – internal model.

This article focuses on the advanced measurement approach because this is the only “model” that truly reflects a risk-based approach, as it mirrors (in many cases, also adds greater discipline to) the approach taken by the banks’ management themselves.

The decision to include operational risk computation as a contributor to banks’ minimum regulatory capital requirements (a so-called Pillar 1 risk under Basel II) was always likely to be controversial for a number of reasons:

- Conceptually, operational risk is not well defined. Indeed, it is not hard to find risk professionals who feel it is not really a risk discipline at all, and is often thought of as simply a “catch all” description for the residual risk not covered by market or credit risk.
- Operational risk events are normally categorized, for data collection purposes, as either high-frequency low-severity events or low-frequency high-severity events. Data collection under both

descriptors poses both conceptual and practical problems that may be difficult for banks to resolve.

- Data on operational risks is often said to be both unreliable and frequently unavailable.
- Some of the approaches to Operational Risk Measurement, particularly Op. VaR, are criticized for failing to provide any useful insight into the management of operational risk.

Defining Operational Risk

The Basel II definition of operational risk, “The risk of loss resulting from inadequate or failed internal processes, people and systems, or from external events” has a distinctly “catch all” flavor to it.

Whatever the technical definition of risk, anyone who deals with credit or market risk (*see* **Credit Risk Models; Securitization/Life; Default Risk; Compliance with Treatment Allocation**) will have a clear conception of risk and reward as “two sides of the same coin”. For instance, a rise in the dealing position limits within a dealing room will, all other factors remaining the same, lead to an overall greater return, but at the expense of greater volatility of return, and if that volatility produces a particularly bad outcome, the bank will need more capital to survive it. The same concept applies to credit risk; however, operational risk is different.

Efficient Operation of Bank Processes

When we look at operational risk, we observe that in some ways it conforms to many of the characteristics of typical process control issues. For instance, if we look at the efficient operation of a call center, one frequent observation is that inefficient centers are more costly to run (often because of the need to run a parallel set of “repair” processes), produce lower sales, and have a far greater incidence of losses owing to administrative errors. Improvements in the management and control of the operational processes can frequently produce higher returns, owing to lower costs and higher sales being accompanied by a reduced likelihood of incurring administrative losses.

It is this characterization of operational risk that tends to take it away from the discipline of risk management as it is applied to market or credit risk and firmly embed it in the discipline of process management and control. Thus, for management purposes,

this often leads banks to focus their operational risk reporting and management on ensuring that bank processes meet the best international industry standards, such as those covered by relevant International Organization for Standardization (ISO) standards, and by the use of well-trying process management techniques such as “six sigma”.

It also should lead the banking sector to learn from how other industries look at operational risk. For instance, the manufacturing industry is well advanced in its use of real-time signaling of potential problems as happens, say, when heat and vibration sensors on machines feed predictive models of machine breakdown and thus preinform maintenance and repair schedules. While applying such techniques to the banking sector requires research, original thought, and imagination, the potential for operational risk to materialize is surely, in part, increased when banking processes (be they machine driven or manual) “run hot”.

Categorizing Operational Risk Events

If positively defining operational risk is difficult, perhaps recourse to the data can help us characterize events and thus, by observation, develop both a more encompassing definition of operational risk and the data from which to feed appropriate models of operational risk.

An immediate conceptual problem is, however, that operational risk is perhaps not best defined simply in terms of bad outcomes (i.e., losses).

Indeed, by defining operational risk simply in terms of loss events, a bank could actually fail to record important operational risk events. The Basel II definition uses the term “risk of loss”, but this can be misleading, as not all operational risk events lead to banks incurring losses. It is possible for operational risk events to result in a profit, but such events should not be ignored, because they have the potential to result in a loss if they occur again.

Take the example of a bank that has a dealing desk that undertakes foreign exchange transactions. After one deal, the trader mistakenly books that he purchased yen against a sale of US dollars instead of what actually took place, a purchase of US dollars against a sale of yen. This makes it appear to the dealer that he is “long” in yen. To resolve the mismatch position, he decides to sell the yen he thinks he has and purchase dollars.

As a result of the original mistake, the dealer has actually doubled his US dollar mismatch instead of “squaring off” the position. Later in the day, the error is realized, and the trader sells his large US dollar position. Luckily for the trader, the dollar exchange rate has risen against the yen, and so he actually makes a profit.

In this example, the operational risk event (the incorrect recording of a spot foreign exchange transaction) resulted in the bank making a profit, not a loss. It should be recorded as an operational risk event, however, (usually termed a *near miss*) to help the bank improve its processes, because, it may not be so lucky next time. In this particular instance, it is important that the profit is not recorded as a trading profit.

This example, like that in defining operational risk, above, is an example of a high-frequency/low-impact event. These events tend to be readily understood and are viewed by many banks simply as “the cost of doing business”, particularly where the bank is convinced its processes are efficient.

Many financial products, particularly those in retail banking, will include a factor for these kinds of losses within their pricing structure. For example, many banks that offer credit card products adjust their pricing structure to account for a level of fraud they cannot avoid.

Unreliable and Unavailable Data

There is much concern among banks that they have not recorded data on operational risk events. While it must be true that all causes of loss are correctly recorded in the annual accounts (and thus operational risk data must exist), it is nonetheless true

- that such data has often not been recorded other than for accounting purposes and such data is often not suitable for operational risk analysis purposes. For example, operational risk events that are often a problem as profitable events (see the foreign exchange example above) are often systematically underrecorded (would the above profit be recorded as a dealing profit?).
- that individual banks have often suffered too few low-frequency/high-impact events to make any analysis of them meaningful.

In the former case, there are a number of changes to data recording practice that can improve the

situation. One of the world's largest banks has not only reviewed the recording of operational risk events and losses but has also ensured that it has a range of appropriate profit accounts with appropriate procedures to mitigate the systematic underrecording problem. While a certain amount of reconstruction of data is possible, it has to be admitted that for many banks initial operational risk data would be deficient against the standards their supervisors and their own model builders are likely to require.

Modeling Extreme Outcomes

Characterizing operational risk as closely related to process control is clearly necessary but is also clearly not sufficient. While we may observe that the bulk, by number, of bad outcomes in operational risk are so related, we can also see that many of the very worst outcomes, especially those that have led to banking collapses, are not solely explained by breakdowns in processes. In the case of Barings, the US Savings and Loans Crisis, and Long-Term Capital Management (LTCM) (and many more), clearly other factors were at work.

However, it is the low-frequency/high-impact events that are, by far, the most challenging for banks, both in terms of understanding their underlying ("root") causes and in terms of establishing a meaningful base of relevant data. By their very nature, such events are the least understood and the most difficult to predict. Furthermore, they have the potential to do severe damage and may even result in the collapse of the bank.

One problem of low-frequency/high-impact event recording (*see* **Potency Estimation; Extreme Values in Reliability; Near-Miss Management: A Participative Approach to Improving System Reliability**) is that it has "survivor bias" issues in that banks that suffer catastrophic loss events are likely, as a result of the event, to leave the industry. While this problem is well known, it is, nonetheless, very real, and leads to a systematic underestimation of the likelihood of such losses.

To mitigate this issue, and to ensure that banks can use distributions of outcomes that reflect industry experience rather than single bank experience, several data sharing projects have been initiated. This external data is used to perform benchmark analyses, and may also be used to perform scenario and stress analyses. To obtain external data, banks have entered into

data pooling arrangements with other banks. Such arrangements include the following:

- The Operational Riskdata eXchange Association (ORX)
- The British Bankers Association (BBA) Global Operational Loss Database (GOLD).

It is also important that banks use both stress tests and scenario analysis (*see* **Operational Risk Modeling**) to add insight to an understanding of their potential vulnerability to operational risk events. Stress tests, because of the paucity of even industry-wide data are often produced using Monte Carlo simulation (*see* **Environmental Monitoring; Risk Measures and Economic Capital for (Re)insurers; Simulation in Risk Management; Reliability Optimization**) techniques; such techniques, however, often show their greatest weakness when it comes to explaining the model results to senior management. It is very hard for anyone to give due weight in their decision making to events with very low probabilities of occurrence. For example, when traveling, how many of us assess the relative mortality rates for different modes of transport? In practice, most people confine the statistics on the likelihood of their dying as a result of traveling by car, train, ship, or airplane to a meeting to one "box" called *unlikely*. Focusing senior management attention on events with say a 1 : 10 000 likelihood of occurring is never easy, as they suffer from human behavior just like everyone else.

Scenario analysis on the other hand, whatever the probability of the scenarios occurrence, can bring any risk debate alive, especially when the scenario has some "official" backing. The Financial Services Authority of the United Kingdom produces each year their "Financial Risk Outlook", which makes a very useful basis for any banks scenario analysis.

Operational Risk Measurement – Op. VaR

From a technical, statistical perspective, it is possible to concatenate different distributions to produce a distribution that brings together high-frequency/low-impact events and low-frequency/high-impact events. However, such distributions may have limited explanatory power if the resulting distribution has outcomes that are a result of a wide variety of differing causes.

From this, we may conclude that a measure of risk such as value at risk (VaR) when used to measure

operational risk (Op. VaR) may not produce the same level of insight into the management of operational risk that similar measures produce in the case of market risk (M. VaR) or credit risk (C. VaR).

An example of this is that operational losses, by number, in banks will be from high-frequency/low-impact events, individually small, many resulting from “mistakes” in processing. Moreover, there will be a lot of such events, but their overall value will often be surprisingly stable over time, in relation to the number of transactions undertaken by the bank. This data may be extracted from the bank’s sundry profit and loss accounts and are easily incorporated into the price the bank charges its customers for a relevant service. As such, and provided they can evidence to its supervisor that the bank does incorporate the average level of such losses into its product pricing, then these losses will not affect the level of capital the bank must hold.

An Op. VaR measure of these outcomes is likely to produce a “well-behaved” elliptical distribution where “pricing in” the mean is relatively easy to achieve, and, if not “normal”, an analysis of the percentiles will show a high percentage of all outcomes within 2 to 3 standard deviations of the mean.

The incorporation of low-frequency/high-impact events in the distribution that frequently have very different causes will produce a tail of the distribution that needs to be focused on for cause of the obviously extreme outcomes. Indeed, given the size of these extreme outcomes and the contribution of these outcomes to capital requirements, it is this part of the distribution that is of greatest interest to chief risk officers.

This points to another limitation of Op. VaR, for if it is the tail of the distribution of outcomes in operational risk that is most interesting, then one would have strong reservations about using VaR to measure it, as VaR is not a measure of tail risk, but a measure of variance around central tendency. It would seem more logical to start with a measure that looks directly at the tail data, say expected shortfall (ES) (*see **Extreme Value Theory in Finance; Compliance with Treatment Allocation***). As ES also has the feature of being a measure used extensively in the insurance industry to price risks, it should also inform a common framework for banks and insurers to allow banks to insure some of their tail risk through insurance contracts. This contingent capital approach, called so because the payout under the

insurance contract is effectively an injection of free capital available to absorb losses, should grow in the future, as Basel II allows insurance as a risk mitigation technique.

Conclusions

- Only the implementation of an advanced measurement approach to operational risk is likely to add value to a bank’s management of operational risk.
- The measurement and management of high-frequency low-impact operational risk events is best treated as an exercise in process management and control, but there is room for innovative approaches based on methodologies used in other industries.
- Data on operational risk presents problems both in definition and collection that are not commonly met in collecting credit and market risk data. To improve operational risk data requires cooperation between banks and change to internal management and accounting processes.
- Operational value at risk (Op. VaR) may have limited explanatory powers, as outcomes are the result of widely differing causes, and Op. VaR, as a pricing tool, is inferior to measures such as ES, a measure widely used in the insurance industry. As Basel II allows insurance as a risk mitigation technique, this difference in measurement techniques may inhibit the insurance of operational risk.

Further Reading

- Basel II (2005). *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*.
- GARP (2006). *International Certificate in Banking Risk and Regulation Module 1 and Module 2*, ISBN 1-933861-00-2 and -01-0.
- UK Financial Services Authority (2007). *Financial Risk Outlook 2007*, at <http://www.fsa.gov.uk/pubs/plan/financialriskoutlook2007.pdf>.
- Bank for International Settlements (2003). *Operational Risk Management Practices*, at <http://www.bis.org/publ/bcbs96.pdf?noframes=1> (accessed Feb 2003).
- Bank for International Settlements (2006). *Basel II: International Convergence of Capital Measurement and Capital Standards: A revised Framework – Comprehensive Version*, at <http://www.bis.org/publ/bcbs128.pdf?noframes=1> (accessed Jun 2006).

Operational Risk Modeling

Operational Risk Definition and Database Modeling

An operational risk (*see Solvency; Extreme Value Theory in Finance; Near-Miss Management: A Participative Approach to Improving System Reliability*) loss can be defined as a negative impact on the earnings or equity value of the firm owing to an operational risk event. An operational risk event is an incident that happens because of inadequate or failed processes, people or systems, or due to external facts and circumstances; basically it involves any procedural error in processing transactions or any external events (legal suit, terrorism, etc.).

The tracking of internal operational loss data is essential to the entire operational risk management and measuring process. This data must be collected and organized in a time series. The internal loss database is fundamental in the estimation of economic capital to cover against operational risk.

The operational loss data should be collected straight from the profit and loss (P&L) and related systems whenever it is possible and economically feasible. The basis for this logic lies in the fact that when an operational risk loss affects assets or accounts maintained on a mark-to-market basis, the economic impact of the event is usually the same as the accounting impact. In such cases, the basis for measurement is the loss adjustment as recognized in the P&L system.

Internal Loss Data Characteristics

To model operational risk, a financial institution needs to first establish a database with internal historical losses. There are a number of regulatory and technical prerequisites that the firm must follow to attend minimum supervisory and modeling requirements. A bank's internal loss data must be thorough in that it captures all material activities and exposures from all appropriate subsystems and geographic locations; therefore, the internal loss data program covers firmwide operational losses with the participation of every business unit in all the banks' locations

throughout the world. A minimum threshold for data collection could also be established for the collection of internal loss data.

Definition of Gross Loss Amount

The gross operational loss should include the following:

- charges to the P&L and write downs owing to operational risk events;
- market losses owing to operational risk events;
- payments made to third parties for lost use of funds, net of amounts earned on funds held, pending a late payment;
- in cases where a single event causes both positive and negative impacts, the two should be netted. If the net amount is negative and exceeds the reporting threshold, it should be submitted;
- loss is determined at the time of the event, without regard to any eventual remediation. Thus, for example, if a security is bought when a sale was intended, the market value as on the day of the transaction is considered for purposes of calculating the losses, even if the security is held for a period of time until a more favorable market environment develops;
- it is understood that certain losses happen in core functions not directly involved with the P&L, like the IT department, for example. Losses happening in such departments should also be reported as long as they meet the conditions mentioned above;
- operational risk losses that are related to market risk are to be treated as operational risk for the purposes of calculating minimum regulatory capital (e.g., misrepresentation of market positions or P&L caused by operational errors, stop loss violations, losses produced by market valuation models that fail to work properly for any reason, losses taken from market positions taken in excess of limits etc.);
- legal risk is considered operational risk and should be reported. Events related to legal risk usually are disputes that may include litigation in court, arbitration, claims negotiation, etc.
- Some operational risks embedded in credit risk should be reported such as:
 - procedure failures – where processing errors prevent recovery on a loan or actually enable

a loss, as where a cash advance is made on a credit facility that was earlier cancelled by a credit officer;

- legal issues – loan documents that may contain legal errors (invalid clauses or terms etc.);
- scoring models – errors in scoring models might result in the approval of transactions that would otherwise not be admitted.

Exclusions from the Definition of Gross Loss Amounts

The following events should *not* be included in the gross loss:

- internal reworks (cost of repair and/or replacement), overtime, investments, etc.;
- near misses (potential losses that have not materialized);
- costs of general maintenance contracts on equipment, etc. even if used for repairs or replacements in connection with an operational risk event;
- events causing only reputational damage. Reputational risk is *not* considered operational risk. Reputational risk is the risk of any damage to the firm's reputation caused by internal or external factors;
- losses arising from flawed strategic or discretionary processes are *not* recordable, as such losses are considered the result of business or strategic risks. This type of risk is associated often with senior management decision making (e.g., merger and acquisition decisions, regional or global strategy, etc.).

Firms should also collect external loss data, which can be used either by mixing it with the internally collected data points or used to develop scenario analysis and/or stress tests. Self-assessment type of subjective opinion on potential losses is also approved by the regulators although it is not indicated how it should be used in the framework. Firms should also establish and collect business and control environment factors. All these elements (internal data, external data, scenario analysis, and business and control factors) should be used in the measurement framework.

Measuring Operational Risk

Introduction to the Measurement Framework

According to the new Basel Accord established in June 2004, financial institutions should allocate capital against operational risk. They have three options to choose from – the basic indicator, the standardized approach, and the advanced measurement approach (AMA). In the AMA approach, financial institutions may develop their own models to measure and manage operational risk. To establish some ground rules for such models, there are a series of standards that must be followed by the financial institutions to be able to use them as determined by the Basel Accord.

It is possible that a number of models would have to be developed to cope with the different types of data that need to be used by regulatory requirement – the risk measurement (value at risk (VaR) model (*see Equity-Linked Life Insurance; From Basel II to Solvency II – Risk Management in the Insurance Sector; Credit Value at Risk; Compliance with Treatment Allocation*)), the causal model, and the scenario analysis model. Each model covers a different aspect of operational risk, thereby supplementing one another. The overall risk measurement outline might be thought of in terms of capital estimation, in which the following three layers of coverage are usually considered by firms: expected losses, unexpected losses, and catastrophic losses. Expected losses are not considered risks as these are amounts a firm presumes to lose in a given period and, therefore, would not demand capital but would be covered through provisions, or would be embedded in the pricing of the product. Unexpected losses are considered as risk and would demand capital. There are also the high, very infrequent losses (“catastrophic”) that, would not necessarily demand capital because they have a very low probability of occurrence, but, nevertheless, need to be fully understood to be avoided.

In regard to the models used to assess these three levels of estimation, the expected losses can be seen as simply the expected average of the losses for a certain period. The unexpected losses can be estimated by VaR type of models and the “catastrophic” (or extreme) losses can be estimated by scenario analysis or by performing stress tests. The causal model includes key risk indicators to the analysis and tries to find relationships and correlations between these and the losses. This type of analytical framework helps

to identify the indicators that play an influential role in determining the losses.

Operational VaR Model

The VaR type of models, whose development began in the early 1990s, is currently considered as the standard measure for market risk and has even been extended to credit risk measurement. From the viewpoint of market risk, VaR measures the estimated losses that can be expected to be incurred within a certain confidence interval, in the market value of a portfolio until the position can be neutralized. In other words, VaR estimates losses resulting from holding a portfolio for a determined period using the asset prices over the last n days as a measure of the volatility.

A similar logic can be applied in operational risk. However, as the underlying stochastic processes in market and operational risks are different, changes have to be made in the framework developed for the market VaR. In operational risk, two separate stochastic processes that are investigated are as follows: the frequency (*see Experience Feedback; Operational Risk Development*) and the severity (*see Compensation for Loss of Life and Limb; Potency Estimation; Risk Measures and Economic Capital for (Re)insurers*) of the losses. The operational VaR is the aggregation of these processes, i.e., the aggregated loss distribution. Putting it simply, $\text{VaR} = \text{frequency} \times \text{severity}$.

In more formal terms, the aggregated losses at time t given by $X(t) = \sum_{i=1}^{N(t)} U_i$ have the distribution function (where U represents the individual operational losses)

$$F_{I(t)}(x) = \Pr(X(t) \leq x) = \Pr\left(\sum_{i=1}^{N(t)} U_i \leq x\right) \quad (1)$$

The derivation of an explicit formula for $F_{I(t)}(x)$ is, in most cases, impossible. It is usually assumed that the processes $\{N(t)\}$ and $\{U_n\}$ are stochastically independent. Deriving the formula above, we observe the following fundamental relationship:

$$\begin{aligned} \text{VaR}_{OR} &= F_{I(t)}(x) = \Pr(X(t) \leq x) \\ &= \Pr\left(\sum_{t=0}^{\infty} P_t(t) F_U^{\bullet t}(x)\right) \end{aligned} \quad (2)$$

where $F_U^{\bullet t}$ refers to the t th convolution of U with itself, i.e., $F_U^{\bullet t}(x) = \Pr(U_1 + \dots + U_t \leq x)$, the distribution function of the sum of k independent random variables with the same distribution as U .

More practically, the operational VaR model takes a time series of internal loss data as inputs and tries to use this data set to estimate losses, at several confidence intervals, for different periods ahead.

Extreme Value Theory. The typical operational loss database presents a distribution that is not Gaussian. In general, an operational risk database is composed of a few large events and several smaller ones. For risk-management purposes, the interest is to understand the behavior of the tail of the curve (or the most significant losses). One way of dealing with this type of situation is to use certain types of distributions that fit these patterns of events very well. These distributions are consolidated under the extreme value theory (EVT) (*see Individual Risk Models; Mathematics of Risk and Reliability: A Select History; Copulas and Other Measures of Dependency; Compliance with Treatment Allocation*). The application of EVT is used for severity distributions and it only works with the largest events of a database (above a certain threshold). For more details, please refer to [1].

Causal Model (Multifactor Model)

Although the operational VaR model estimates the risk level, it fails to explain the influence that internal factors play in determining these estimates. This is an important feature in risk-management models. Just for comparison, in the market risk VaR, risk analysts can easily verify the impact of any changing factor (e.g., increasing interest rates) in the VaR figures by stress testing the model. Considering the structural differences between market and operational VaR models, this would not be possible in the operational model because the necessary factors are not embedded in the model as they are in the market risk model. To overcome this limitation, an additional model would have to be developed incorporating the use of internal factors (key risk indicators) as inputs to explain which manageable factors are important for determining the operational losses.

One way of dealing with causal models (*see Modifiable Areal Unit Problem (MAUP); Risk from Ionizing Radiation; Uncertainty Analysis*

and Dependence Modeling; Statistical Arbitrage) is to consider the relationship between the dependent and independent variables as linear and apply multifactor models to explain the losses. Basically, such models try to relate losses to a series of key risk indicators suggesting the following relationship:

$$Y_t = \alpha_t + \beta_1 X_{1t} + \dots + \beta_n X_{nt} + \varepsilon_t \quad (3)$$

where Y represents the operational losses in a particular business during a particular period and X represents the key risk indicators. The α 's and β 's are the estimated parameters.

There are several approaches to estimate the parameters of the model, but one of the most popular ones is the ordinary least-squares (OLS) method. In summary, this method basically solves the equation above for ε_t and minimizes

$$\hat{\varepsilon}_t = Y_t - \hat{\alpha}_1 - \hat{\alpha}_2 X_{1t} \quad \text{or} \quad \hat{\varepsilon} = Y_i - \hat{Y}_i \quad (4)$$

The OLS method squares the residuals and minimizes their sum by solving the problem

$$\min \sum \hat{\varepsilon}_t^2 = \sum (Y_t - \hat{\alpha}_1 - \hat{\alpha}_2 X_{1t})^2 \quad (5)$$

The parameters that solve the above equation for the minimum value of the residuals are selected.

The Scenario Analysis Model

Owing to fact that large/catastrophic events, because of their exceptionality, should not happen frequently, historical loss data alone might not be a good indicator of the operational risk level of the firm, at least in the early stages of the measurement process. Considering these facts, the operational VaR model might also have some limitations in performing robust estimation of large/infrequent events. By including scenario analysis (*see Managing Infrastructure Reliability, Safety, and Security; Dynamic Financial Analysis*) as a supplementary model, decisions on both capital and operational risk management can be based on more information than these initially scarce data collected internally. Scenario analysis can provide useful information about a firm's risk exposure that VaR methods can easily miss, particularly if VaR models focus on the "regular" risk rather than the risks associated with rare or extreme events. Such information can then be fed into strategic planning, capital allocation, hedging, and other major decisions.

The measure of operational risk can be seen as the aggregation of severity and frequency of losses over a given time horizon. The severity and frequency of the losses can easily be linked to the evaluation of scenarios, crystallized as potential future events. Their evaluation would involve learning how the frequency and severity would be affected by extreme environments.

To start the modeling process of a scenario analysis, it is necessary to develop a clear procedure for generating a representative set of scenarios that takes into account all relevant risk drivers that might influence its control environment, which ultimately determines the operational risk level. Understanding the relationship of these risk drivers with the frequency, severity, or aggregated operational losses is important in the development of these scenarios as they make it easier to incorporate expert opinions into the model. The risk drivers are then categorized and give rise to scenario classes (e.g., system crashes).

There are quite a few ways to generate these scenarios. Some of them include stressing the parameters or results of the VaR and causal models. In accordance with the operational risk measure, the scenario analysis process must lead to an evaluation of the potential frequency and severity of the financial impact of particular scenarios. The scenario analysis would consider every input available such as expert opinion, key risk indicators, historical internal losses, and any relevant external events. The weighting attributed to each would depend on the quality of the information available. An overview of the scenario analysis model framework would be as in Table 1.

The envisaged scenario analysis model would deal with different ways to generate scenarios from the different inputs and models and help in the validation process.

The scenario analysis process is a complex one and involves more than just quantitative analysis. It implicates also the analysis of shocks and how they would affect a particular business unit or the entire firm.

The scenario analysis process starts by defining the risk profile of a certain business unit and then proceeds by investigating how the business would behave given a series of hypotheses that might be arbitrary or historical (based on events that happened in the past with this firm or another firm), and

Table 1 Scenario analysis framework

Inputs	Scenario generation	Model	Validation
Expert opinion	Scenario analysis model	Estimate loss distributions from expert opinions	Use of internally generated distributions to compare
External data	Scenario analysis model	Analyze sensitivity of external events in the firm Compare external loss distributions with internally generated ones	Use of loss data and expert opinions to compare Use of internally generated distributions to compare (statistical tests may apply)
Key risk indicators	Causal model	Turn explanatory variables stochastic Use upper range of model parameters confidence interval	Quantitative Quantitative
Internal losses	VaR model	Stress model parameters Use of extreme value theory	Quantitative Quantitative

Table 2 Hypothesis tests for scenario analysis and sources of data

Test of hypothesis	Source of data	Note
Historical shocks	Internal experience, external data	The external data available in the databases can be analyzed to estimate their eventual impact on the firm considering the firm's specific control environment. In addition to this, in case the firm experienced large events in a particular type of situation, these can also be used.
Arbitrary shocks	Expert opinions, internal data, KRIs	The arbitrary shocks might be designed on the basis of expert opinions collected from the BSA or even be developed from workshops.
Sensitivity tests	All available	Aim at understanding the changing variance in the correlation between the several KRIs and the loss data, and the eventual extreme losses arising from that. These relationships can be stressed to verify their impact on the P&L.

sensitivity tests. A summary of the hypothesis tests can be seen in Table 2.

The scenario analysis for operational risk involves all data sources, in addition to sensitivity tests that aim at estimating extreme losses that might arise from negative situations. These hypotheses defined by the shocks and the sensitivity tests are also further tested and benchmarked in a second stage.

Framework Integration

The overall framework should be composed of the following three key elements: the database, the models, and the outcomes. The database, as described in the Section titled "Operational Risk Definition and Database Modeling" of this article, is composed not only of the time series of the internal losses and key risk indicators, but also includes expert opinion collected from the business self-assessment process and

external data. These inputs feed the operational VaR, the causal, and the scenario analysis models. Interfaces between the database and the models have to be developed, allowing the users of the models to load data from different databases. These models allow the estimation of economic and regulatory capital, performance of sensitivity and cost/benefit analysis, and performance of a more proactive operational risk management. Among the outcomes of the models are also the official periodic reports to be submitted to the regulators.

Conclusion

Operational risk modeling has evolved significantly since it was first introduced in the mid-1990s. Virtually every financial institution of a reasonable size would have a team or at least someone responsible for the management of operational risk. The

models presented here are just a blueprint of what can really be done to measure and manage this risk. Many other aspects should be taken into consideration like the use of hedge (insurance), correlations—dependence, etc. This is still a young area with quite a lot of potential to develop. Books like [2] describe a number of approaches and techniques that can be used in operational risk.

References

- [1] Cruz, M. (2002). *Modeling, Measuring and Hedging Operational Risk*, John Wiley & Sons, Chichester.
- [2] Cruz, M. (2004). *Operational Risk Modeling and Analysis: Theory and Practice*, Risk Publications, London.

MARCELO CRUZ

Optimal Risk-Sharing and Deductibles in Insurance

Deductibles, in some form or other, are often part of real-world insurance contracts, but most of the classical theory of risk sharing is unable to explain this phenomenon. In this paper, we discuss deductibles from the perspective of economic risk theory and try to point out when optimal contracts contain deductibles.

The model we consider is a special case of the following: consider a group of I agents that could be consumers, investors, reinsurers (syndicated members), insurance buyers etc., having preferences over a suitable set of random prospects. These preferences are assumed to be represented by expected utility (*see Utility Function*), which means that there is a set of utility functions $u_i : R \rightarrow R$, such that X is preferred to Y if and only if $Eu_i(X) \geq Eu_i(Y)$. Here, the random variables X and Y may represent terminal wealth or consumption in the final period. They could also represent insurance risks by a proper reformulation. We assume smooth utility functions with the properties that more is preferred to less, i.e., these functions are strictly increasing, and that the agents are risk averse, i.e., these functions are all concave.

Each agent is endowed with a random payoff X_i called his *initial portfolio*. Uncertainty is objective and external, and there is no informational asymmetry. All parties agree upon the space $(\Omega, \mathcal{F}, P)$ as the probabilistic description of the stochastic environment, the latter being unaffected by their actions. Here, Ω is the set of states of the world, $\mathcal{F}$ is the set of events, formally a σ field generated by the given initial portfolios X_i , and P is the common belief probability measure. For the sake of convenience, it is proposed that both expected values and variances exist for all the initial portfolios, since this is true in real life applications of the theory.

We suppose that the agents can negotiate any affordable contracts among themselves, resulting in a new set of random variables Y_i , representing the possible final payout to the different members of the group, or final portfolios. The transactions are carried out right away at “market prices”, where $\pi(Y)$ represents the market price for any nonnegative random variable Y , and signifies the group’s common valuation of Y relative to the other random variables

having finite variance. The essential objective is then to determine the following:

1. the market price $\pi(Y)$ of any “risk” Y from the set of preferences of the agents and the joint probability distribution of the X_i ’s;
2. the final portfolio Y_i most preferred by each individual, from among those satisfying his budget constraint. Finally, market clearing requires that the sum of the Y_i ’s equals the sum of the X_i ’s, denoted X_M .

The outline of the paper is as follows: In the section titled “Risk Tolerance and Aggregation”, we define key concepts like Pareto optimality and risk tolerance (aversion), and give Borch’s characterization of Pareto optimality. In our subsequent treatment, we more or less take for granted the concepts of a competitive equilibrium, Borch’s theorem, as well as the representative agent construction that leads to the solution of 1 and 2 above. With this as a starting point, in the section titled “The Risk Exchange between an Insurer and a Policy Holder”, the above general formulation is applied to the problem of finding the optimal risk exchange between an insurer and a policy holder. In this section, we also present a simple, motivating example. Deductibles in the optimal contracts are treated in the section titled “Deductibles”. First, we discuss deductibles stemming from no frictions, then those arising from administrative costs, and finally deductibles from moral hazard (*see Moral Hazard*). The final section “Conclusions” is a brief summing up.

A glaring omission in our treatment of deductibles is that of adverse selection. Rothschild and Stiglitz [1] showed in their seminal paper that the presence of high-risk customers caused the low-risk customers to accept a deductible. Since the analysis of this problem deviates quite substantially from the treatment in the rest of the paper, we have omitted it. Adverse selection is, however, another prominent example of asymmetric information that causes deductibles and less than full insurance to be optimal.

Risk Tolerance and Aggregation

We first define what is meant by a Pareto optimum. The concept of *Pareto optimality* offers a minimal and noncontroversial test that any socially optimal

economic outcome should pass. In words, an economic outcome is Pareto optimal if it is impossible to make some individuals better off without making some other individuals worse off.

Karl Borch's characterization of a Pareto optimum $Y = (Y_1, Y_2, \dots, Y_I)$ simply says that there exist positive "agent" weights λ_i such that the marginal utilities at Y of all the agents are equal modulo these constants, i.e.,

$$\begin{aligned} \lambda_1 u'_1(Y_1) &= \lambda_2 u'_1(Y_2) \cdot \dots = \lambda_I u'_1(Y_I) \\ &= u'_\lambda(X_M) \quad a.s. \end{aligned} \quad (1)$$

The function $u'_\lambda(\cdot)$ is interpreted as the marginal utility function of the representative agent. The risk tolerance function of an agent $\rho(x) : R \rightarrow R_+$, is defined by the reciprocal of the absolute risk aversion function $R(x) = -\frac{u''(x)}{u'(x)}$, where the function u is the utility function in the expected utility framework. Consider the following nonlinear differential equation:

$$Y'_i(x) = \frac{R_\lambda(x)}{R_i(Y_i(x))} \quad (2)$$

where $R_\lambda(x) = -\frac{u''_\lambda(x)}{u'_\lambda(x)}$ is the absolute risk aversion function of the representative agent, and $R_i(Y_i(x)) = -\frac{u''_i(Y_i(x))}{u'_i(Y_i(x))}$ is the absolute risk aversion of agent i at the Pareto optimal allocation function $Y_i(x)$, $i \in \mathcal{I}$. There is a neat result connecting the risk tolerances of all the agents in the market to the risk tolerance of the representative agent in a Pareto optimum: *1. The risk tolerance of "the market", or the representative agent, $\rho_\lambda(X_M)$ equals the sum of the risk tolerances of the individual agents in a Pareto optimum, or*

$$\rho_\lambda(X_M) = \sum_{i \in \mathcal{I}} \rho_i(Y_i(X_M)) \quad a.s. \quad (3)$$

2. The real, Pareto optimal allocation functions $Y_i(x) : R \rightarrow R$, $i \in \mathcal{I}$ satisfy the nonlinear differential equations (2).

The result in 1 was first stated by Wilson [2], and later also by Borch [3]. It is, in fact, a direct consequence of Borch's Theorem; see also Bühlmann [4] for the special case of exponential utility functions.

The result has several important consequences, one of which is as follows: suppose there is a risk neutral agent. In a Pareto optimum the entire risk will be borne by this agent.

The Risk Exchange between an Insurer and a Policy Holder

Some of the issues that may arise in finding optimal insurance contracts are, perhaps, best illustrated by an example:

Example 1 A potential insurance buyer can pay a premium αp and thus receive insurance compensation $I(x) = \alpha x$ if the loss is equal to x . He then obtains the expected utility

$$U(\alpha) = E\{u(w - X - \alpha p + I(X))\} \quad (4)$$

where $0 \leq \alpha \leq 1$ is a constant of proportionality. It is easy to show that if $p > EX$, it will *not* be optimal to buy full insurance, i.e., $\alpha^* < 1$ (see e.g., Mossin [5]).

On the other hand, the following approach was suggested by Karl Borch [6]. The situation is the same as in the above, but the premium functional is $p = (\alpha EX + c)$ instead, where c is a nonnegative constant. It follows that here $\alpha^* = 1$, i.e., full insurance is indeed optimal:

$$\begin{aligned} Eu(w - X + I(X)) \\ - EI(X) - c \leq u(w - EX - c) \end{aligned} \quad (5)$$

which follows by Jensen's inequality, in fact for *any* compensation function $I(\cdot)$. Equality is obtained in the above when $I(x) = x$, i.e., full insurance is optimal (if c is small enough such that insurance is preferred to no insurance). In other words, we have just demonstrated that the optimal value of α equals one.

The seeming inconsistency between the solutions to these two problems has caused some confusion in the insurance literature. In the two situations of Example 1, we only considered the problem of the insured. Let us, instead, take the step of including both the insurer and the insurance customer, in which case "optimality" would mean Pareto optimality.

Consider a policy holder having initial capital w_1 , a positive real number, and facing a risk X , a nonnegative random variable. The insured has the utility function u_1 , where monotonicity and strict risk aversion prevails, $u'_1 > 0$, $u''_1 < 0$. The insurer has the utility function u_2 , $u'_2 > 0$, $u''_2 \leq 0$, so that risk neutrality is allowed; initial reserve is given by w_2 , also a positive real number. These parties

can negotiate an insurance contract, stating that the indemnity $I(x)$ is to be paid by the insurer to the insured if claims amount to $x \geq 0$. It seems reasonable to require that $0 \leq I(x) \leq x$ for any $x \geq 0$. Notice that this implies that no payments should be made if there are no claims, i.e., $I(0) = 0$. The premium p for this contract is payable when the contract is initialized.

We recognize that we may employ our established theory for generating Pareto optimal contracts. In doing so, Moffet [7] was the first to show the following:

The Pareto optimal, real indemnity function $I: R_+ \rightarrow R_+$, satisfies the following nonlinear, differential equation

$$\frac{\partial I(x)}{\partial x} = \frac{R_1(A)}{R_1(A) + R_2(B)} \quad (6)$$

where $A = w_1 - p - x + I(x)$ and $B = w_2 + p - I(x)$, and the functions $R_1 = -\frac{u_1''}{u_1'}$, and $R_2 = -\frac{u_2''}{u_2'}$ are the absolute risk aversion functions of the insured and the insurer, respectively.

In our setting, this result is an immediate consequence of the result 2 of the previous section, provided the premium p is taken as a given constant.

From this result, we conclude the following: If $u_2'' < 0$, it follows that $0 < I'(x) < 1$ for all x , and, together with the boundary condition $I(0) = 0$, by the mean value theorem it follows that

$$0 < I(x) < x, \quad \text{for all } x > 0 \quad (7)$$

implying that *full insurance is not Pareto optimal when both parties are strictly risk averse*. We note that the natural restriction $0 \leq I(x) \leq x$ is not binding at the optimum for any $x > 0$, once the initial condition $I(0) = 0$ is employed.

We also observe that *contracts with a deductible d cannot be Pareto optimal when both parties are strictly risk averse*, since such a contract means that $I_d(x) = x - d$ for $x \geq d$, and $I_d(x) = 0$ for $x \leq d$ for $d > 0$, a positive real number. Thus either $I_d' = 1$ or $I_d' = 0$, contradicting $0 < I'(x) < 1$, for all x .

However, when $u_2'' = 0$, we notice that $I(x) = x$, for all $x \geq 0$. *When the insurer is risk neutral, full insurance is optimal and the risk neutral part, the insurer, takes all the risk*. Clearly, when R_2 is uniformly much smaller than R_1 , this will approximately be true even if $R_2 > 0$.

This gives a neat resolution of the ‘‘puzzle’’ in Example 1. We see that the premium p does not really enter the discussion in any crucial manner when it comes to the actual form of the risk-sharing rule $I(x)$, although this function naturally depends on the parameter p .

Deductibles

No Frictions

The usual explanation of deductibles in insurance goes *via* the introduction of some frictions of the neo-classical model. There are a few notable exceptions, one of which is as follows: the common model of insurance takes the insurable asset as given, and then finds the optimal amount of insurance given some behavioral assumptions. To explain the wealth effect in insurance, Aase [8] uses a model where the amount in the insurable asset is also a decision variable. In this situation, solutions can arise where the insurance customer would like to short the insurance contract, i.e., $I^*(x) < 0$ for some losses $x > 0$, for the optimal solution I^* . Since this violates $I^* \in [0, x]$, a deductible naturally arises, since no insurance then becomes optimal at strictly positive losses.

In fact, this seems to be the most common reason why deductibles also arise under frictions; when some constraint becomes binding at the optimum, usually caused by a friction, a deductible may typically arise.

The first case we discuss in some detail is that of administrative costs.

Administrative Costs

The most common explanation of deductibles in insurance is, perhaps, provided by introducing costs in the model. Intuitively, when there are costs incurred from settling claim payments, costs that depend on the compensation and are to be shared between the two parties, the claim size ought to be beyond a certain minimum in order for it to be Pareto optimal to compensate such a claim.

First we note that Arrow [9] has a result where deductibles seem to be a consequence of risk neutrality of the insurer.

Following Raviv [10], let the costs $c(I(x))$ associated with the contract I , and claim size x , satisfy $c(0) = a \geq 0$, $c'(I) \geq 0$ and $c''(I) \geq 0$, for all $I \geq 0$;

in other words, the costs are assumed increasing and convex in the indemnity payment I , and are *ex post*.

Raviv then shows that if the cost of insurance depends on the coverage, then a nontrivial deductible is obtained. Thus Arrow's [9, Theorem 1] deductible result was not a consequence of the risk-neutrality assumption (see also Arrow [11]). On the other hand, it was obtained because of the assumption that insurance cost is proportional to coverage. His theorem is then a direct consequence of the above-cited result. A consequence of this is the following: *If $c(I) = kI$ for some positive constant k , and the insurer is risk neutral, the Pareto optimal policy is given by*

$$I(x) = \begin{cases} 0, & \text{if } x \leq d \\ x - d, & \text{if } x > d \end{cases} \quad (8)$$

where $d > 0$ if and only if $k > 0$.

Here, we obtain full insurance above the deductible. If the insurer is strictly risk averse, a nontrivial deductible would still result if $c'(I) > 0$ for some I , but now there would also be coinsurance (further risk sharing) for losses above the deductible.

Risk aversion, however, is not the only explanation for coinsurance. Even if the insurer is risk neutral, coinsurance might be observed, provided the cost function is a strictly convex function of the coverage I . The intuitive reason for this result is that cost function nonlinearity substitutes for utility function nonlinearity.

A more natural cost function than the one considered above, would, perhaps, be a zero cost if there is no claim, and a jump at zero to $a > 0$ say, if a small claim is incurred. Thus, the overall cost function is $a\chi_{[I>0]} + c(I)$, where $c(0) = 0$, χ_B being the indicator function of B . Then, as long as $I < a$, it is not optimal for the insured to get a compensation, since the cost, through the premium, outweighs the benefits, and a deductible will naturally occur, even if $c' \equiv 0$.

To conclude this section, we note that a strictly positive deductible occurs if the insurance cost depends on the insurance payment, and also if a cost at zero is initiated by a strictly positive claim. Coinsurance above the deductible $d \geq 0$ results from either insurer risk aversion or cost function nonlinearity. For further results on costs and optimal insurance, see e.g., Spaeter and Roger [12].

Moral Hazard

A situation involving moral hazard is characterized by a decision taken by one of the parties involved (the insured), that only he can observe. The other party (the insurer) understands what decision the insured will take, but cannot force the insured to take any particular decision by a contract design, since he cannot monitor the insured.

The concept of moral hazard has its origin in marine insurance. The old standard marine insurance policy of Lloyd's – known as *SG* (ship and goods) policy – covered “physical hazard”, more picturesquely described as “the perils of the sea”. The “moral hazard” was supposed to be excluded, but it seemed difficult to give a precise definition of this concept. Several early writers on marine insurance (e.g., Dover [13], Dinsdale [14] and Winter [15]) indicate that situations of moral hazard were met with underwriters imposing an extra premium.

The idea that follows was initiated by Holmström [16]. To explain this, as in the above, let $u_2(x)$ and w_2 denote the insurer's utility function and initial wealth; $u_1(x)$, w_1 are similarly the corresponding utility function and initial wealth of the insured. In the latter case, the function $v(a)$ denotes *disutility* of level of care (or effort for short) and a the effort designated to minimize or avoid the loss. Only the insured can observe a . The loss facing the insured is denoted by X , having probability density function $f(x, a)$. We consider here, the part of the level of care that is unobservable by the insurer. Many insurance contracts are contingent upon certain actions taken by the insured, actions that can be verified or observed by the insurer. In this instance, we only consider nonobservable level of care.

In doing so, we deviate from the neoclassical assumption that uncertainty is exogenous, since the stochastic environment is now affected by the insured's actions.

The problem may be formulated as follows:

$$\max_{I(x), a, p} Eu_2(w_2 - I(X) + p)$$

where $I(x) \in [0, x]$, $p \geq 0$, subject to

$$Eu_1(w_1 - X + I(X) - p) - v(a) \geq \bar{h} \quad (9)$$

and

$$a \in \operatorname{argmax}_{a'} \{Eu_1(w_1 - X + I(X) - p) - v(a')\}$$

The first constraint is called the *participation Constraint* (individual rationality), and the last one is called the *incentive compatibility constraint*. This latter is the moral hazard constraint, and implies that the customer will use the effort level a that suits him best; this is feasible, since the insurer does not observe a . Technically speaking, the natural requirement $0 \leq I(x) \leq x$ will be binding at the optimum, and this causes a strictly positive deductible to occur. We illustrate this using an example:

Example 2 Consider the case with $u_2(x) = x$, $u_1(x) = \sqrt{x}$, $v(a) = a^2$, and the probability density of claims $f(x, a) = ae^{-ax}$ is exponential with parameter a (effort). Notice that $P(X > x) = e^{-ax}$ decreases as effort a increases: an increase in effort decreases the likelihood of a loss X larger than any given level x .

We consider a numerical example where the initial certain wealth of the insurer $w_1 = 100$, and his alternative expected utility $\bar{h} = 19.60$. This number equals his expected utility without any insurance. In this case, the optimal effort level is $a^* = 0.3719$.

In the case of no moral hazard, we solve the problem without the incentive compatibility constraint, and obtain what is called *the first best solution*. As expected, since the insurer is risk neutral, full insurance is optimal: $I(x) = x$, and the first best level of effort $a^{\text{FB}} = 0.3701$, smaller than that without insurance, which is quite intuitive.

The expected utility of the representative agent may be denoted by the welfare function. It is given by

$$Eu_2(w_2 + p - I(X)) + \lambda(Eu_1(w_1 - p - X + I(X)) - v(a)) = w_2 + 195.83 \quad (10)$$

since $\lambda = \lambda^{\text{FB}} = 9.8631$. Here $p^{\text{FB}} = 2.72$. Moving to the situation with moral hazard, full insurance is no longer optimal. We get a contract with a deductible d , and less than full insurance above the deductible:

$$I(x) = \begin{cases} 0, & \text{if } 0 \leq x \leq d \\ x + p - w_1 + (\lambda + \frac{\mu}{a} - \mu x)^2, & \text{if } d < x \leq \frac{\lambda}{\mu} + \frac{1}{a} \\ x + p - w_1, & \text{if } x > \frac{\lambda}{\mu} + \frac{1}{a} \end{cases} \quad (11)$$

Here the deductible $d = 11.24$, and the second best level of effort $a = a^{\text{SB}} = 0.3681$. We note that this is lower than a^{FB} . The Lagrangian multipliers of the two constraints are $\lambda^{\text{SB}} = 9.7214$ and $\mu = 0.0353$. The second best premium in this case is $p^{\text{SB}} = 0.0147$.

Owing to the presence of moral hazard, there is now a welfare loss; the expected utility of the representative agent has decreased:

$$Eu_2(w_2 + p - I(X)) + \lambda^{\text{SB}}(Eu_1(w_1 - p - X + I(X)) - v(a^{\text{SB}})) = w_2 + 187.73 \quad (12)$$

implying a welfare loss of 8.10 compared to the first best solution. We note that the premium p^{SB} here is lower than that for the full insurance case of the first best solution, owing to a smaller liability, by the endogenous contract design, for the insurer in the situation with moral hazard.

In the above example, we have used *the first-order approach*, justified in Jewitt [17].

Again, we see that deductibles may result, and also coinsurance above the deductible, when the classical model predicts full insurance and no deductible.

A slightly different point of view is the following: if the insured can gain by breaking the insurance contract, moral hazard is present. In such cases, the insurance company will often check that the terms of the insurance contract are being followed. Note that in the above model the insurance company *could not* observe the action a of the insured. This was precisely the cause of the problem in Example 2. Here, checking is possible, but will *cost money*, so it follows that the mere existence of moral hazard will lead to costs. These costs are different from the type of costs in Example 2, but of the same origin – moral hazard. The present situation clearly invites analysis as a two-person game played between the insurance company and its customer.

Borch [18] takes up this challenge, and finds a Nash equilibrium in mixed strategies. In this model, the insured pays the full cost of moral hazard through the premium.

It seems to have been Karl Borch's position that moral hazard ought to be met by imposing an extra premium.

From the above example, we note that increased premiums may not really be the point in dealing

with this problem: here we see that the second best premium p^{SB} is actually smaller than the first best premium p^{FB} . The important issue is that the contract creates incentives for the insured to protect his belongings, since he carries some of the risk himself under the second best risk-sharing rule. This is brought out very clearly in the above example, if the first best solution is implemented when moral hazard is present. Then the insured will set his level of effort $a = 0$, i.e., to its smallest possible level, since he has no incentive to avoid the loss, resulting in a very large loss with high probability (a singular situation with a Dirac distribution at infinity). Since the second best solution is the best when moral hazard is present, naturally this leads to a low profitability, in particular, for the insurer.

Conclusions

Optimal risk sharing was considered from the perspective of Pareto optimality. Relying on the concepts of a competitive equilibrium, Pareto optimality, Borch's Theorem, and the representative agent construction, we discussed some key results regarding risk tolerance and aggregation, which we applied to the problem of optimal risk sharing between an insurer and a customer. The development was motivated by a simple example, illustrating some of the finer points of this theory. It turned out that the concept of a Pareto optimal risk exchange is exactly the right notion of optimality in the present situation.

To explain deductibles, which do not readily follow from the neoclassical theory, we separately introduced (a) the insurable asset as a decision variable, (b) administrative costs, and (c) moral hazard, with the latter also illustrated by an example.

References

- [1] Rothschild, M. & Stiglitz, J. (1976). Equilibrium in competitive insurance markets: an essay in the economics of imperfect information, *Quarterly Journal of Economics* **90**, 629–650.
- [2] Wilson, R. (1968). The theory of syndicates, *Econometrica* **36**, 119–131.
- [3] Borch, K.H. (1985). A theory of insurance premiums, *The Geneva Papers on Risk and Insurance* **10**, 192–208.
- [4] Buhlmann, H. (1980). An economic premium principle, *ASTIN Bulletin* **11**, 52–60.
- [5] Mossin, J. (1968). Aspects of rational insurance purchasing, *Journal of Political Economy* **76**, 553–568.
- [6] Borch, K.H. (1990). *Economics of Insurance*, *Advanced Textbooks in Economics* 29, K.K. Aase & A. Sandmo, eds, North Holland, Amsterdam, New York, Oxford, Tokyo.
- [7] Moffet, D. (1979). The risk sharing problem, *Geneva Papers on Risk and Insurance* **11**, 5–13.
- [8] Aase, K.K. (2007). *Wealth Effects on Demand for Insurance*, working paper, Norwegian School of Economics and Business Administration, Bergen.
- [9] Arrow, K.J. (1970). *Essays in the Theory of Risk-Bearing*, North Holland, Chicago, Amsterdam, London.
- [10] Raviv, A. (1979). The design of an optimal insurance policy, *American Economic Review* **69**, 84–96.
- [11] Arrow, K.J. (1974). Optimal insurance and generalized deductibles, *Skandinavisk Aktuarietidsskrift* **1**, 1–42.
- [12] Spaeter, S. & Roger, P. (1995). *Administrative costs and Optimal Insurance Contracts*, Preprint Université Louis Pasteur, Strasbourg.
- [13] Dover, V. (1957). *A Handbook to Marine Insurance*, 5th Edition, Witherby, London.
- [14] Dinsdale, W.A. (1949). *Elements of Insurance*, Pitman & Sons, London.
- [15] Winter, W.D. (1952). *Marine Insurance: Its Principles and Practice*, 3rd Edition, McGraw-Hill.
- [16] Holmström, B. (1979). Moral hazard and observability, *Bell Journal of Economics* **10**, 74–91.
- [17] Jewitt, I. (1988). Justifying the first-order approach to principal-agent problems, *Econometrica* **56**(5), 1177–1190.
- [18] Borch, K. (1980). The price of moral hazard, *Scandinavian Actuarial Journal* 173–176.

Further Reading

- Aase, K.K. (1990). Stochastic equilibrium and premiums in insurance, *Approche Actuarielle des Risques Financiers, 1^{er} Colloque International AFIR*, Approche Actuarielle des Risques Financiers, Paris, pp. 59–79.
- Aase, K.K. (1992). Dynamic equilibrium and the structure of premiums in a reinsurance market, *The Geneva Papers on Risk and Insurance Theory* **17**(2), 93–136.
- Aase, K.K. (1993). Equilibrium in a reinsurance syndicate; existence, uniqueness and characterization, *ASTIN Bulletin* **22**(2), 185–211.
- Aase, K.K. (1993). Premiums in a dynamic model of a reinsurance market, *Scandinavian Actuarial Journal* **2**, 134–160.
- Aase, K.K. (2002). Perspectives of risk sharing, *Scandinavian Actuarial Journal* **2**, 73–128.
- Aase, K.K. (2004). Optimal risk sharing, in *Encyclopedia of Actuarial Science*, J.L. Teugels & B. Sundt, eds, John Wiley & sons, Chichester, pp. 1212–1215.
- Baton, B. & Lemaire, J. (1981). The core of a reinsurance market, *ASTIN Bulletin* **12**, 57–71.
- Bewley, T. (1972). Existence of equilibria in economies with infinitely many commodities, *Journal of Economic Theory* **4**, 514–540.

- Borch, K.H. (1960). Reciprocal reinsurance treaties, *ASTIN Bulletin* **1**, 170–191.
- Borch, K.H. (1960). Reciprocal reinsurance treaties seen as a two-person cooperative game, *Skandinavisk Aktuarietidskrift* 29–58.
- Borch, K.H. (1960). The safety loading of reinsurance premiums, *Skandinavisk Aktuarietidskrift* 163–184.
- Borch, K.H. (1962). Equilibrium in a reinsurance market, *Econometrica* **30**(3), 442–444.
- DuMouchel, W.H. (1968). The Pareto optimality of an n-company reinsurance treaty, *Skandinavisk Aktuarietidskrift* **51**, 165–170.
- Gerber, H.U. (1978). Pareto-optimal risk exchanges and related decision problems, *ASTIN Bulletin* **10**, 25–33.
- Lemaire, J. (1990). Borch's theorem: a historical survey of applications, in *Risk, Information and Insurance. Essays in the Memory of Karl H. Borch*, H. Loubergé, ed, Kluwer Academic Publishers, Boston, Dordrecht, London.
- Lemaire, J. (2004). Borch's theorem, in *Encyclopedia of Actuarial Science*, J.L. Teugels & B. Sundt, eds, John Wiley & sons, Chichester Vol. 1, pp. 195–200.
- Mas-Colell, A. (1986). The price equilibrium existence problem in topological vector lattices, *Econometrica* **54**, 1039–1054.
- Nash, J.F. (1950). The bargaining problem, *Econometrica* **18**, 155–162.
- Nash, J.F. (1951). Non-cooperative games, *Annals of Mathematics* **54**, 286–295.
- Stiglitz, J.E. (1983). Risk, incentives and insurance: the pure theory of moral hazard, *The Geneva Papers on Risk and Insurance* **26**, 4–33.
- Wyler, E. (1990). Pareto optimal risk exchanges and a system of differential equations: a duality theorem, *ASTIN Bulletin* **20**, 23–32.

KNUT K. AASE

Optimal Stopping and Dynamic Programming

Brief History of Optimal Stopping

Optimal stopping problems originated in Wald's sequential analysis (see **Clinical Dose–Response Assessment**) [1], being a method of statistical inference (sequential probability ratio test) where the number of observations is not determined in advance of the experiment (see pp. 1–4 in the book for a historical account).

Snell [2] formulated a general optimal stopping problem for (discrete-time) stochastic processes, and using the methods suggested in the papers of Wald and Wolfowitz [3] and Arrow *et al.* [4], he characterized the solution by means of the smallest supermartingale (called *Snell's envelope*) dominating the gain (or reward) process. Studies in this direction (often referred to as *martingale methods*) (see **Insurance Pricing/Nonlife; Weather Derivatives; Mathematics of Risk and Reliability: A Select History**) are summarized in [5].

Dynkin [6] formulated a general optimal stopping problem for Markov processes and characterized the solution by means of the smallest superharmonic function dominating the gain function. Dynkin treated the case of discrete time in detail and indicated that the analogous results also hold in the case of continuous time. Studies in this direction (often referred to as *Markovian methods*) (see **Dependent Insurance Risks**) are summarized in [7, 8].

Formulation of Optimal Stopping Problem

Let $G = (G_t)$ be a stochastic process defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$. We interpret G_t as the *gain* (or *loss*) if the observation of G is stopped at time t , and we interpret $\mathcal{F}_t$ as the *information* available up to time t . It is assumed that G is *adapted* to the filtration $(\mathcal{F}_t)$ in the sense that each G_t is $\mathcal{F}_t$ measurable (i.e., the value G_t is known if $\mathcal{F}_t$ is known). Recall that each $\mathcal{F}_t$ is a σ algebra of subsets of Ω such that $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$ for $s \leq t$. Typically, $(\mathcal{F}_t)$ coincides with the natural filtration (meaning that we know $\mathcal{F}_t$ if we know the sample path of G up to time t), but generally, it may also be larger. All our decisions in regard to optimal stopping at time t must be

based on the information $\mathcal{F}_t$ only (no anticipation is allowed).

The following definition formalizes the previous requirement and plays a key role in the study of optimal stopping.

Definition 1 A random variable τ defined on Ω and taking values in the time set is called a *stopping time* if the event $\{\tau \leq t\}$ belongs to $\mathcal{F}_t$ for all t .

Thus τ is a stopping time if (and only if) the information available up to time t is sufficient to determine (decide) whether τ is (or should be) smaller than or equal to t .

The *optimal stopping problem* seeks to solve

$$V = \sup_{0 \leq \tau \leq T} EG_\tau \quad (1)$$

where the supremum is taken over stopping times τ , and E denotes the expectation with respect to the measure P . Note that equation (1) involves two tasks: (a) to evaluate the *value function* V as explicitly as possible; (b) to exhibit an *optimal* stopping time τ_* at which the supremum is attained. The *horizon* T in equation (1) may be either *finite* or *infinite*. In the latter case, the value X_∞ needs to be specified (unless the supremum is taken only over stopping times τ that are finite valued). The most common practice is to set $G_\infty = \limsup_{t \rightarrow \infty} G_t$, but other choices are also possible. It is assumed throughout that $E \sup_t |G_t| < \infty$ so that EG_τ is well defined for every τ , but the former integrability condition can also be relaxed.

To solve the problem (1) in such a generality, one uses *martingale methods* mentioned following [2] above. Recall that a real-valued process $M = (M_t)$ defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ is a *martingale* if $E(M_t | \mathcal{F}_s) = M_s$ for all $s \leq t$, meaning that the predicted value for a future position of the process, given the entire past, equals the present position of the process. Recall also that the process $S = (S_t)$ defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ is a *supermartingale* if $E(S_t | \mathcal{F}_s) \leq S_s$ for all $s \leq t$, meaning that the predicted value for a future position of the process, given the entire past, is smaller than or equal to the present position of the process.

Very often, however, the problem (1) has a more specific structure where

$$G_t = G(X_t) \quad (2)$$

for some *time-homogeneous strong Markov process* $X = (X_t)$ taking values in a state space E and some real-valued (deterministic) function G on E .

Recall that X is a *Markov process* on $(\Omega, \mathcal{F}, (\mathcal{F}_t), P_x)$ if $P_x(X_t \in B | \mathcal{F}_s) = P_{X_s}(X_{t-s} \in B)$ for any measurable subset B of E and any $s \leq t$, meaning that the future states of the process, given the entire past, depend only on the present state of the process. Recall also that X is a *strong Markov process* if the deterministic t in the previous identity can be replaced by any stopping time τ (e.g., the first hitting time of X to a point or a set). Recall finally that X is said to be *time-homogeneous* if $P_x(X_t \in B | \mathcal{F}_s)$ does not depend on s and t but only on $t - s$ when $s \leq t$.

A profound step toward deeper understanding of the resulting problem (1) is to conduct its studies *via* functions of the initial point x of X in E , yielding the extended problem

$$V(x) = \sup_{0 \leq \tau \leq T} E_x G(X_\tau) \quad (3)$$

where E_x denotes the expectation with respect to the measure P_x under which the process X starts at x . This leads to *Markovian methods* mentioned following [6] above.

Solution to Optimal Stopping Problem

1. Consider the optimal stopping problem (1) when the time set $\{0, 1, \dots, N\}$ is discrete and the horizon $T = N$ is finite. Introduce the auxiliary problems

$$V_n = \sup_{n \leq \tau \leq N} E G_\tau \quad (4)$$

for $0 \leq n \leq N$ and note that $V_0 = V$. To solve the problem (1), one can let the time n go backward and proceed recursively as follows:

For $n = N$ we have to stop immediately and our gain S_N equals G_N . For $n = N - 1$ we can either stop or continue. If we stop, our gain S_{N-1} will be equal to G_{N-1} , and if we continue optimally, our gain S_{N-1} will be equal to $E(S_N | \mathcal{F}_{N-1})$. The latter conclusion reflects the fact that our decision about stopping or continuation at time $n = N - 1$ must be based on the information contained in $\mathcal{F}_{N-1}$ only. It follows that if $G_{N-1} \geq E(S_N | \mathcal{F}_{N-1})$, then we need to stop at time $n = N - 1$, and if $G_{N-1} < E(S_N | \mathcal{F}_{N-1})$ then we need to continue at time $n = N - 1$. For $n = N - 2, \dots, 0$, the considerations are continued analogously.

The method of *backward induction* just explained leads to a sequence of random variables $(S_n)_{0 \leq n \leq N}$ defined recursively as follows:

$$S_n = G_N \quad \text{for } n = N \quad (5)$$

$$S_n = \max \{G_n, E(S_{n+1} | \mathcal{F}_n)\} \quad \text{for } n = N - 1, \dots, 0 \quad (6)$$

The method also suggests considering optimality of the stopping time

$$\tau_n = \inf \{n \leq k \leq N : S_k = G_k\} \quad (7)$$

for $0 \leq n \leq N$. Note that the infimum in equation (7) is always attained.

The first part of the following theorem shows that S_n and τ_n solve the problem in a stochastic sense. The second part of the theorem shows that this leads to a solution of the initial problem (1). The third part of the theorem provides a supermartingale characterization of the solution. The method of backward induction and the results presented in the theorem play a central role in the theory of optimal stopping.

Theorem 1 *Consider the optimal stopping problem (4) for $0 \leq n \leq N$. Then we have:*

$$S_n \geq E(G_\tau | \mathcal{F}_n) \quad \text{for all } n \leq \tau \leq N \quad (8)$$

$$S_n = E(G_{\tau_n} | \mathcal{F}_n) \quad \text{for all } 0 \leq n \leq N \quad (9)$$

Moreover, if $0 \leq n \leq N$ is given and fixed, then we have:

The stopping time τ_n from equation (7) is optimal in equation (4) and $V_n = E S_n$ (10)

If τ_ is an optimal stopping time in equation (4),*

$$\text{then } \tau_n \leq \tau_* \text{ } P\text{-a.s.} \quad (11)$$

The sequence $(S_k)_{n \leq k \leq N}$ is the smallest

$$\text{supermartingale that dominates } (G_k)_{n \leq k \leq N} \quad (12)$$

The stopped sequence $(S_{k \wedge \tau_n})_{n \leq k \leq N}$

$$\text{is a martingale} \quad (13)$$

It follows from Theorem 1 that the optimal stopping problem (1) is solved inductively by solving the problems V_n for $n = N, N - 1, \dots, 0$. Moreover, the optimal stopping rule τ_n for V_n satisfies $\tau_n = \tau_k$ on

$\{\tau_n \geq k\}$ for $0 \leq n \leq k \leq N$, where τ_k is the optimal stopping rule for V_k . This, in other words, means that if it was not optimal to stop within the time set $\{n, n+1, \dots, k-1\}$, then the same optimality rule applies in the time set $\{k, k+1, \dots, N\}$. In particular, when specialized to the problem V_0 , the following general principle is obtained: if the stopping rule τ_0 is optimal for V_0 and it was not optimal to stop within the time set $\{0, 1, \dots, n-1\}$, then starting the observation at time n and based on the information $\mathcal{F}_n$, the same stopping rule is still optimal for the problem V_n . This principle of solution for optimal stopping problems has led to the general principle of *dynamic programming* (see **Repair, Inspection, and Replacement Models; Reliability Optimization**) (often referred to as *Bellman's principle*) in the theory of optimal stochastic control (see **Longevity Risk and Life Annuities; Stochastic Control for Insurance Companies; Repair, Inspection, and Replacement Models**) (see [9]).

The method of backward induction, by its nature, requires that the horizon N be finite so that the case of infinite horizon N remains uncovered. It turns out, however, that the random variables S_n defined by the recurrence relations (5) and (6) admit a different characterization that can be directly extended to the case of infinite horizon N . This characterization forms the basis for the second method that will now be discussed. With this aim, note that equations (10) and (11) in Theorem 1 suggest that $S_n = \sup_{n \leq \tau \leq N} E(G_\tau | \mathcal{F}_n)$ for all $0 \leq n \leq N$. A difficulty arises, however, from the fact that both equations (8) and (9) hold only P -a.s. so that the exceptional P -null set may depend on the given τ . Thus, if the supremum is taken over uncountably many τ , then the right-hand side need not define a measurable function, and the identity may fail as well. To overcome this difficulty, it turns out that the concept of *essential supremum* proves useful.

Recall that if $\{Z_\alpha : \alpha \in I\}$ is a family of random variables defined on $(\Omega, \mathcal{G}, P)$ where the index set I can be arbitrary, then there exists a countable subset J of I such that the random variable $Z^* : \Omega \rightarrow \overline{\mathbb{R}}$, defined by $Z^* = \sup_{\alpha \in J} Z_\alpha$ satisfies the following two properties: (a) $P(Z_\alpha \leq Z^*) = 1$ for each $\alpha \in I$ given and fixed; (b) If $\tilde{Z} : \Omega \rightarrow \overline{\mathbb{R}}$ is another random variable satisfying all the previous identities in place of Z^* , then $P(Z^* \leq \tilde{Z}) = 1$. The random variable Z^* is called the *essential supremum* of $\{Z_\alpha : \alpha \in I\}$

relative to P and is denoted by $Z^* = \text{ess sup}_{\alpha \in I} Z_\alpha$. It is determined by the properties (a) and (b) uniquely up to a P -null set.

With the concept of essential supremum, one can then rewrite equations (8) and (9) as follows:

$$S_n = \text{ess sup}_{n \leq \tau \leq N} E(G_\tau | \mathcal{F}_n) \quad (14)$$

for all $0 \leq n \leq N$. This identity provides an additional characterization of the sequence of random variables $(S_n)_{0 \leq n \leq N}$ introduced initially by means of the recurrence relations (5) and (6). Its advantage, in comparison with the recurrence relations, lies in the fact that the identity (14) can naturally be extended to the case of infinite horizon N . The results of this extension are analogous to the results of Theorem 1 (including equations (5) and (6)). [One may observe that the problem of finding S_n in equation (14) is no simpler than the problem of finding V_n in equation (4). The point is that the construction of S_n leads to a construction of the optimal stopping time in equation (7).]

Likewise, in the case of continuous-time set $[0, T]$, one extends equation (14) by setting

$$S_t = \text{ess sup}_{t \leq \tau \leq T} E(G_\tau | \mathcal{F}_t) \quad (15)$$

where the horizon T can be either finite or infinite. If the gain process G is right continuous and left continuous over stopping times, then the process $S = (S_t)$ admits a right-continuous modification and the results of Theorem 1 remain valid without major changes (it is also assumed that the filtration $(\mathcal{F}_t)$ is right continuous and that $\mathcal{F}_0$ contains all P -null sets in $\mathcal{F}$). Note that the analog of equation (12) then states that S is the smallest right-continuous supermartingale that dominates G . The process S is often referred to as the *Snell envelope* of G (both in discrete and continuous time).

2. Consider the optimal stopping problem (3) when the time set $\{0, 1, \dots, N\}$ is discrete and the horizon $T = N$ is finite. Introduce the auxiliary problems

$$V^{N-n}(x) = \sup_{0 \leq \tau \leq N-n} E_x G(X_\tau) \quad (16)$$

for $0 \leq n \leq N$ and note that $V^N(x) = V(x)$ for all x .

Since the problem (16) is equivalent to the problem (4), we can apply the method of backward induction equations (5) and (6) which lead to a

sequence of random variables $(S_n)_{0 \leq n \leq N}$ and a stopping time τ_n defined in equation (7). The key identity linking equation (16) to equation (4) is

$$S_n = V^{N-n}(X_n) \quad (17)$$

for $0 \leq n \leq N$, and the results of Theorem 1 are directly applicable. It follows that

$$\tau_0 = \inf\{0 \leq n \leq N : X_n \in D_n\} \quad (18)$$

where $D_n = \{x \in E : V^{N-n}(x) = G(x)\}$ for $0 \leq n \leq N$. Recall that the transition operator P_1 of X is defined by $P_1 F(x) = E_x F(X_1)$ for $x \in E$, whenever $F : E \rightarrow \mathbb{R}$ is a (bounded) measurable function.

Theorem 2 *Consider the optimal stopping problem (3) when the time set is discrete and the horizon is finite. Then the value function V^N solves the Wald–Bellman equation*

$$V^d(x) = \max\{G(x), P_1 V^{d-1}(x)\} \quad (19)$$

for $d = 1, \dots, N$ and $x \in E$ where $V^0 = G$. Moreover, we have:

The stopping time τ_0 from equation (18) is optimal in equation (3) (20)

If τ_ is an optimal stopping time in equation (3), then $\tau_0 \leq \tau_*$ P_x -a.s. for every $x \in E$* (21)

The sequence $(V^{N-n}(X_n))_{0 \leq n \leq N}$ is the smallest supermartingale that dominates $(G(X_n))_{0 \leq n \leq N}$ under P_x for every $x \in E$ (22)

The stopped sequence $(V^{N-n \wedge \tau_0}(X_{n \wedge \tau_0}))_{0 \leq n \leq N}$ is a martingale under P_x for every $x \in E$ (23)

The Wald–Bellman equation (19) can be written in a more compact form by setting $MF(x) = \max\{G(x), P_1 F(x)\}$ for $x \in E$ and noting that equation (19) reads $V^d(x) = M^d G(x)$ for $d = 0, 1, \dots, N$ where M^d denote the d -th power of M . These recurrence relations form a constructive method for finding $V = V^N$ when $\text{Law}(X_1|P_x)$ is known for $x \in E$.

The results of Theorem 2 extend to the case when the horizon $T = N$ is infinite without major changes. Noting that $N \mapsto V^N$ is increasing, it follows that $\lim_{N \rightarrow \infty} V^N$ exists and equals the

value function V , which solves the Wald–Bellman equation (19) with $V(x)$ in place of $V^d(x)$ and $V^{d-1}(x)$, respectively (when x is given and fixed). Moreover, it can be shown that if $x \mapsto \tilde{V}(x)$ is any (measurable) solution to this equation (satisfying also $E_x \sup_n \tilde{V}(X_n) < \infty$ for all x), then $V = \tilde{V}$ if and only if $\limsup_{n \rightarrow \infty} \tilde{V}(X_n) = \limsup_{n \rightarrow \infty} G(X_n)$ P_x -a.s. for all x . The optimal stopping time τ_0 from equation (18) is defined analogously by $\tau_D = \inf\{n \geq 0 : X_n \in D\}$ where $D = \{x \in E : V(x) = G(x)\}$.

3. To indicate how the results of Theorem 2 extend to the case of continuous-time set $[0, T]$, let us first note that if the horizon T is finite, then one needs to replace the process X_t by the process $Z_t = (t, X_t)$ so that the problem reads

$$V(t, x) = \sup_{0 \leq \tau \leq T-t} E_{t,x} G(t + \tau, X_{t+\tau}) \quad (24)$$

where the “rest of time” $T - t$ changes when the initial state $(t, x) \in [0, T] \times E$ changes in its first variable. It turns out, however, that no argument from the infinite horizon formulation is more seriously affected by this change, and the results will be identical if we simply think of a new process X to be Z , with a new “two-dimensional” state space E equal to $\mathbb{R}_+ \times E$. For these reasons, we will omit any reference to horizon in what follows.

Consider the optimal stopping problem (3) when the time set $[0, T]$ is continuous upon assuming that the following regularity conditions are satisfied: (a) E is a locally compact Hausdorff space with a countable base and the family of measurable sets equals the Borel σ -algebra on E ; (b) the sample paths of X are right continuous and left continuous over stopping times; and (c) the filtration $(\mathcal{F}_t)_{t \geq 0}$ is right continuous (implying that the first entry times to open and closed sets are stopping times) and that $\mathcal{F}_0$ contains all P -null sets from $\mathcal{F}$ (implying also that the first entry times to Borel sets are stopping times). Introduce the continuation set $C = \{x \in E : V(x) > G(x)\}$ and the stopping set $D = \{x \in E : V(x) = G(x)\}$. Define the first entry time τ_D of X into D by setting

$$\tau_D = \inf\{t : X_t \in D\} \quad (25)$$

The following definition plays a fundamental role in solving the optimal stopping problem (3). A

measurable function $F : E \rightarrow \mathbb{R}$ is said to be *superharmonic* if $E_x F(X_\tau) \leq F(x)$ for all stopping times τ and all $x \in E$. If F is lsc (lower semicontinuous) and $(F(X_t))_{t \geq 0}$ is uniformly integrable, then the following stochastic characterization of superharmonic functions is valid (recall equations (12) and (22)): F is superharmonic if and only if $(F(X_t))$ is a right-continuous supermartingale under P_x for every $x \in E$.

Theorem 3 *Consider the optimal stopping problem (3) when the time set is continuous, and suppose that the gain function G is continuous. Then we have:*

The stopping time τ_D from equation (25) is optimal in equation (3) (26)

If τ_ is an optimal stopping time in equation (3), then $\tau_D \leq \tau_*$ P_x -a.s. for every $x \in E$* (27)

The value function V is the smallest superharmonic function that dominates the gain function G on E (28)

The stopped sequence $(V(X_{t \wedge \tau_D}))$ is a martingale under P_x for every $x \in E$ (29)

The results of Theorem 3 remain valid if one assumes that G is finely continuous, and it can also be shown that this hypothesis implies that V is finely continuous. A function F is *finely continuous* (see [10], Chapter XI) if, roughly speaking, the process X sees F as continuous (even though F may be discontinuous). On the other hand, if X is a one-dimensional diffusion process, then V is continuous whenever G is measurable (see [8, p. 157]).

4. Theorem 3 shows that the optimal stopping problem (3) is equivalent to the problem of finding the smallest superharmonic function $\hat{V}$ that dominates G on E . Once $\hat{V}$ is found it follows that $\hat{V} = V$ and τ_D from equation (25) is optimal. This yields the following representation:

$$\hat{V}(x) = E_x G(X_{\tau_D}) \quad (30)$$

for the unknown function $\hat{V}$ and the unknown set D when $x \in E$. Because of the Markovian structure of X , any function of the form (30) is intimately related to a deterministic equation that governs X in mean (parabolic/elliptic *partial differential equations* when

X is continuous, or more general *partial integro-differential equations* when X has jumps).

To describe this connection, recall that the mean-value kinematics of the process X is described by the *characteristic operator* $\mathbb{L}_X$ defined on a function $F : E \rightarrow \mathbb{R}$ as follows:

$$\mathbb{L}_X F(x) = \lim_{U \downarrow x} \frac{E_x F(X_{\tau_{U^c}}) - F(x)}{E_x \tau_{U^c}} \quad (31)$$

where the limit is taken over a family of open sets U shrinking down to x in E . Under rather general conditions, it is possible to establish that the characteristic operator is an extension of the *infinitesimal operator* $\mathbb{L}_X$ defined as in equation (31) with deterministic t in place of random τ_{U^c} and with $t \downarrow 0$ in place of $U \downarrow x$ under the limit sign. Very often, one refers to $\mathbb{L}_X$ as the *infinitesimal generator* of the process X . Its infinitesimal role is uniquely determined through its action on sufficiently regular (smooth) functions F for which both limits in equation (31) exist and coincide. From general theory of Markov processes, one knows that (on a given relatively compact subset of E which is assumed to be $\mathbb{R}^d$) the infinitesimal generator $\mathbb{L}_X$ takes the following integro-differential form (cf. [11]):

$$\begin{aligned} \mathbb{L}_X F(x) = & \lambda(x)F(x) + \sum_{i=1}^d \rho_i(x) \frac{\partial F}{\partial x_i}(x) \\ & + \sum_{i,j=1}^d \sigma_{ij}(x) \frac{\partial^2 F}{\partial x_i \partial x_j}(x) \\ & + \int_{\mathbb{R}^d \setminus \{0\}} \left(F(y) - F(x) - \sum_{i=1}^d (y_i - x_i) \right. \\ & \times \left. \frac{\partial F}{\partial x_i}(x) \right) \nu(x, dy) \end{aligned} \quad (32)$$

where λ corresponds to killing (when negative) or creation (when positive), ρ is the drift coefficient, σ is the diffusion coefficient, and ν is the compensator of the measure μ of jumps of X . By the strong Markov property of X , it follows that any function $\hat{V}$ satisfying equation (30) solves the *Dirichlet problem*

$$\mathbb{L}_X \hat{V} = 0 \quad \text{in } C \quad (33)$$

$$\hat{V}|_D = G \quad (34)$$

Since the optimal stopping set D in this problem is unknown as well (making it a free-boundary

problem), the crucial question becomes to specify additional conditions that will single out the optimal stopping set D among all possible candidates.

5. The previous considerations lead to two traditional ways for finding $\widehat{V}$ when X and G are given: (a) *Wald–Bellman equation* and (b) *free-boundary problems*.

The book [7] provides numerous examples of (a), under various conditions on G and X . For example, it is known that if G is lsc and $E_x \inf_{t \geq 0} G(X_t) > -\infty$ for all $x \in E$, then $\widehat{V}$ can be computed as follows:

$$M_n G(x) := G(x) \vee E_x G(X_{1/2^n}) \quad (35)$$

$$\widehat{V}(x) = \lim_{n \rightarrow \infty} \lim_{N \rightarrow \infty} M_n^N G(x) \quad (36)$$

for $x \in E$, where M_n^N denotes the N -th power of M_n . The method of proof relies upon discretization of the time set $[0, T]$ and making use of discrete-time results reviewed above.

The book [8] presents various examples of (b). The basic idea is that $\widehat{V}$ and D should solve the free-boundary problem

$$\mathbb{L}_X \widehat{V} \leq 0 \quad (\widehat{V} \text{ minimal}) \quad (37)$$

$$\widehat{V} \geq G \quad (\widehat{V} > G \text{ on } C \ \& \ \widehat{V} = G \text{ on } D) \quad (38)$$

where $\mathbb{L}_X$ is the infinitesimal generator of X given by equation (32). Note that the inequality in equation (37) represents an analytic characterization of superharmonic functions.

The following “rule of thumb” specifies the additional conditions referred to following equations (33) and (34). If G is smooth and X after starting at ∂C (the boundary of C) enters $\text{int}(D)$ (the interior of D) immediately (e.g., when X is a diffusion process and ∂C is sufficiently nice), then the condition (37) (under equation (38)) splits into the two conditions

$$\mathbb{L}_X \widehat{V} = 0 \quad \text{in } C \quad (39)$$

$$\frac{\partial \widehat{V}}{\partial x} \Big|_{\partial C} = \frac{\partial G}{\partial x} \Big|_{\partial C} \quad (\text{smooth fit}) \quad (40)$$

On the other hand, if G is continuous and X after starting at ∂C does not enter $\text{int}(D)$ immediately (e.g., when X has jumps and no diffusion component, while ∂C may still be sufficiently nice) then the condition (37) (under equation (38)) splits into the two conditions

$$\mathbb{L}_X \widehat{V} = 0 \quad \text{in } C \quad (41)$$

$$\widehat{V} \Big|_{\partial C} = G \Big|_{\partial C} \quad (\text{continuous fit}) \quad (42)$$

A more precise meaning of these conditions is discussed in [8], Chapter IV.

The resulting free-boundary problems lead to a unique solution $\widehat{V}$ that coincides with the value function V . A powerful method for solving these problems consists in expressing V in terms of ∂C and then deriving a nonlinear (integral) equation for ∂C . Using *local time–space calculus* techniques, it is then possible to show that the latter equation characterizes ∂C uniquely in the class of admissible functions. For further details, see [8], Section 14 and the references therein.

6. The geometric interpretation of superharmonic functions (when X is a Brownian motion process) leads to an appealing physical interpretation of the value function V associated with the gain function G : if G depicts an obstacle, and a rope is put above G with both ends pulled to the ground, the resulting shape of the rope coincides with V (see Figure 1). Clearly, the fit of the rope and the obstacle should be smooth whenever the obstacle is smooth (smooth fit). A similar interpretation (as in Figure 1) extends to dimension two (membrane) and higher dimensions. This leads to a class of problems in mathematical physics called the “*obstacle problems*”.

Another class of problems coming from mathematical physics fits into the free-boundary setting above. These are processes of melting and solidification leading to the “*Stefan free-boundary problem*”. Imagine a chunk of ice (at temperature G) immersed in water (at temperature V). Then the ice–water interface (as a function of time and space) will coincide with the optimal boundary (surface)

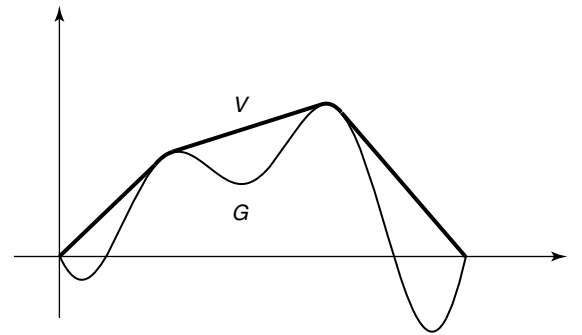


Figure 1 An obstacle G and the rope V depicting the superharmonic characterization

∂C. The thermal energy transfer (heat) from water to ice follows the conservation of energy law and should happen smoothly (smooth fit). This illustrates a basic link between optimal stopping and the Stefan problem.

An Example of Optimal Stopping Problem

The book [8] presents various examples of optimal stopping problems arising from stochastic analysis (Chapter V), mathematical statistics (Chapter VI), mathematical finance (Chapter VII), and financial engineering (Chapter VIII). We conclude this exposition with one example that deals with *optimal prediction* and is closely related to the questions of *risk*.

Example 1 (Optimal prediction of the ultimate maximum) Imagine the real-line movement of a Brownian particle started at 0 during the time interval $[0, 1]$. Let S_1 denote the *maximal* positive height that the particle ever reaches during this time interval. As S_1 is a random quantity whose values depend on the *entire* Brownian path over the time interval, its *ultimate* value is unknown at any given time $t \in [0, 1]$. Following the Brownian particle from the initial time zero onward, the question arises, naturally, of how to determine a time when the movement should be terminated so that the position of the particle at that time is as “close” as possible to the *ultimate maximum* S_1 . In this example, we present the solution to this problem if “closeness” is measured by a mean-square distance.

To formulate the problem above more precisely, let $B = (B_t)_{0 \leq t \leq 1}$ be a standard Brownian motion, and let $(\mathcal{F}_t^B)_{0 \leq t \leq 1}$ denote the natural filtration generated by B . Setting $S_1 = \max_{0 \leq t \leq 1} B_t$, the optimal stopping/prediction problem is to find

$$V = \inf_{0 \leq \tau \leq 1} E(B_\tau - S_1)^2 \quad (43)$$

where the infimum is taken over all stopping times τ of B (i.e., with respect to $(\mathcal{F}_t^B)_{0 \leq t \leq 1}$). Note that the gain process in equation (43) depends on the future values of B as well.

The optimal stopping problem (43) is of interest, for example, in financial engineering, where an optimal decision (i.e., optimal stopping time) should be based on a prediction of the time when the observed process take its maximal value (over a given time interval). The argument also carries over to

many other applied problems where such predictions play a role.

It can be shown (see [12]) that the value V in equation (43) is given by the formula

$$V = 2\Phi(z_*) - 1 = 0.73 \dots \quad (44)$$

where $z_* = 1.12 \dots$ is the unique root to the equation $4\Phi(z_*) - 2z_*\varphi(z_*) - 3 = 0$, and the following stopping time is optimal:

$$\tau_* = \inf \left\{ 0 \leq t \leq 1 : S_t - B_t \geq z_* \sqrt{1-t} \right\} \quad (45)$$

where $S_t = \max_{0 \leq s \leq t} B_s$ for $0 \leq t \leq 1$. Recall that $\varphi(x) = (1/\sqrt{2\pi}) e^{-x^2/2}$ and $\Phi(x) = \int_{-\infty}^x \varphi(y) dy$ denote the density and distribution function of a standard normal variable.

It is interesting to note that

$$V_0 = \inf_{0 \leq t \leq 1} E(B_t - S_1)^2 = \frac{1}{\pi} + \frac{1}{2} = 0.81 \dots \quad (46)$$

with the infimum being attained at $t = 1/2$. Thus, by using the optimal stopping time (instead of an optimal deterministic time), *one makes savings of about 10% on average*.

The following identity holds (see [13]):

$$E|\tau - \theta| = E(B_\tau - B_\theta)^2 - \frac{1}{2} \quad (47)$$

for all stopping times τ of B satisfying $0 \leq \tau \leq 1$, where θ is the (P -a.s. unique) time at which the maximum of B on $[0, 1]$ is attained (i.e., $B_\theta = S_1$).

Taking the infimum on both sides of equation (47) over all stopping times τ of B satisfying $0 \leq \tau \leq 1$, we see that the stopping time τ_* given in equation (45) is optimal for both equation (43) as well as the optimal stopping/prediction problem

$$V_1 = \inf_{0 \leq \tau \leq 1} E|\tau - \theta| \quad (48)$$

where the infimum is taken over all stopping times τ of B . Thus, not only is τ_* the optimal time to stop *as close as possible to the ultimate maximum*, but also τ_* is the optimal time to stop *as close as possible to the time at which the ultimate maximum is attained*. This is indeed the most extraordinary feature of this particular stopping time. For further studies of optimal prediction problems, see [14, 15].

References

- [1] Wald, A. (1947). *Sequential Analysis*, John Wiley & Sons, New York; Chapman & Hall, London.

- [2] Snell, J.L. (1952). Applications of martingale system theorems, *Transactions of the American Mathematical Society* **73**, 293–312.
- [3] Wald, A. & Wolfowitz, J. (1949). Bayes solutions of sequential decision problems, *Proceedings of the National Academy of Sciences of the United States of America* **35**, 99–102; *Annals of Mathematical Statistics* **21**, 82–99, 1950.
- [4] Arrow, K.J., Blackwell, D. & Girshick, M.A. (1949). Bayes and minimax solutions of sequential decision problems, *Econometrica* **17**, 213–244.
- [5] Chow, Y.S., Robbins, H. & Siegmund, D. (1971). *Great Expectations: The Theory of Optimal Stopping*, Houghton Mifflin, Boston.
- [6] Dynkin, E.B. (1963). The optimum choice of the instant for stopping a Markov process, *Soviet Mathematics Doklady* **4**, 627–629.
- [7] Shiryaev, A.N. (1978). *Optimal Stopping Rules*, Springer, New York.
- [8] Peskir, G. & Shiryaev, A.N. (2006). *Optimal Stopping and Free-Boundary Problems, Lectures in Mathematics, ETH Zürich*, Birkhäuser, Basel.
- [9] Bellman, R. (1952). On the theory of dynamic programming, *Proceedings of the National Academy of Sciences of the United States of America* **38**, 716–719.
- [10] Doob, J.L. (1984). *Classical Potential Theory and Its Probabilistic Counterpart*, Springer, New York.
- [11] Skorokhod, A.V. (1967). Homogeneous Markov processes without discontinuities of the second kind, *Theory Probability Applied* **12**, 222–240.
- [12] Graversen, S.E., Peskir, G. & Shiryaev, A.N. (2001). Stopping Brownian motion without anticipation as close as possible to its ultimate maximum, *Theory of Probability and its Applications* **45**, 125–136.
- [13] Urusov, M. (2005). On a property of the moment at which Brownian motion attains its maximum and some optimal stopping problems, *Theory of Probability and its Applications* **49**, 169–176.
- [14] Pedersen, J.L. (2003). Optimal prediction of the ultimate maximum of Brownian motion, *Stochastics and Stochastics Reports* **75**, 205–219.
- [15] Du Toit, J. & Peskir, G. (2007). The trap of complacency in predicting the maximum, *Annals of Probability*, **35**, 340–365.

NICK H. BINGHAM AND GORAN PESKIR

Options and Guarantees in Life Insurance

A life insurance policy is a contract in which the policyholder pays premiums to the insurance company and, in return, the insurance company promises to pay benefits in the event of certain contingencies, for example, death or survival to retirement. The amount of the premiums and benefits may be fixed in advance or they may be variable. A key feature of these contracts is that the payment of the premiums and/or the benefits normally depends on whether the policyholder is alive or not. As the timing and amount of these payments is normally outside of the policyholder's control, the insurance company can take advantage of the law of large numbers through risk pooling in its determination of premiums (i.e., assuming that moral hazard is not present).

Many modern life insurance contracts often incorporate complex design features in the benefit structure that involve options and guarantees. These are of financial significance, and pricing, reserving, and risk management need to take into account the presence of these features.

Life insurance contracts can be classified on the basis of the interest rate offered to the policyholder, the associated guarantees, and the options contained in the benefit. Hence, we distinguish between the following types of contracts.

Unit-linked contract

The interest rate credited to the policyholder's account is linked directly and without time lags to the return on some reference portfolio (whose value is expressed in terms of units). In the case in which the reference portfolio is composed of equities only, the contract is described as equity linked.

Participating contract

The interest rate is calculated according to some smoothing surplus distribution mechanism. Examples include the conventional with profit contract, the accumulating with profit contract (now common in the United Kingdom), and the revalorization contract (common in Europe, excluding the United Kingdom). See Booth *et al.* [1] for more details.

Maturity guarantee

The contract promises to pay at least some absolute amount at maturity (for example, 80% of the initial premium).

Interest rate guarantee

The contract promises to credit the policyholder's account with some minimum return every period.

European-type life insurance contract

The embedded option(s) can be exercised only at maturity, which implies that the policyholder receives the benefit paid by the contract only at the end of its term.

American-type life insurance contract

The policy can, in this case, be terminated (exercised) at the policyholder's discretion (according to his/her circumstances) at any time during the life of the contract. In other words, the policyholder has an American-style option to sell back the policy to the issuing company any time he/she likes. This feature is known as a *surrender option*.

Participating Contracts with Minimum Guarantee

In its basic form, the contract comes into effect at time zero (i.e., the beginning of the first year) with the payment of a single premium, P_0 , from the policyholder to the insurance company (the extension to periodic premium is straightforward, but omitted owing to space constraints). The premiums is invested by the insurer in a fund, S , in the financial market together with the capital provided by the shareholders, E_0 , so that $P_0 = \theta S(0)$, $E_0 = (1 - \theta)S(0)$, where $S(0)$ is the initial value of the reference fund and θ is the cost allocation parameter.

The insurer commits to crediting interest to the policy's account balance according to a scheme linked to the annual returns on the fund until the contract expires. The benefit is payable if the policyholder dies during the policy term of T years, or (at maturity) after T years. In either event, the account is settled by a single payment from the insurance company to the policyholder, unless the policyholder is entitled to terminate the contract at his/her discretion prior to time T , through the surrender option.

This single payment is formed by a guaranteed component, P , which is generated by the cumulated interest credited during the term of the contract, and a discretionary bonus, B .

Guaranteed Benefit

At the beginning of each period, over the lifetime of the contract, the account P accumulates at rate r_P so that

$$\begin{cases} P(t) = P(t-1)(1 + r_P(t)) & t = 1, 2, \dots, T \\ P(0) = P_0 \end{cases} \quad (1)$$

The annual rate r_P at which the interest is credited usually includes a minimum-guaranteed rate, r_G , and a smoothed share of the annual returns generated by the reference fund, r_S . This latter component follows the fluctuations of the financial market; however, the presence of the guarantee imposes a floor limit on the level of the guaranteed benefit. In fact, if β denotes the so-called participation rate, then a common feature of $r_P(t)$ is

$$r_P(t) = \max\{r_G, \beta r_S(t)\}, \quad 0 \leq \beta < 1 \quad (2)$$

which can be rewritten as

$$r_P(t) = r_G + \max\{\beta r_S(t) - r_G, 0\} \quad (3)$$

On the one hand, equation (3) highlights the fact that the benefit is guaranteed never to fall below the contractually specified rate r_G ; hence, as mentioned above, the contract offers a downside protection to the policyholder, which is usually referred to in the literature as the risk-free bond element (see, for example Grosen and Jorgensen [2]). On the other hand, equation (3) also shows the presence of a European call option element in the design of the contract, which affects the risk features of the liabilities generated by the policy. Hence, special care has to be taken when reserving decisions are made.

The specification of the floating rate r_S varies from country to country. For example, in France, the balance P in the policyholder's account grows annually at the guaranteed rate, so that $\beta = 0$ (see, for example [3, 4]). In the United Kingdom, on the other hand, r_S may be calculated as the arithmetic

average of the returns on the reference fund over the last τ years

$$r_S(t) = \frac{1}{n} \left(\frac{S(t)}{S(t-1)} + \dots + \frac{S(t-n+1)}{S(t-n)} - n \right) \quad (4)$$

where n is the length of the smoothing period chosen as $n = \min(t, \tau)$. Alternatively, the floating rate in UK policies may be calculated using the “smoothed asset share” principle: let P^1 denote the “unsmoothed asset share” such that

$$P^1(t) = P^1(t-1)(1 + r_P(t)) \quad (5)$$

$$r_P(t) = \max \left\{ r_G, \beta \frac{S(t) - S(t-1)}{S(t-1)} \right\} \quad (6)$$

then the policy reserve is defined as the average of the value at time t of the unsmoothed asset share with weight α , and the value at time $(t-1)$ of the smoothed asset share, i.e., the policy reserve itself, with weight $(1 - \alpha)$. In other words,

$$P(t) = \alpha P^1(t) + (1 - \alpha)P(t-1) \quad (7)$$

with $\alpha \in (0, 1)$ playing the role of the smoothing parameter (see, for example, Ballotta [5] and Ballotta *et al.* [6] for further details). In Scandinavia, the policy design becomes more complex owing to the presence of a buffer mechanism aimed at stabilizing the annual net profit of the company. For example, in Denmark, insurance companies try to maintain the balance of the bonus account at a fixed, predetermined ratio (described by Grosen and Jorgensen [7] as the target buffer ratio) of the sum of the balances of the guaranteed benefit and the equity capital; consequently, the return on the policy depends on the excess return produced by the business over this target, i.e.,

$$r_S(t) = \frac{B(t-1)}{P(t-1) + E(t-1)} - \xi \quad (8)$$

where ξ is the fixed target buffer ratio.

Discretionary Bonus

Participating contracts usually offer, in addition to the guaranteed benefit, a discretionary bonus that depends on the surplus earned by the reference fund. In the United Kingdom, and in France, this additional benefit is called the *terminal bonus* (see **Value at**

Risk (VaR) and Risk Measures) or *participation bénéficiaire* (see Briys and de Varenne [3] and Ballotta *et al.* [6]); it is paid only at the expiration of the contract, and it is defined as

$$B(T) = \gamma(\theta S(T) - P(T))^+ \quad (9)$$

where γ is the so-called terminal bonus rate and denotes the percentage of the surplus distributed by the insurance company to the policyholders. The terminal bonus is therefore a European call option on the reference portfolio, capturing the fact that the surplus is distributed only when it is positive (i.e., the policyholders do not contribute to possible deficits of the insurance company). The terminal bonus may also be payable if the policyholder dies during the T year term.

In Scandinavia, the discretionary bonus has a form that is different from the above description. The account accumulates (i.e., is distributed) annually, and it provides a buffer mechanism that allows investment surplus to be set aside in years with good investment performance, to be used to cover the annual guarantee in years when the investment return is lower than the guaranteed rate, as described for example in Grosen and Jorgensen [2].

Default Option

The previous sections provided a general description of the overall benefit the participating policyholder is entitled to receive. However, if we assume the absence of an external guarantor, the company's promise can be honored only if the company is solvent at time T , i.e., if the insurer's assets are enough to provide the payment of the guaranteed benefit (that is $S(T) > P(T)$). If, instead, default occurs at maturity and the equityholders have limited liability, the policyholder takes only those assets that are available, while the equityholders are left with nothing. Thus, the liability, L , of the insurance company at maturity can be described as follows:

$$L(T) = \begin{cases} S(T) & \text{if } S(T) < P(T) \\ P(T) & \text{if } P(T) < S(T) < \frac{P(T)}{\theta} \\ P(T) + B(T) & \text{if } S(T) > \frac{P(T)}{\theta} \end{cases} \quad (10)$$

In a more compact way, we can write this as

$$L(T) = P(T) + B(T) - D(T) \quad (11)$$

where

$$D(T) = (P(T) - S(T))^+ \quad (12)$$

represents the payoff of the so-called default option. The default option is a European put option capturing the market loss that the policyholder suffers if a shortfall occurs and, as is described later on, it plays an important role in the valuation process of the insurance contract for the definition of appropriate premiums.

The supervisory authorities might impose more restrictive conditions and monitor the position of the insurance company regularly during the lifetime of the contract. In this case, if the market value of the assets is critically low during the lifetime of the contract, the regulatory authorities might decide to close down the company immediately and distribute the recovered wealth to stakeholders. A possible model for this situation of early default is provided, for example, by Grosen and Jorgensen [7] and Jorgensen [8].

Surrender Option

Insurance policies sometimes give the policyholder the right to terminate the contract prior to maturity and convert the future payments into an immediate cash payment. This feature is known as the *surrender option*, and the immediate cash payment is the so-called surrender value. To avoid financial losses, it is common practice for insurance companies to calculate the surrender value by applying an appropriate discount factor to the cumulated benefit, or mathematical reserve being held, at the time of termination.

The setting of the surrender conditions when designing a new insurance policy is crucial, especially when the financial component of the policy is predominant. The discount factor, in fact, could be fixed in such a way that the cash surrender values are penalizing; however, this might have non desirable effects on the marketability of the contract.

Other Options

Other optional features of life insurance contracts include the right to change the benefits, the premiums or the remaining duration of the policy without necessarily terminating it; the right to cease

premiums, but maintain the policy with an amended benefit level (the so-called paid-up option); the right to begin further insurance contracts without providing evidence of health (the so-called guaranteed insurability option).

Unit-Linked Contracts with Minimum Guarantee

In a unit-linked contract the benefit payable is directly related to the market value of the reference fund. A common feature is to include also a minimum guarantee $G(t)$, so that the benefit payable at time t (on death if $t < T$ or on maturity of $t = T$) is given by

$$P(t) = G(t) + \max [S(t) - G(t), 0] \quad (13)$$

$$= S(t) + \max [G(t) - S(t), 0] \quad (14)$$

which may thus be written as either a call option or a put option on $S(t)$.

Contracts with these features were common in the United Kingdom in the 1970s until the difficulties associated with the provision of these guarantees and in particular, the size of reserves needed were identified [9]. The financial-engineering approach to these embedded options was pioneered by Brennan and Schwartz [10], and Boyle and Schwartz [11]. These contracts remain common in the many markets and a full review of the literature is given by Hardy [12].

Guaranteed Annuity Options

A guaranteed annuity option (GAO) provides the policyholder the right to receive at retirement, at time T , either a cash benefit equal to the current value of the reference portfolio, $S(T)$, or an annuity which would be payable throughout his/her remaining lifetime and which is calculated at a guaranteed rate, g , depending on which has the greater value. This guarantee of the conversion rate between cash and pension income was a common feature of pension policies sold in the United Kingdom during the 1970s and 1980s.

Hence, if the policyholder is aged x_0 at time 0, when the contract is initiated, and N is the normal

retirement age, the GAO pays out at maturity $T = N - x_0$:

$$L(T) = (gS(T)a_{x_0+T} - S(T))^+ \quad (15)$$

$$= gS(T)(a_{x_0+T} - K)^+ \quad (16)$$

where $K = 1/g$ and a_{x_0+T} represents the “annuity factor”, i.e., the expected present value at time T of a life annuity which pays £1 per year throughout the remaining lifetime of the policyholder.

Other Insurance Contracts

Options and guarantees are implicit in many other types of insurance contracts. Thus, long-term contracts providing protection in the event of critical illness (for example, occurrence of a heart attack or incidence of cancer) may include a long-term guarantee regarding the definition of the medical conditions included. More details are provided by Booth *et al.* [1].

Pricing and Reserving

Following the recent developments in terms of accounting standards and capital requirements, insurance companies are required to mark-to-market the liabilities generated by the contracts sold to policyholders (*see Fair Value of Insurance Liabilities*). In the case of contracts containing minimum guarantees, and therefore embedded options, the market consistent pricing can be arrived at by following the theory of contingent claim valuation, and adopting the no-arbitrage principle. This implies that the premiums charged by the insurance company have to be fair when compared to the nominal value of the contract acquired by the policyholder in exchange for those premiums. Hence, for a policy offering a benefit at maturity T , the initial (single) premium P_0 paid by the policyholder has to satisfy the following

$$P_0 = \mathbb{E}^{\mathbb{Q}} [\tilde{L}(T)] \quad (17)$$

where $\mathbb{E}^{\mathbb{Q}}$ denotes the expectation calculated under the risk neutral probability $\mathbb{Q}$ (*see Risk-Neutral Pricing: Importance and Relevance*) and $\tilde{L}$ denotes the value of the overall liability due at maturity, discounted at the risk-free rate of interest. The rhs

of equation (17) represents the nominal market value of the insurance policy originating a liability $L(T)$ at time T ; consequently, if equation (17) is satisfied, the policyholder receives value for the money paid in the form of the actual premium, but at the same time, the insurer is not offering the benefits too cheaply, which would compromise the position of the equityholders.

From the point of view of the calculations, equation (17) can be solved once suitable assumptions are made for the model of the underlying risk drivers, like the reference portfolio and the term structure of interest rates. Further, the valuation framework has to be calibrated to the prevailing market conditions, to estimate the relevant parameters of the model (*see Nonparametric Calibration of Derivatives Models*).

Participating and Unit-Linked Contracts

The application of the market consistent pricing framework to the case of participating contracts with liability of the type described by equation (11) implies that the initial single premium has to satisfy the following condition

$$P_0 = V^P(0) + V^B(0) - V^D(0) \quad (18)$$

where

$$\begin{aligned} V^P(0) &= \mathbb{E}^Q [\tilde{P}(T)]; V^B(0) = \mathbb{E}^Q [\tilde{B}(T)]; \\ V^D(0) &= \mathbb{E}^Q [\tilde{D}(T)] \end{aligned} \quad (19)$$

Equation (18) can be rewritten as

$$P_0 + V^D(0) = V^P(0) + V^B(0) \quad (20)$$

from which it follows that the price of the default option, $V^D(0)$, represents an additional premium that the policyholder should pay to gain an “insurance” against a possible default of the company. In this sense, the default option premium can be interpreted as a safety loading, as discussed by Ballotta *et al.* [13].

A simple pricing formula for this type of life insurance contract is derived by Briys and de Varenne [3] for the case in which $\beta = 0$, as an application of the Black–Scholes option pricing formula. For more complex contract designs, though, the pricing

framework requires the implementation of suitable numerical algorithms, like for example, Monte Carlo simulation (*see Numerical Schemes for Stochastic Differential Equation Models*). The extension of the pricing framework to the case of a jump-diffusion economy (*see Lévy Processes in Asset Pricing*) has been covered by Ballotta [5].

The pricing of the surrender option, however, is a more complex problem. The surrender option is, in fact, an American-style contingent claim whose valuation relies on the application of optional stopping theory (*see Optimal Stopping and Dynamic Programming*). Examples of valuation framework for this part of the contract proposed in the literature make use of either the binomial tree model introduced by Cox *et al.* [14], or the “original” no-arbitrage approach of Black and Scholes [15] extended to incorporate the complex features of insurance contracts. Thus, under the first group of contributions, we mention the backward recursive valuation formulae developed by Bacinello [16] and Grosen and Jørgensen [2] for single premium participating contracts, Bacinello [17] for the case of contracts with annual premium, and Bacinello [18] for the case of unit-linked policies. The no-arbitrage argument has been instead adopted by Jensen *et al.* [19], who propose an approach to reduce the dimensionality of the problem to allow for the development and implementation of a finite difference algorithm for fast and accurate numerical evaluation of the contract. Finally, we cite Steffensen [20], who addresses the valuation problem of the surrender option as an application of the more general framework developed in this paper for intervention options. The value of the surrender part of the contract is hence obtained as the solution to a quasi-variational inequality, which generalizes the classical variational inequality used for the calculation of the price of American options.

A similar approach would be appropriate for unit-linked contracts with guarantees: see Jørgensen [12] for more details.

Guaranteed Annuity Options

The application of the market consistent pricing framework to the case of GAOs requires a stochastic model for the term structure of interest rates. Let r_t be the stochastic short rate of interest. Then, risk neutral valuation implies that the fair value at time T of the

annuity a_{x_0+T} can be calculated as

$$a_{x_0+T} = \mathbb{E}^{\mathbb{Q}} \left[\sum_{j=0}^{w-(T+x_0)} e^{-\int_T^{T_j} r_u du} 1_{(\tau_{x_0+T} > T_j - T)} \middle| \mathcal{F}_T \right] \quad (21)$$

where w is the largest survival age, T_j , $j = 1, \dots, w - (T + x_0)$, are the times of the annuity payments, $\mathcal{F}_T$ is the information flow at maturity and τ_y is a random variable representing the remaining lifetime of a policyholder aged y . Under the assumption of mortality independent of the market risk, it follows from equation (21) that

$$a_{x_0+T} = \sum_{j=0}^{w-(T+x)} {}_j p_{T+x} P(T, T + j) \quad (22)$$

where $P(t, \tau)$ is the price at time t of a zero-coupon bond with redemption date at τ , and ${}_j p_t$ is the survival probability. Equations (16) and (22) imply that the GAO can be seen as a call option written on a coupon-bearing bond, with coupons equal to the survival probabilities, and “coupon dates” $T < T + 1 < \dots < w - x$.

Pricing formulae for this contract have been derived under different market frameworks by Ballotta and Haberman [21], Boyle and Hardy [22], and Pelsser [23], among others. These contributions use fixed mortality tables to calculate the survival probabilities; possible extensions to the case of stochastic mortality have been proposed by Ballotta and Haberman [24] and Biffis and Millosovich [25].

References

- [1] Booth, P., Chadburn, R., Haberman, S. James, D., Khorasane, Z., Plumb, R.H. & Rickayzen B. (2005). *Modern Actuarial Theory and Practice*, 2nd Edition, Chapman & Hall/CRC.
- [2] Grosen, A. & Jorgensen, P.L. (2000). Fair valuation of life insurance liabilities: the impact of interest rate guarantees, surrender options, and bonus policies, *Insurance: Mathematics and Economics* **26**, 37–57.
- [3] Briys, E. & de Varenne, F. (1994). Life insurance in a contingent claim framework: pricing and regulatory implications, *Geneva Papers on Risk and Insurance Theory* **19**, 53–72.
- [4] Briys, E. & de Varenne, F. (1997). On the risk of life insurance liabilities: debunking some common pitfalls, *Journal of Risk and Insurance* **64**, 673–694.
- [5] Ballotta, L. (2005). Lévy process-based framework for the fair valuation of participating life insurance contracts, *Insurance: Mathematics and Economics* **37**, 173–196.
- [6] Ballotta, L., Haberman, S. & Wang, N. (2006). Guarantees in with-profit and unitised with profit life insurance contracts: fair valuation problem in presence of the default option, *Journal of Risk and Insurance* **73**, 97–121.
- [7] Grosen, A. & Jorgensen, P.L. (2002). Life insurance liabilities at market value: an analysis of investment risk, bonus policy and regulatory intervention rules in a barrier option framework, *Journal of Risk and Insurance* **69**, 63–91.
- [8] Jorgensen, P.L. (2001). Life insurance contracts with embedded options: valuation, risk management and regulation, *Journal of Risk Finance* **3**, 19–30.
- [9] Maturity Guarantees Working Party. (1980). The Report, *Journal of the Institute of Actuaries* **107**, 103–209.
- [10] Brennan, M.J. & Schwartz, E.S. (1976). The pricing of equity-linked life insurance policies with an asset value guarantee, *Journal of Financial Economics* **3**, 195–213.
- [11] Boyle, P.P. & Schwartz, E.S. (1977). Equilibrium prices of equity linked insurance policies with an asset value guarantee, *Journal of Risk and Insurance* **44**, 639–660.
- [12] Hardy, M. (2003). *Investment Guarantees: Modelling and Risk Management of Equity Linked Life Insurance*, John Wiley & Sons, New York.
- [13] Ballotta, L., Esposito, G. & Haberman, S. (2006). The IASB insurance project for life insurance contracts: impact on reserving methods and solvency requirements, *Insurance: Mathematics and Economics* **39**, 356–375.
- [14] Cox, J.C., Ross, S.A. & Rubinstein, M. (1979). Option pricing: a simplified approach, *Journal of Financial Economics* **7**, 229–263.
- [15] Black, F. & Scholes, M. (1973). The pricing of options on corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- [16] Bacinello, A.R. (2003). Fair valuation of a guaranteed life insurance participating contract embedding a surrender option, *Journal of Risk and Insurance* **70**, 461–487.
- [17] Bacinello, A.R. (2003). Pricing guaranteed life insurance participating policies with annual premiums and surrender option, *North American Actuarial Journal* **7**, 1–17.
- [18] Bacinello, A.R. (2005). Endogenous model of surrender conditions in equity-linked life insurance, *Insurance: Mathematics and Economics* **37**, 270–293.
- [19] Jensen, B., Grosen, A. & Jorgensen, P.L. (2001). A finite difference approach to the valuation of path dependent life insurance liabilities, *Geneva Papers on Risk and Insurance Theory* **26**, 57–84.
- [20] Steffensen, M. (2002). Intervention options in life insurance, *Insurance: Mathematics and Economics* **31**, 71–85.
- [21] Ballotta, L. & Haberman, S. (2003). Valuation of guaranteed annuity conversion options, *Insurance: Mathematics and Economics* **33**, 87–108.

-
- [22] Boyle, P.P. & Hardy, M.R. (2003). Guaranteed annuity options, *ASTIN Bulletin* **33**, 125–152.
- [23] Pelsser, A. (2003). Pricing and hedging guaranteed annuity options via static option replication, *Insurance: Mathematics and Economics* **33**, 283–296.
- [24] Ballotta, L. & Haberman, S. (2006). The fair valuation problem of guaranteed annuity options: the stochastic mortality case, *Insurance: Mathematics and Economics* **38**, 195–214.
- [25] Biffis, E. & Millossovich, P. (2005). The fair value of guaranteed annuity options, *Scandinavian Actuarial Journal* **1**, 23–41.

LAURA BALLOTTA AND STEVEN HABERMAN

Ordering of Insurance Risk

Comparing and ordering of risks is a basic problem of actuarial theory and practice. Risks are generally modeled by random variables or distribution functions. There is a great variety of stochastic models in use, which reflect the diversity of insurances like theft insurance, car insurance, liability insurance, etc. The diversity of insured population is reproduced in the stochastic models by introducing individual or collective risk models, which include internal, external, and group risk factors. There are risk models with rare extreme events and, on the other hand, models with moderate or even bounded risks. A particular problem in risk theory is dependence on a portfolio of insurance policies that may lead to a drastic increase in the risk of the portfolio.

Ordering of risks gives a guideline to many of the basic tasks of risk theory like measuring of risk or equivalent to the choice and analysis of risk premium principles (see **Risk Classification/Life; Equity-Linked Life Insurance**). It is also a basic tool for estimating the ruin probability (see **Individual Risk Models; Longevity Risk and Life Annuities; Ruin Theory**), the effect of various bonus–malus (see **Bonus–Malus Systems**) and credibility systems (see **Credibility Theory**), the confidence of statistical estimates forecasting the total of claims, etc. The ordering approach is an extension of the classical mean-variance approach of Markovitz to obtain a more specific analysis of essential risk features. Premium principles and risk measures should be consistent with reference to natural risk orderings.

In the following we shall concentrate on two of the most important orderings.

Stochastic Order and the Stop Loss Order

The basic question is, when does a risk X represent a riskier situation than another risk Y ? The answer to this question depends on the attitude toward risk aversion (see **Optimal Risk-Sharing and Deductibles in Insurance; Risk Attitude**). The most simple and obvious postulate is that stochastically larger risks describe more dangerous situations. Here, the

stochastic ordering $X \leq_{\text{st}} Y$ is defined by the postulate that the expectation of all increasing functions is larger for Y than for X , i.e.,

$$Ef(X) \leq Ef(Y) \quad \text{for all } f \in \mathcal{F} \quad (1)$$

where $\mathcal{F} = \mathcal{F}_i$ is the set of increasing functions. Equivalent to condition (1) is that the distribution functions F_X, F_Y of X, Y are comparable in the sense that

$$\begin{aligned} F_X(x) = P(X \leq x) &\geq P(Y \leq x) \\ &= F_Y(x) \end{aligned} \quad (2)$$

for all x . For most of the usual models in insurance it is not difficult to check whether $X \leq_{\text{st}} Y$. In particular, pointwise comparison of risks $X \leq Y$ implies stochastic ordering $X \leq_{\text{st}} Y$ (see Figures 1 and 2).

A common way to describe risk aversion behavior is in terms of utility functions (see **Mathematics of Risk and Reliability: A Select History; Risk Attitude; Clinical Dose–Response Assessment**). Risk X is preferred to risk Y by all risk averse agents (traders, persons) if inequality (equation (1)) holds true for all $f \in \mathcal{F} = \mathcal{F}_{\text{icx}}$, the class of increasing convex functions. In this case we write

$$X \leq_{\text{icx}} Y \quad \text{increasing convex order} \quad (3)$$

Increasing convex order combines the increase in stochastic order with an increase in the diffusiveness of risks. It is equivalently described by the comparison of the expectation of the angle (call) functions. For all real a

$$E(X - a)_+ \leq E(Y - a)_+ \quad (4)$$

i.e., the expected stop loss (see **Premium Calculation and Insurance Pricing; Risk Measures and Economic Capital for (Re)insurers**) risks are bigger for Y than for X . Therefore, stop loss contracts for the risk Y should have a higher premium than those for X . Equation (4) defines the *stop loss ordering*

$$X \leq_{\text{sl}} Y \quad (5)$$

The equivalence to the increasing convex order $\leq_{\text{icx}}$ is a basic justification for considering the stop loss order. For risks X, Y with the same expectation $EX = EY$ a sufficient condition for the stop loss

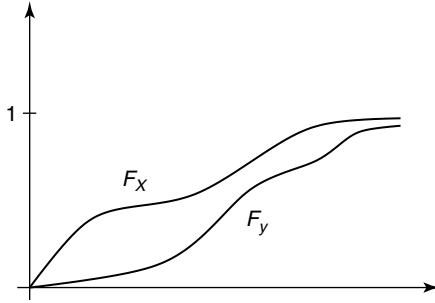


Figure 1 Stochastic ordering

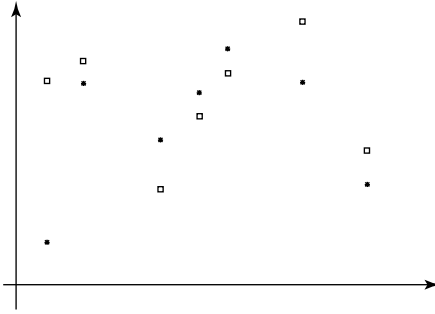


Figure 2 Typical sample where $X \leq_{st} Y$; $\square : Y$; $\cdot : X$

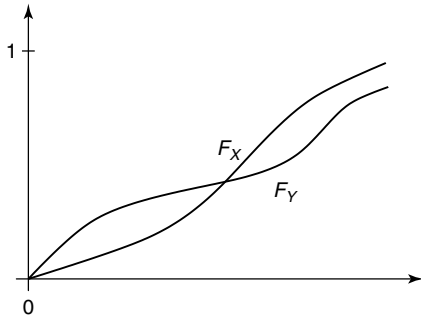


Figure 3 Cut criterion

order in equation (5) is the Karlin–Novikov cut criterion saying that the distribution functions F_X, F_Y cross exactly one time, i.e., for some x_0

$$\begin{aligned} F_X(x) &\leq F_Y(x) \quad \text{for } x < x_0 \\ \text{and } F_X(x) &\geq F_Y(x) \quad \text{for } x > x_0 \end{aligned} \quad (6)$$

Inequality (6) is a consequence of twice crossing of the densities f_X, f_Y (see Figures 3 and 4).

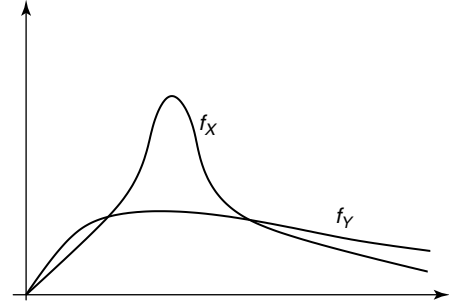


Figure 4 Cut criterion for densities

Applications

Individual and Collective Risk Model

The classical individual model of risk theory has the form

$$X_{\text{ind}} = \sum_{i=1}^n b_i I_i \quad (7)$$

where $I_i \sim B(1, p_i)$ are independent Bernoulli distributed random variables. With probability p_i contract i will yield a claim of size $b_i \geq 0$ for any of the n policies. Replacing the claims $b_i I_i$ by $b_i N_i$ with Poisson-distributed $N_i \sim \text{Poisson}(\lambda_i)$ we obtain the classical approximation of the individual risk model by the collective model

$$X_{\text{coll}} = \sum_{i=1}^n b_i N_i \quad (8)$$

X_{coll} is called *collective model* since it has a representation of the form

$$X_{\text{coll}} = \sum_{i=1}^N X_i \quad (9)$$

with $N \sim \text{Poisson}(\lambda)$, $\lambda = \sum_{i=1}^n \lambda_i$ where (X_i) are independent and identically distributed with point masses $\frac{\lambda_i}{\lambda}$ at b_i , $1 \leq i \leq n$. If we choose $\lambda_i = p_i$, then the expected payments coincide; $EX_{\text{coll}} = EX_{\text{ind}}$ since $E I_i = p_i = E N_i$.

As an application of stochastic and stop loss ordering we get that the collective risk model X_{coll} leads to an overestimate of the risks and, therefore, also to an increase of the corresponding risk premiums for the whole portfolio

$$X_{\text{ind}} \leq_{sl} X_{\text{coll}} \quad (10)$$

From the cut criterion it follows that $I_i \leq_{sl} N_i$ and, therefore, by convolution stability of the stop loss order we obtain the comparison in equation (10). Choosing the parameter λ_i in the collective model as $\lambda_i = -\log(1 - p_i) > p_i$ we obtain a collective model, which is even more on the safe side for the insurer. With this choice, even stochastic ordering holds.

$$X_{ind} \leq_{st} X_{coll} \quad (11)$$

Reinsurance Contracts

As a second application of the stop loss ordering, we consider reinsurance contracts $I(X)$ for a risk X , where $0 \leq I(X) \leq X$ is the reinsured part of the risk X and $X - I(X)$ is the retained risk of the insurer. Consider the stop loss reinsurance contract $I_a(X) = (X - a)_+$, where a is chosen such that $E I_a(X) = E I(X)$. Then it follows from the cut criterion (equation (6)) that for any reinsurance contract $I(X)$

$$X - I_a(X) \leq_{sl} X - I(X) \quad (12)$$

holds, i.e., the stop loss contract $I_a(X)$ minimizes the retained risk of the insurer. Thus it is the *optimal reinsurance contract* for the insurer in the class of all contracts $I(X)$, which have the same expected risk.

Diversification of Risks

The stop loss ordering also gives a clue to the *diversification of risk* problem. Let $X_1, \dots, X_n$ be n independent and identically distributed risks. For any diversification strategy (p_i) , $0 \leq p_i$ with $\sum_{i=1}^n p_i = 1$, we obtain a diversified risk portfolio $\sum_{i=1}^n p_i X_i$ with p_i relative shares of the i th risk. Then by stochastic ordering techniques, it is easy to establish that under all diversification schemes (p_i) the *uniform diversification is optimal*, i.e., it has the lowest risk

$$\frac{1}{n} \sum_{i=1}^n X_i \leq_{sl} \sum_{i=1}^n p_i X_i \quad (13)$$

In fact, since the expectations of both sides in (equation (13)) coincide we even get the ordering in the stronger sense that equation (1) holds for $\mathcal{F} = \mathcal{F}_{cx}$, the *convex ordering*. Thus it also allows, in particular, comparison of the angle (put) function $(a - x)_+$.

Dependent Portfolios Increase Risk

The following example demonstrates the strong influence that dependence, between individual risks, may have on the risk of the joint portfolio (*see Dependent Insurance Risks*). Let $X_i = \Theta Y_i + (1 - \Theta) Z_i$, $1 \leq i \leq 10^5$, be a mixed model for a large portfolio with Bernoulli distributed $Y_i \sim B(1, \frac{1}{100})$, $Z_i \sim B(1, \frac{1}{1100})$, and $\Theta \sim B(1, \frac{1}{100})$, where all Θ, Y_i, Z_i are independent. It is easy to calculate that $X_i \sim B(1, \frac{1}{1000})$, i.e., each individual contract X_i yields a unit loss with small probability $\frac{1}{1000}$. The presence of Θ in the model implies an increase of the risk of all contracts (*positive dependence*) in rare cases. Thus typically, $\Theta = 0$ and the risks in the portfolio produce independently with small probability $\frac{1}{1100}$ a unit loss to the insurer. With small probability, however, a bad event $\Theta = 1$ happens, which causes all contracts to undergo an increase in risk, which now yields independently, with probability $\frac{1}{100}$, a unit loss for any of the contracts.

The common risk factor Θ introduces a small positive correlation of magnitude $\frac{1}{1000}$ between the individual risks. It is interesting to compare the total risk $T_n = \sum_{i=1}^n X_i$ in the mixed model (X_i) with the total risk $S_n = \sum_{i=1}^n W_i$ in an independent portfolio model (W_i), where $W_i \sim B(1, \frac{1}{1000})$ are distributed identically to X_i . Then, from stochastic ordering results as above, we obtain that the risk of T_n is bigger than that of S_n

$$S_n \leq_{sl} T_n \quad (14)$$

In fact, also $S_n \leq_{cx} T_n$ since $ES_n = ET_n = 100$. What is the magnitude of this difference? By the central limit theorem S_n is approximately normal distributed with mean $\mu = 100$ and dispersion $\sigma = 10$. Thus, $t = \mu + 5\sigma = 150$ is a *safe* retention limit for S_n . $P(S_n > t)$ is extremely small and the net premium is approximately

$$E(S_n - t)_+ \approx 2.8 \times 10^{-8} \quad (15)$$

The positive dependence in the mixed model (X_i), which is small in terms of correlation, causes a big increase of risk. For the mixed model we get, using the same retention limit t as in the independent model, the considerable stop loss premium

$$E(T_n - t)_+ \approx 8.5 \quad (16)$$

The presence of positive dependence in the mixed model implies that with probability about $\frac{1}{100}$, the

risk of the joint portfolio T_n is greater than 800. Thus, neglecting the common risk factor Θ and basing the calculation of premiums on the incorrect independent model (W_i) would lead to a disaster for the insurance company. The effect demonstrated in this example is present in a similar form in many related mixture models and clearly shows the necessity and importance of correct modeling of risks and also the necessity to use more advanced stochastic ordering tools going beyond mean-variance analysis.

An Outlook

Many further tools and orderings have been investigated in the literature to describe specific classes of risk models, how they compare with reference to different kinds of risk measures, and in which sense one distribution represents a more dangerous situation than another distribution. Several orderings have been developed for risk vectors or risk portfolios in particular, for measuring the degree of positive dependence in multivariate portfolios and its influence on various risk functionals. Important examples of positive dependence orderings are the supermodular, the directionally convex, the Δ -monotone orderings, and the positive orthant dependence orderings.

In various circumstances and for various risk measures results of the type, *more positive dependence implies higher risk* have been established.

A main actual topic of ordering of risks is to obtain sharp bounds on the risk based on partial knowledge of the dependence structure. For detailed exposition of ordering of insurance risks we refer the reader to the following references.

Further Reading

- Kaas, R., Goovaerts, M., Dhaene, J. & Denuit, M. (2001). *Modern Actuarial Risk Theory*, Kluwer Academic Publishers.
- Müller, A. & Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*, John Wiley & Sons.
- Denuit, M., Dhaene, J., Goovaerts, M. & Kaas, R. (2005). *Actuarial Theory for Dependent Risks Measures, Orders, and Models*, John Wiley & Sons.
- Rüschendorf, L. (2005). Stochastic ordering of risks, influence of dependence, and a.s. constructions, in *Advances on Models, Characterizations, and Applications*, N. Balakrishnan, I.G. Bairamov and O.L. Gebizlioglu, Chapman & Hall, pp. 15–56.
- Rolski, T., Schmidli, H., Schmidt, V. & Teugels, J. (1998). *Stochastic Processes for Insurance and Finance*, John Wiley & Sons.

LUDGER RÜSCHENDORF

Parametric Probability Distributions in Reliability

In many applications of reliability theory and quantitative risk assessment, one is interested in random quantities for which one assumes a parametric probability distribution, as is generally often done in statistics and stochastics. In this paper, we briefly introduce some of the most commonly used parametric probability distributions in reliability, restricting ourselves to univariate random quantities and the most basic forms of the distributions. Far more detailed presentations of these distributions, with historical notes, discussions of properties and applications, and further generalizations, can be found in the *Wiley Encyclopedia of Statistical Sciences* [1]. Key text books in reliability and related topics tend to present these (and more) parametric probability distributions in greater detail, see e.g., Hougaard [2], Lawless [3], Martz and Waller [4], or Meeker and Escobar [5]. These main distributions are presented in a very brief manner in this article; they are used in many other articles in this encyclopedia, providing examples of their applications.

Parametric probability distributions are used both in stochastic analyses of reliability of systems, where they are mostly assumed to be fully known and corresponding properties of the system are analyzed, and in statistical inference, where process data are used to estimate the parameters of the distribution, often followed by a specific inference of interest. In the latter case, one must take care to also diagnose the assumptions underlying the specific parametric distribution assumed, to ensure a reasonable fit with the empirical data [1].

We present some main probability distributions for both continuous and discrete random quantities. For the former, parametric probability distributions can be uniquely specified via the probability density function (pdf) $f(t)$, the cumulative distribution function (cdf) $F(t)$, the survival function $S(t)$ (see **Insurance Applications of Life Tables; Risk Classification/Life; Mathematics of Risk and Reliability: A Select History**), or the hazard rate $h(t)$ (see **Mathematical Models of Credit Risk; Imprecise Reliability**). For example, for a nonnegative random quantity, as often used in reliability if one is interested in a random lifetime, these functions are related

as follows: $F(t) = \int_0^t f(u) du$, $S(t) = 1 - F(t)$ and $h(t) = f(t)/S(t)$. The hazard rate is often considered to be particularly attractive in reliability, as it represents an instantaneous rate of the occurrence of a failure, conditioned on no failure until now. For example, wear-out of a unit over time can be represented by an increasing hazard rate.

In the section titled “Normal and Related Distributions”, we briefly consider the normal distribution, and some related distributions, which, although important, play perhaps a lesser role in reliability than in other areas of application of statistics. In the section “Exponential and Related Distributions”, we present the exponential distribution, and the Weibull and gamma distributions, which are popular models in reliability and which can be considered as generalizations of the exponential distribution. The section titled “Distributions for Discrete Random Quantities” presents some important distributions for discrete random quantities; and in the section “Other Distributions”, some further important distributions for reliability are briefly mentioned. The paper concludes with some further remarks in the final section.

Normal and Related Distributions

The normal distribution (also known as the *Gaussian distribution*) is arguably the most important probability distribution in statistics, as it occurs as the limiting distribution when a sum of random quantities is considered (central limit theorem). In general, it also plays a big role in quantitative risk assessment, e.g., when risks of portfolios of investments are considered, but its role in reliability is somewhat less important. However, it is frequently used as a suitable probability distribution for the (natural) logarithm of a random lifetime, in which case the distribution of the lifetime is called *lognormal*. It is also related to the inverse Gaussian distribution (see **Lifetime Models and Risk Assessment**), which is of importance in some processes in reliability.

Normal Distribution

The normal distribution has two parameters, $-\infty < \mu < \infty$ and $\sigma^2 > 0$, which are equal to its mean and

variance, so its standard deviation is σ , and its pdf is

$$f(x|\mu, \sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for } -\infty < x < \infty \quad (1)$$

The cdf of the normal distribution is not available in closed form, so computations often used tables of the cdf of the standard normal distribution, which has $\mu = 0$ and $\sigma^2 = 1$, using the fact that $(X - \mu)/\sigma$ is standard normally distributed if X is normally distributed with parameters μ and σ^2 . Such tables are included in statistics text books, but all main statistical and mathematical software nowadays have good routines for the computation of the cdf of the normal distribution. Its direct use in reliability is often restricted to modeling residual or error terms in regression models, but it plays an important role as a model for log-transformed lifetimes.

Lognormal Distribution

A random quantity $T > 0$ is lognormally distributed if $X = \ln T$ has a normal distribution, so that $T = \exp(X)$ has pdf

$$f(t|\mu, \sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma t} \exp\left(-\frac{(\ln t - \mu)^2}{2\sigma^2}\right) \quad \text{for } t > 0 \quad (2)$$

The mean and variance of T are $\exp(\mu + \sigma^2/2)$ and $[\exp(\sigma^2) - 1][\exp(2\mu + \sigma^2)]$, respectively. This distribution is quite popular as a model for lifetimes, even though its hazard rate has the somewhat unattractive property that it increases, from $h(0) = 0$, to a maximum, and thereafter decreases to 0 for $t \rightarrow \infty$. However, if attention is particularly directed toward early failure times, yet with a wear-out effect, then this model might be appropriate. One possible argument justifying the use of this model is related to the central limit theorem argument for the normal distribution. Informally stated, the latter implies that, if a random quantity can be considered to be the sum of many independent random quantities, then its distribution will be approximately normal. Hence, the same property holds for $\ln T$ if T has a lognormal distribution, with the log transform here meaning that T , for this argument, can be interpreted as the product of many independent random quantities, which might be attractive in certain types of failure processes. An

obvious argument for the popularity of the lognormal distribution was always the wide availability of statistical tables for the normal distribution, an argument that is less relevant nowadays.

Inverse Gaussian Distribution

Contrary to what the name might perhaps suggest, this is not a distribution for a simple transformation of a normally distributed random quantity. However, it becomes ever more important in reliability theory because of the fact that it is appropriate for stopping times in Brownian motion (*see Optimal Stopping and Dynamic Programming*) (“Gaussian”) processes, and these (and related) processes are playing an increasingly important role in reliability modeling. Suppose that a Brownian motion, starting at 0 at time $t = 0$, has drift $\psi \geq 0$ and variance σ^2 , then the time to reach the value $a > 0$ for the first time has an inverse Gaussian distribution with parameters $\beta = \frac{\psi}{2\sigma^2}$ and $\alpha = \frac{a}{\sigma\sqrt{2}}$. The pdf of this distribution is

$$f(t|\alpha, \beta) = \frac{\alpha \exp(2\alpha\beta^{1/2}) \exp(-\beta t - \alpha^2/t)}{\pi^{1/2} t^{3/2}} \quad \text{for } t > 0 \quad (3)$$

Padgett and Tomlinson [6] present an excellent example of the use of such processes to describe degradation, providing a model of continuous cumulative damage. They present a general accelerated test model in which failure times and degradation measures are combined for inference about system lifetime, with the drift of the process depending on the acceleration variable. Their paper includes an illustrative example using degradation data observed in carbon-film resistors.

Exponential and Related Distributions

The probability distributions in this section, for a continuous positive random quantity T , are very popular models for lifetimes, used in a wide variety of applications in reliability and beyond. We briefly introduce these distributions in basic forms, they can be generalized in a variety of ways. The most obvious generalization is inclusion of a location parameter τ , such that $T > \tau$, so effectively this shifts the start of the distribution from 0 to τ .

Mathematically, this is straightforward, but one must be careful in case of statistical inference based on the likelihood function, as for some models the maximum-likelihood estimator of τ is equal to the smallest observation in the available data, which is unreasonable from several perspectives.

Exponential Distribution

The exponential distribution has a constant hazard rate, say $h(t|\lambda) = \lambda > 0$, so its pdf is

$$f(t|\lambda) = \lambda \exp(-\lambda t) \quad \text{for } t > 0 \quad (4)$$

It is the unique model, for continuous T , with the so-called memory-less property, that is, the probability for the event to take place in a future interval does not depend on the current age of the item or individual considered. This distribution models the random times between events in homogeneous Poisson processes.

Weibull Distribution

The Weibull distribution is a very widely used probability distribution in reliability, and it has two parameters: scale parameter $\alpha > 0$ and shape parameter $\beta > 0$. It is easiest introduced *via* its hazard rate, which is

$$h(t|\alpha, \beta) = \alpha\beta(\alpha t)^{\beta-1} \quad \text{for } t > 0 \quad (5)$$

The corresponding survival function is

$$S(t|\alpha, \beta) = \exp(-(\alpha t)^\beta) \quad \text{for } t > 0 \quad (6)$$

For $\beta = 1$ this distribution is an exponential distribution, whereas $\beta > 1$ leads to an increasing hazard rate, and hence can be considered to model wear-out, as often deemed appropriate for mechanical units, and $\beta < 1$ leads to decreasing hazard rate, hence modeling wear-in of a product, as often advocated for electronic units. One Weibull distribution cannot model both wear-in and wear-out at different stages, e.g., it cannot resemble a “bath-tub shaped hazard rate”, but different Weibull distributions are often advocated for the different stages of a unit’s life. Although its form is mathematically straightforward, statistical inference based on the Weibull distribution is less trivial owing to the fact that there are no reduced-dimensional sufficient statistics for the shape

parameter β . This distribution is routinely available in standard mathematical and statistical software. The Weibull distribution with $\beta = 2$, hence with hazard rate a linear function of t , is also known as the *Rayleigh distribution*. Weibull distributions are also frequently used in regression-type inferences in reliability, enabling information from covariates to be taken into account. There is also a useful relationship between the Weibull distribution and an extreme value distribution (also called the *Gumbel distribution*), which occurs as a limiting distribution for extreme values of a sample, under some assumptions on the tail of the sample distribution: if T has a Weibull distribution, then $\ln T$ has such an extreme value distribution.

Gamma Distribution

The gamma distribution is also a useful lifetime model in many reliability applications, and has two parameters: scale parameter $\lambda > 0$ and shape parameter $\kappa > 0$. Its pdf is

$$f(t|\kappa, \lambda) = \frac{\lambda(\lambda t)^{\kappa-1} e^{-\lambda t}}{\Gamma(\kappa)} \quad \text{for } t > 0 \quad (7)$$

where $\Gamma(\kappa)$ is the gamma function, which of course equals the integral of the numerator of this pdf over $t > 0$. Computation of the cdf, survival function, and hazard rate is somewhat awkward as it involves the incomplete gamma function, yet this distribution is available in the leading mathematical and statistical software. The gamma distribution with $\kappa = 1$ is an exponential distribution, and gamma distributions with integer κ are also known as *Erlang distributions*. The pdf of the gamma distribution can have many different shapes, but it is not as commonly used in reliability as the Weibull distribution, possibly because of the lack of a clearly interpretable hazard rate. An attractive interpretation, however, occurs for integer κ , in which case a gamma distributed lifetime can be interpreted as the sum of κ independent exponentially distributed random quantities, which might, for example, be suitable if one wishes to model the lifetime of a system with spare units. This latter property also makes the gamma distribution popular in models for queuing systems. With increased popularity of Bayesian methods in statistics, the gamma distribution is used frequently as it is the conjugate prior distribution for both exponential and Poisson sampling distributions.

Distributions for Discrete Random Quantities

In reliability, as in many other application areas of stochastic modeling and statistical inference, one is also regularly interested in discrete random quantities, e.g., when counting the number of components that have performed a task successfully. The binomial distribution is probably the most frequently used distribution, whereas the negative binomial distribution and the Poisson distribution also deserve explicit mention. Of course, there are many variations and generalizations to these and other distributions for discrete random quantities; for these we refer again to the literature mentioned in the introductory paragraphs.

Binomial Distribution

Suppose that X denotes the number of successes in n independent trials, where each trial results either in a success with probability θ or in a failure with probability $1 - \theta$. The probability distribution of X , for a given value of n , is given by

$$P(X = x|\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \quad \text{for } x \in \{0, 1, \dots, n\} \quad (8)$$

The expected value and variance of X are $n\theta$ and $n\theta(1 - \theta)$, respectively. Computation of these probabilities is straightforward if n is not too large, for large values of n two approximations can be used: for values of x close to 0 or n , the Poisson distribution with expected value $n\theta$ can be used as a reasonable approximation, whereas for other values of x an approximation based on the normal distribution, with the same expected value and variance is suitable. A standard “text-book” example for a situation where the binomial distribution is appropriate is n tosses of a possibly biased coin, where the coin lands heads up with probability θ on each toss, and where the number of tosses with heads up is counted. In situations where the possible outcomes are in more than two unordered categories, the multinomial distribution provides a suitable generalization of the binomial distribution [1].

Negative Binomial Distribution

The negative binomial distribution is a variation to the binomial distribution, with the total number of trials not a predetermined constant, but instead one counts the number of trials until a specified number of successes have occurred. Suppose again that trials are independent, each resulting either in a success with probability θ or a failure with probability $1 - \theta$, and let N denote the number of trials required to obtain the r th success. Then the probability distribution of N , for fixed r , is given by

$$P(N = n|\theta) = \binom{n-1}{r-1} \theta^r (1 - \theta)^{n-r} \quad \text{for } n = r, r+1, \dots \quad (9)$$

Sometimes the negative binomial distribution is defined slightly differently, that is, as counting the number of trials prior to the r th success. This distribution is important in its own right in probabilistic risk assessment, in reliability but also, for example, in quality control, but it is also increasingly used in Bayesian statistical inferences [4], as it occurs as the posterior predictive distribution for Poisson sampling with a conjugate gamma prior distribution. The special case of $r = 1$ is known as the *geometric distribution*, which simply counts the number of trials until the first success has occurred.

Poisson Distribution

A random quantity X has a Poisson distribution with expected value μ if its probability distribution is given by

$$P(X = x|\mu) = \frac{e^{-\mu} \mu^x}{x!} \quad \text{for } x = 0, 1, 2, \dots \quad (10)$$

This distribution is particularly important in stochastic processes, where it counts the number of events in a given period of time. For example, for a nonhomogeneous Poisson process with failure rate function $\lambda(t)$, the number of events in time interval $[t_1, t_2]$ has a Poisson distribution with expected value $\mu = \int_{t_1}^{t_2} \lambda(u) du$.

Other Distributions

Many other parametric probability distributions play an important role in some specific reliability applications. We briefly mention some of them, but refer to

the literature, e.g., the sources mentioned in the introductory paragraphs, for more details. Historically, the log-logistic distribution (*see* **Lifetime Models and Risk Assessment**) was frequently used in reliability, with a lifetime T having a log-logistic distribution if $Y = \ln T$ has a logistic distribution with pdf

$$f(y|\mu, \sigma) = \frac{\exp[(y - \mu)/\sigma]}{\sigma(1 + \exp[(y - \mu)/\sigma])^2} \quad \text{for } -\infty < y < \infty \quad (11)$$

with parameters $-\infty < \mu < \infty$ and $\sigma > 0$. This distribution is very similar to the normal distribution, but its survival function is available in closed form and hence it makes it easier to deal with right-censored observations, which often occur in reliability applications.

The uniform distribution, which has a constant pdf over a finite interval, is also useful in some reliability applications. For example, it models the times of events in a homogeneous Poisson process (*see* **Product Risk Management: Testing and Warranties; Reliability Growth Testing**) if the total number of events over a time interval of predetermined length is given. It is also frequently used in Bayesian statistics, as a prior distribution which, as is sometimes argued, can reflect quite limited prior knowledge being available. The beta distribution is also popular in Bayesian statistics, as it is a conjugate prior for binomial sampling models.

The Gompertz distribution is characterized by an attractive form for the hazard rate, given by $h(t) = \lambda\phi^t$ for $t > 0$, with scale parameter $\lambda > 0$ and shape parameter $\phi > 0$. Clearly, $\phi = 1$ gives the Exponential distribution, $\phi > 1$ models an increasing hazard rate (and hence can be used to model wear-out), but for $\phi < 1$, so decreasing hazard rate (“wear-in”), a problem occurs as the corresponding probability distribution is improper (i.e., its density does not integrate to 1 for $t > 0$). This latter aspect could be interpreted as if a proportion of the population considered cannot experience the event of interest. We mention this distribution mostly due to its apparently attractive hazard rate, and to emphasize the complications that can occur already with such rather simple mathematical forms. The Gompertz distribution has been used in actuarial mathematics since the early 19th century.

When we presented the Weibull distribution in the section “Exponential and Related Distributions”, we

briefly referred to an extreme value distribution. Generally, there are several extreme value distributions (three main functional forms), which occur for the maximum (or minimum) of n identically distributed real-valued random quantities if $n \rightarrow \infty$, and which often provide good approximations for the distribution of this maximum (or minimum) if n is large. These distributions are useful in a variety of reliability applications, for example, those related to reliability of systems or to structural reliability and overloads.

In this paper, we have restricted attention to univariate random quantities. Of course, in many reliability applications one is interested in multivariate random quantities, we refer to Hougaard [2] for an excellent presentation of related statistical theory, including useful parametric distributions for multivariate data.

Concluding Remarks

Basic parametric probability distributions, such as those presented in this paper, are also often used as part of more complex statistical models, such as mixture models, Bayesian hierarchical models, models with covariates or parametric models for processes. We refer the reader to the literature (*see* the introductory paragraphs of this article) for further discussion of such models; several examples are provided in other articles in this encyclopedia. Parametric models are certainly not always suitable. In particular, if one has many data available the use of nonparametric or semiparametric methods might be preferable because of their increased flexibility to adapt to specific features of the data. With increased computational power, applications of non- and semiparametric methods have become more widely available, yet parametric distributions are likely to remain important tools in reliability and risk assessment.

It is often an advantage if the parameters have an intuitive interpretation. Several of the parametric probability distributions discussed in this paper, including normal, exponential and gamma (but unfortunately not the Weibull distribution, unless its shape parameter is assumed to be a known constant), belong to the so-called exponential family of distributions [7], for which the parameters can be related to one- or two-dimensional sufficient statistics that summarize the data; this has the added benefit of available

conjugate priors to simplify computation in Bayesian statistics.

When choosing a parametric distribution as a model in a specific application, one often has to look for a suitable balance between simplicity of the model and corresponding computational aspects, and how flexible and realistic the model is. As mentioned before, for trustworthiness of statistical inferences based on an assumed parametric model, it is important that the choice of the model is well explained and, wherever possible, diagnostic methods (e.g., “goodness-of-fit tests” [1]) are used to check whether the model fits well with the available data.

References

- [1] Kotz, S., Read, C.B., Balakrishnan, N. & Vidakovic, B. (2006). *Wiley Encyclopedia of Statistical Sciences*, John Wiley & Sons, New York.
- [2] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer, New York.
- [3] Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.
- [4] Martz, H.F. & Waller, R.A. (1982). *Bayesian Reliability Analysis*, Wiley, New York.
- [5] Meeker, W.Q. & Escobar, L.A. (1998). *Statistical Methods for Reliability Data*, John Wiley & Sons, New York.
- [6] Padgett, W.J. & Tomlinson, M.A. (2004). Inference from accelerated degradation and failure data based on Gaussian process models, *Lifetime Data Analysis* **10**, 191–206.
- [7] Lee, P.M. (1997). *Bayesian Statistics: An Introduction*, 2nd Edition, Arnold, London.

FRANK P.A. COOLEN

Persistent Organic Pollutants

Widespread concern about persistent organic pollutants (POPs) dawned with the demonstration that the organochlorine pesticide Dichloro-diphenyl-trichloro-ethane (DDT) was causing widespread reproductive failures among fish-eating birds in the 1960s and 1970s [1, 2]. The endocrine-disrupting properties of DDT and its metabolites caused eggshell thinning, and nesting females crushed their unwitting offspring before they had a chance to hatch [3, 4].

This fundamental lesson in wildlife toxicology launched a rocky start along a path that would scrutinize and eventually regulate chemicals possessing three fundamental properties: those which are persistent, bioaccumulative, and toxic (PBT). Hence the catchall term *POP* amalgamated a wide group of chemicals for the purposes of scientific investigation, risk assessment, management scrutiny, and regulation. The extent to which individual POPs are deemed also to be “PBT” represents the basis for their regulation and inclusion in international treaties.

The Persistent Organic Pollutants (POPs)

It is important to first review some of the basic features of the POPs. POPs include such notable members as the *polychlorinated biphenyls* (PCBs) (see **Polychlorinated Biphenyls**), *dioxins* (see **Health Hazards Posed by Dioxin**) (polychlorinated dibenzo-*p*-dioxins (PCDDs)), furans (polychlorinated dibenzofurans (PCDFs)), organochlorine pesticides, and polybrominated diphenyl ethers (PBDEs). While small quantities of dioxins and furans originate from natural sources including forest fires and volcanoes, the spatial and temporal trends in the levels of these contaminants reveal human activities to dominate their release into the biosphere. PCBs and organochlorine pesticides have no known natural sources. POPs of anthropogenic origin are either produced intentionally (e.g., PCBs, organochlorine pesticides) or represent unintentional by-products of processes (e.g., dioxins and furans resulting from incomplete combustion, pulp kraft bleaching, or pesticide manufacture). Some are or were produced for “closed” or partially closed applications (e.g.,

PCBs as dielectric fluids), while others are or were produced for intentional application to the environment (e.g., pesticide application to crops). The physicochemical characteristics of the POPs underlie much of the current concerns about regulation, transport/fate, and effects (see Table 1).

POPs are persistent, as evidenced from long air, water, or soil half-lives, and resistance to photolytic or microbial breakdown [10, 11]. This feature has bearing on the ecological risks associated with POPs, since any release of POPs to the environment can lead to long-term local or global concerns (see **Ecological Risk Assessment**). This also has bearing on remediation measures should site-specific cleanups be instigated, as special measures must be employed to reduce the risks of further environmental contamination (e.g., sediment capping, landfill) or to destroy the chemicals in question (e.g., high-temperature incineration at specialized facilities and/or the use of specialized bacteria capable of breaking down the product). While POPs mobilized from a source or application are generally “beyond” management, site-specific remediation strategies have been partially successful at facilitating degradation of some parent POPs [12–15].

POPs are bioaccumulative, as evidenced by their accumulation in organisms, their magnification with increasing trophic levels in aquatic food webs, and their resistance to metabolic elimination in biota. This feature represents perhaps the most pressing concern from a risk assessment perspective. Fish-eating wildlife including seabirds and marine mammals are particularly vulnerable to contamination, becoming highly contaminated in such industrial regions as the St Lawrence estuary in Quebec, Canada [16, 17], the Baltic Sea in northern Europe [18, 19], and coastal California in the United States [20, 21]. Certain human consumer groups that rely heavily on seafoods have also been found to be more contaminated than their supermarket-oriented counterparts [22–24], a finding that drew home concerns about risk that had previously been seen to be largely an ecological issue. POPs emerged as a human concern, changing the nature of the perception, the risk, and the risk paradigm.

POPs are toxic, in that members can disrupt endocrine (hormone) systems and alter the growth and development of skeletal, reproductive, neurological, and immune systems. While adverse effects

2 Persistent Organic Pollutants

Table 1 The physicochemical properties of the persistent organic pollutants (POPs) serve to guide regulatory scrutiny. Of these POPs twelve have been targeted for international action under the terms of the Stockholm Convention, while others remain as “candidates” for possible inclusion. The discovery of many of these POPs in remote regions raised concerns about possible health effects in human and wildlife^(a)

	Common name	Use	Log K_{ow}	Molecular weight
Stockholm POPs	Aldrin	Insecticide; seed treatment	5.17–7.4 ^(b)	364.92
	Chlordane	Insecticide; earthworm control	6.00 ^(b)	409.78
	DDT	Insecticide – primarily for malaria control	4.89–6.914 ^(b)	354.49
	Dieldrin	Insecticide	3.692–6.2 ^(b)	380.91
	Dioxins	By-product of manufacturing/combustion	4.75–8.20 ^(c)	218.5–460.0
	Furans	By-product of manufacturing/combustion	5.44–8.0 ^(c)	237.1–443.8
	Endrin	Insecticide; rodenticide	5.2 ^(d)	380.92
	HCB	Pesticide; by-product	5.73 ^(d)	284.78
	Heptachlor	Insecticide; earthworm control	4.40–5.5 ^(b)	373.32
	Mirex	Insecticide; fire retardant	5.28 ^(d)	545.5
	PCBs	Dielectric fluids	4.3–8.26 ^(c)	188.7–498.7
	Toxaphene	Pesticide; control of scabies	3.23–5.50 ^(b)	413.82
Candidate POPs	Chlordecone	Insecticide	5.41 ^(d)	490.6
	HCH	Insecticide; wood treatment	3.72 ^(d)	290.8
	PCP	Biocide; wood preservative	5.12 ^(d)	266.34
	Endosulfan	Insecticide	3.83 ^(d)	406.93
	HCBD	Pesticide; by-product	4.78 ^(d)	260.76
	Dicofol	Pesticide	4.28 ^(d)	370.47
	Methoxychlor	Insecticide; biocide	5.08 ^(d)	345.65
	HBCD	Brominated flame retardant	5.645 ^(d)	631.69
	BDEs			
	Penta-BDE	Brominated flame retardant	6.53–6.76 ^(e)	564.7
	Octa-BDE	Brominated flame retardant	7.90 ^(e)	801.47
	Deca-BDE	Brominated flame retardant	8.70 ^(e)	959.17
	CB			
	Penta-CB	Dielectric fluid; fungicide; flame retardant	5.18 ^(d)	250.34
	Tetra-CB	Insecticide; dielectric fluid; by-product	4.6 ^(d)	215.89
	SCCPs			
	<i>N</i> -Tetradecane	Flame retardant; industrial fluid	7.20 ^(d)	198.4
	Poly-CNs			
	Mono-CN	Heat transfer fluids; wood preservative	4.08 ^(d)	162.61
	Hexa-CN	Heat transfer fluids; wood preservative	7 ^(d)	334.74
	OCS	Unintentional production	6.2 ^(f)	379.71
	PAHs			

Table 1 continued

Common name	Use	Log K_{ow}	Molecular weight
Acenaphthene	Unintentional production	3.92 ^(d)	154.21
Benzo(B)fluoranthene	Unintentional production	6.60 ^(d)	252.32

^(a)DDT: dichloro-diphenyl-trichloroethane; HCB: hexachlorobenzene; PCBs: polychlorinated biphenyls; HCH: hexachlorocyclohexane; PCP: pentachlorophenol; HCBd: hexachlorobutadiene; HBCD: hexabromocyclododecane; BDE: brominated diphenyl ether; PFCs: perfluoro chemicals; CB: chlorobenzene; SCCPs: short-chained chlorinated paraffins; CNs: chlorinated naphthalenes; OCS: octachlorostyrene; PAHs: polycyclic aromatic hydrocarbons

^(b)Reference [5]

^(c)Reference [6]

^(d)Reference [7]

^(e)Reference [8]

^(f)Reference [9]

have been noted in several high-profile accidental poisonings of humans [25–27], the widespread exposure of all humans to low-to-moderate levels of POPs in their diet represents a more insidious challenge to risk assessors and managers. Adverse health effects, including small-for-gestational-age, immune function deficiencies, altered thyroid hormone levels, and impaired neurological development, have been documented in Dutch infants in a long-term study of average (i.e., not belonging to a cohort of concern) citizens [28, 29].

A number of similar health effects noted in contaminated wildlife populations not only provided an epidemiological basis for “real-world” interpretation, but also initiated widespread concerns about possible human health implications. The “human–wildlife” connection was born [30, 31]. In both cases, however, causal linkages were lacking, a feature that led to numerous mechanistic studies of laboratory animals wherein single-chemical, dose–response experiments provided insight into mechanisms of action for different chemicals, evaluated the relative potencies of different chemicals with similar structures (e.g., the toxic equivalency quotient (TEQ) for dioxin-like PCBs, PCDDs and PCDFs, or TEQ) [32], and provided a basis for extrapolation among species.

Many of the concerns about POP-associated toxic risk to humans originated as a result of the studies of wildlife. Associations between POP levels and adverse health effects in seabirds and marine mammals have been described in many industrial regions. Reproductive impairment in Baltic seals, St Lawrence belugas and California sea lions has been described during peak DDT and PCB use in the 1960 and

1970s [33–35]. The onset of skeletal abnormalities in Baltic seal populations after World War II [36], and the appearance of multifaceted abnormalities in Great Lakes fish-eating birds [37, 38], indicated a disruption of developmental processes. Mass mortalities associated with the introduction of new pathogens into vulnerable populations of seals inhabiting contaminated industrial regions are thought to be in part due to a contaminant-related reduction in the effectiveness of their immune systems to disease [39, 40]. Two major captive feeding studies of harbor seals have contributed to a more controlled assessment of POP effects, wherein a contaminated diet of fish reduced reproductive performance, and caused immunotoxicity and endocrine disruption [39, 41–43].

While the “PBT” characteristics largely shape the nature of the movement, fate, and effects of POPs in the environment, one additional feature is considered under the terms of the United Nations Environment Programme (UNEP) sponsored Stockholm Convention. This includes the “potential for long-range environmental transport”. Meeting this criterion involves demonstrating that a chemical is found in remote environments and that it was introduced to that environment *via* air, water, or biological transport. These characteristics of POPs make containment at local hotspots very difficult, and allow for escape *via* volatilization, movement with river or ocean currents, or export *via* migrating wildlife to take place. When one looks, one finds that chemicals of virtually all POP classes have been detected in remote parts of the world. Perhaps the more pressing concerns are the levels to which they accumulate, their behavior in remote food webs, their persistence

in the environment in question, and their toxicity to biota.

Transport, Fate, and Effects of POPs in the Environment

Individual POPs vary in the way in which they behave within each of the four above categories, and hence in the way in which they meet or fail to meet, the criteria set by national and international regulatory bodies. A brief summary of the salient scientific features of POPs as they enter and move through the environment, preferentially accumulate in certain species, and affect the health of organisms will help to shed light upon the links between scientific observations, risk assessments, and regulatory requirements.

POPs are fat soluble, and hence do not dissolve readily in water, but preferentially partition into lipid and organic components of the environment. Since this partitioning represents a relationship between the properties of the chemical in question and the properties of the environmental matrix within which it finds itself, partitioning coefficients have become useful proxies for the behavior of chemicals in the environment, and have greatly facilitated risk assessments at multiple levels (from design and manufacture to regulatory scrutiny to site-specific risk assessments). The partitioning of POPs between air, water, and organic phases are described by the partitioning coefficients for octanol/water (K_{ow}), octanol/air (K_{oa}), and air/water (K_{aw}).

Chemicals with higher $K_{o/w}$ values dissolve more readily in lipids, thereby driving accumulation in biota. However, an “optimal range” exists, where by those POPs falling between Log $K_{o/w}$ values of ~ 5.8 – 7.5 preferentially accumulate in aquatic organisms (see Figure 1a and b). The importance of lipids in determining the magnification of POPs in aquatic food webs is further demonstrated through correlations between POP concentrations and stable nitrogen isotopes (^{15}N : ^{14}N) measured in tissues of different species (see Figure 2; food web). With increasing trophic level, nitrogen isotopes fractionate toward a heavier signature, while POP concentrations increase.

The dispersion of POPs away from source *via* air, water, or biological transport also varies as a function of their physicochemical properties, as evidenced by fractionation as one moves away from the sites

of industrial activities into remote sites. Increasingly “light” PCB and POP signatures have been observed in lake sediments [47], burbot (*Lota lota*) [48], frogs (*Rana temporaria*) [49], and harbor seals (*Phoca vitulina*) [50] along latitudinal gradients. Altitude also influences POP patterns, with a temperature-related condensation and deposition at upper elevations [51], preferentially delivering some of the “lighter” POPs to biota [52, 53]. Moving air masses can readily transport POPs to remote sites in a matter of days, but slow-moving ocean currents and pulsating biological migrations also represent increasingly important modes of transport for POPs [54, 55]. The distillation of POPs in a global context leads to a fractionation that is contingent upon their chemical properties; this process is sometimes referred to as the *grasshopper effect* (see Figure 3) [56].

And while preferential accumulation indeed must reflect the degree to which a given chemical is “lipophilic”, its resistance to metabolic attack is also a requisite feature for magnification in aquatic food webs. Detoxification enzymes in biota possess the ability to eliminate parent POPs through the actions of the cytochrome P450 superfamily of enzymes, processes that lead to the formation of water-soluble, hydroxyl- and methyl-sulfonyl metabolites [57]. Metabolism may therefore be considered as an important driver of POP “patterns” in aquatic food webs, with an increasing dominance of persistent congeners with increasing trophic level, and the consequent loss of more metabolizable congeners [58].

The resemblance between many POPs and natural hormones underlies concerns about the potential for adverse health effects in humans and wildlife. By interacting with natural hormone receptors in biota, POPs can disrupt endocrine processes and influence the growth, development, and maintenance of organ systems. Carefully controlled, single-chemical, dose-response experiments using laboratory animals of cell-based studies have provided a mechanistic understanding of the effects of many individual POPs. Such certainty is not generally available for humans and wildlife, something that has required alternative study designs or approaches. While complex mixtures of POPs complicate a mechanistic interpretation of the “causal” chemical, a “weight of evidence” approach has been applied to evaluate POP risks and identify specific chemical classes of concern, such as the PCBs and DDT [33, 59, 60].

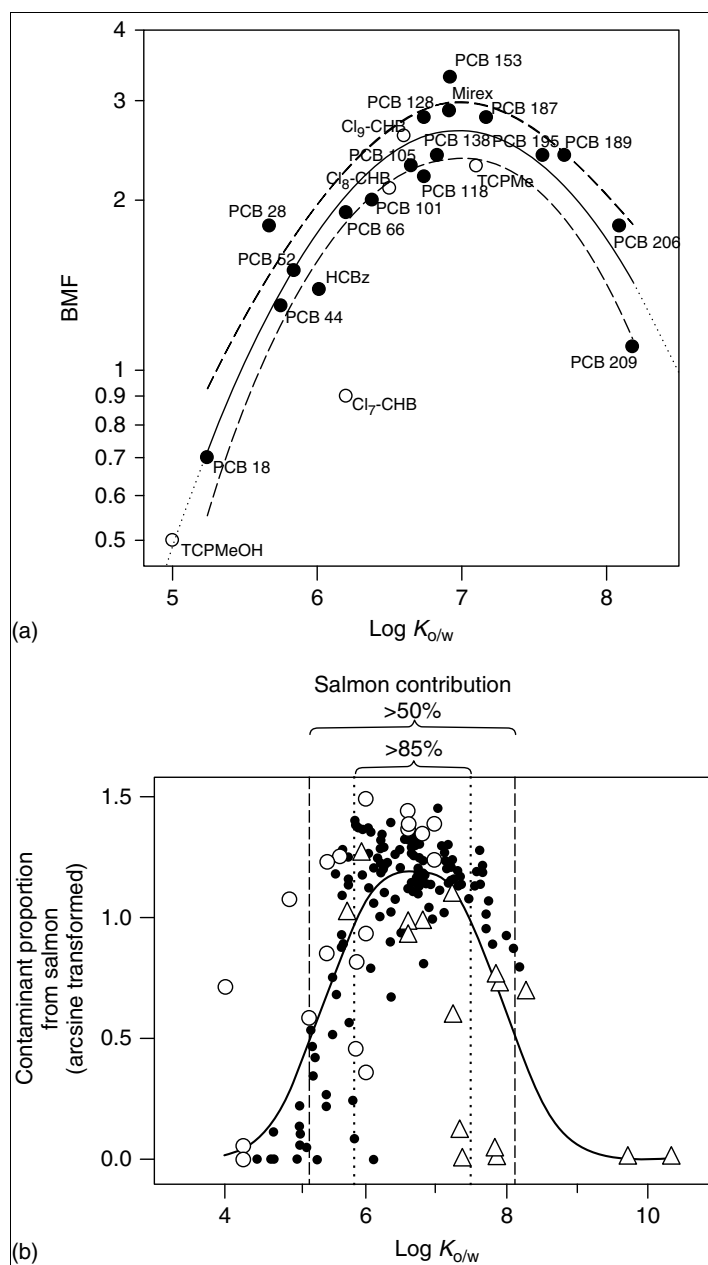


Figure 1 Persistent organic pollutants (POPs) amplify in food webs, but the extent to which this takes place varies as a function of their octanol–water partitioning coefficient ($\log K_{o/w}$). (a) Biomagnification factors (BMFs) for different POPs (including PCB congeners) in trout [Reproduced from [45]. © American Chemical Society, 2005.] (b) The delivery of different POPs from Pacific salmon to British Columbia grizzly bears is shaped by the $\log K_{o/w}$ of different POPs, (open circles: organochlorine pesticides; closed circles: PCB congeners; triangles: PBDE congeners) with an optimal range of 5.8–7.5 [Reproduced from [46]. © American Chemical Society, 2001.]

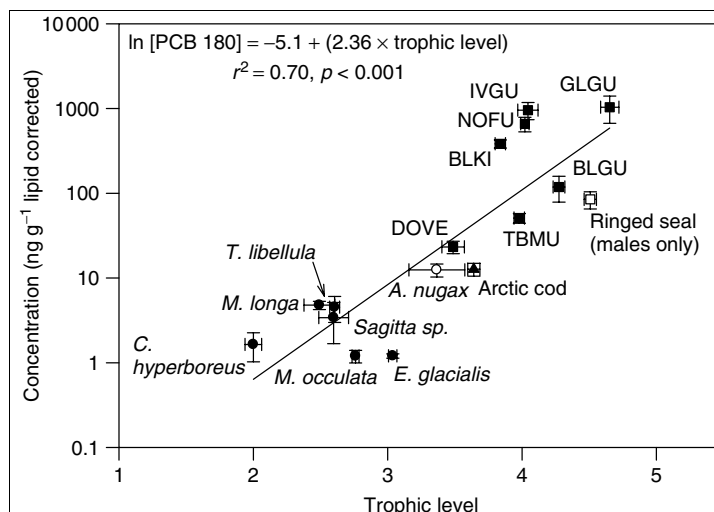


Figure 2 One of the fundamental features of persistent organic pollutants (POPs) is their tendency to biomagnify in aquatic food webs. The relationship observed between the recalcitrant PCB 180 and trophic position in an Arctic food web underscores this feature, and points to the potential for health risks at the top of the food web [46]. (Closed circles: pelagic zooplankton; open circles: benthic amphipods; triangles: Arctic cod; open squares: ringed seals; closed squares: seabirds. Acronyms: DOVE: dovekie; TBMU: thick-billed murre; BLGU: black guillemot; IVGU: ivory gull; GLGU: glaucous gull; NOFU: northern fulmar) [Reproduced from [46]. © American Chemical Society, 2001.]

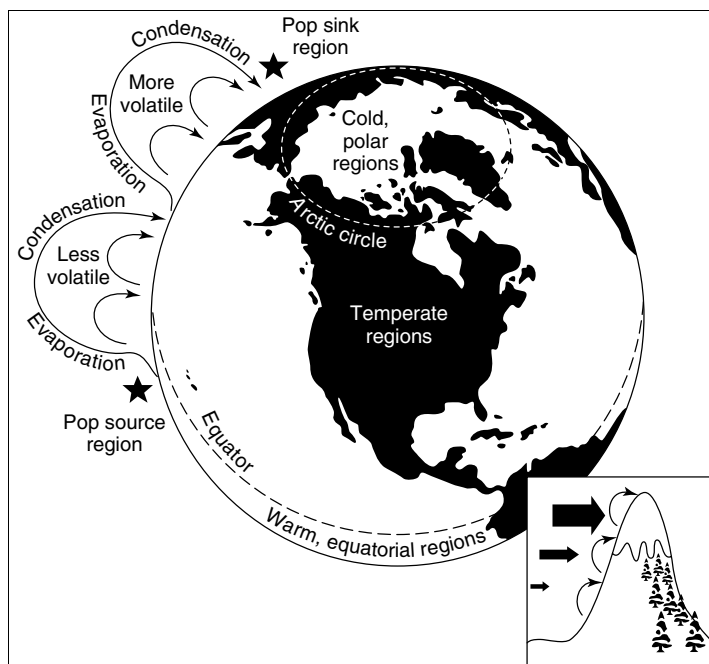


Figure 3 The movement of persistent organic pollutants away from source is governed by the interaction between their physicochemical properties and environmental matrices. The “grasshopper effect” refers to the movement of POPs into remote, typically colder, regions of the world, with a fractionation process along the way shaping the chemical signature. Altitudinal gradients have also been observed, reflecting a cold condensation at higher elevations

Assessing POP Risks: Past, Present, and Future

Every reach of the world has become contaminated with POPs. Remote food webs in alpine lakes, boreal wetlands, polar oceans, and isolated coastal environments are contaminated with POPs as a result of the persistence of these contaminants, their tendency to move large distances, and their attraction to lipids. The degree to which the widespread dispersion of POPs and subsequent contamination of aquatic food webs has impacted the health of humans and wildlife remains a fundamental concern [61, 62]. Observations of reduced immune function in breast-fed Inuit infants, impaired neurological development in infants born to sportfishing families, and endocrine disruption in seabirds and polar bears in the Arctic indicate that POPs threaten the health and viability of populations that are far removed from the origin of the chemicals [63–66].

The protection of humans and wildlife from the unwanted effects of POPs represents a daunting challenge for risk assessors, managers, and regulators. The cumulative evidence from common POP effects in humans and wildlife [67] was used as an example in support of a proposed integrated human and ecological risk assessment framework [68]. Such an example can serve as a form of “forensic” risk assessment that documents past regulatory failures, or can help to guide site-specific remediation efforts, or can contribute to new and improved paradigms for chemical design, regulation, and/or use. The extent to which such a framework could prove effective in preventing future failures remains to be evaluated.

The development and adoption of the Stockholm Convention reflect, in part, the recognition that POPs know no national boundaries, are readily transported to distant reaches, amplify in food webs that support some human and wildlife groups, and can cause adverse health effects. The failure of national regulations to consider or protect remote environments provided much of the original impetus for the Stockholm Convention. And while local or national risk assessment paradigms have afforded some guidance for the cleanup of contaminated sites, or established a basis for the regulation of certain problematic chemicals, they were ill-equipped to consider the characteristics of POPs that emerged as major health and environmental concerns in remote regions.

The “precautionary principle” has been incorporated into policy and legislative agendas of the European Union [69], and into some international treaties, such as the 1992 Rio Declaration on Environment and Development, and the Stockholm Convention on POPs. While it is viewed by some as a constraint to free-market innovation and as lacking a scientific basis [70], it has gained favor among those that point out the failure of the quantitative risk assessment paradigm to predict and/or protect the global environment [71]. The precautionary principle is based on three central elements, including the potential for harm, scientific uncertainty, and precautionary action [72]. Given the overwhelming complexities of interconnected biota and ecosystems, quantitative predictions on fate and effects for all new chemicals are a scientific impossibility. A lack of scientific certainty can be used to do one of two things: to either proceed with chemical use until unequivocal evidence emerges as a result of an adverse effect in the environment, typically decades later, or to invoke the precautionary principle by citing the uncertainty and the potential risks associated with an unknown. Invoking the precautionary principle may be perceived to constrain innovation by forcing industry to find new (and potentially costly) processes. However, its purpose is not only to prevent environmental degradation and to protect human health but also to prevent negative economic impacts “downwind” of the activity. Depending on the nature of the release, cleanup costs associated with POPs that have entered the environment can be exorbitant. For example, the Environmental Protection Agency (EPA) allocated over \$500 million in 2002 for the PCB cleanup of the Hudson River [73]. Exposure to PCBs and other toxins in the United States has been estimated to cost hundreds of billions per year [74]. Fishery closures in British Columbia due to the release of large quantities of dioxins and furans into coastal waters by pulp and paper mills prior to the onset of regulations reduced fisheries opportunities for many communities [75]. Health care costs are increasing for subsistence-oriented Inuit people suffering from elevated incidence of infections and related health problems due to POP exposure through their diet of fish and marine mammals [76]. While the precautionary principle has made a tentative foray into the risk assessment arena, numerous POPs continue to be used on a global scale. The extent to which current risk assessment methods are advanced enough to

protect the biosphere is very much unclear, something that perhaps requires a new way of thinking [77].

POPs are found in remote reaches of the world. Many are endocrine disrupting and can impact growth, brain development, reproduction, and immune function. A weight of evidence indicates that the amplification of POPs within aquatic food webs has adversely affected marine mammals, seabirds, and subsistence-oriented humans in both industrialized regions and in remote regions. Only by being more inclusive and more global in our consideration of POP risks during the lifetime of the chemical, will we truly be in apposition to prevent wanton effects.

References

- [1] Carson, R. (1962). *Silent Spring*, Houghton-Mifflin, Boston.
- [2] Hickey, J.J. & Anderson, D.W. (1968). Chlorinated hydrocarbons and eggshell changes in raptorial and fish-eating birds, *Science* **162**, 271–273.
- [3] Wiemeyer, S.N. & Porter R.D. (1970). DDE thins eggshells of captive American kestrels, *Nature* **227**, 737–738.
- [4] Lundholm, C.E. (1997). DDE-induced eggshell thinning in birds: effects of p,p'-DDE on the calcium and prostaglandin metabolism of the eggshell gland, *Comparative Biochemistry and Physiology* **118C**, 113–128.
- [5] Montgomery, J.H. (1993). *Agrochemicals Desk Reference: Environmental Data*, Lewis Publishers, Ann Arbor, p. 625.
- [6] Mackay, D., Shiu W.Y. & Ma, K.C. (1992). *Illustrated Handbook of Physical–Chemical Properties and Environmental Fate for Organic Chemicals*. Vols 1 and 2. Monoaromatic Hydrocarbons, Chlorobenzenes and PCBs. Lewis Publishers, Chelsea.
- [7] HSDB (2006). *Hazardous Substances Database*, <http://toxnet.nlm.nih.gov> (accessed 2007).
- [8] Wania, F. & Dugani, C.B. (2003). Assessing the long-range transport potential of polybrominated diphenyl ethers: a comparison of four multimedia models, *Environmental Toxicology and Chemistry* **22**(6), 1252–1261.
- [9] Oliver, B.G. & Kaiser, K.L.E. (1986). Chlorinated organics in nearshore waters and tributaries of the St. Clair River, *Water Pollution Research Journal of Canada* **21**(3), 344–350.
- [10] Kao, C.M., Chen, S.C., Liu, J.K. & Wu, M.J. (2004). Evaluation of TCDD biodegradability under different redox conditions, *Chemosphere* **44**, 1447–1454.
- [11] Sinkkonen, S. & Paasivirta, J. (2000). Degradation half-life times for PCDDs, PCDFs and PCBs for environmental fate modelling, *Chemosphere* **40**, 943–949.
- [12] Davis, J.W., Gonsior, S., Marty, G. & Ariano, J. (2005). The transformation of hexabromocyclododecane in aerobic and anaerobic soils and aquatic sediments, *Water Research* **39**, 1075–1084.
- [13] Joner, E.J. & Leyval, C. (2003). Phytoremediation of organic pollutants using mycorrhizal plants: a new aspect of rhizosphere interactions, *Agronomie* **23**, 495–502.
- [14] Manzano, M.A., Perales, J.A., Sales, D. & Quiroga, J.M. (2003). Enhancement of aerobic microbial degradation of polychlorinated biphenyl in soil microcosms, *Environmental Toxicology and Chemistry* **22**, 699–705.
- [15] Kucerova, R. & Fecko, P. (2006). Biodegradation of PAU, PCB, and NEL soil samples from the hazardous waste dump in Pozd'átky (Czech Republic), *Physicochemical Problems of Mineral Processing* **40**, 203–210.
- [16] Martineau, D., Béland, P., Desjardins, C. & Lagacé, A. (1987). Levels of organochlorine chemicals in tissues of beluga whales (*Delphinapterus leucas*) from the St. Lawrence Estuary, Québec, Canada, *Archives of Environmental Contamination and Toxicology* **16**, 137–147.
- [17] Lebeuf, M., Gouteux, B., Measures, L. & Trottier, S. (2004). Levels and temporal trends (1988–1999) of polybrominated diphenyl ethers in beluga whales (*Delphinapterus leucas*) from the St. Lawrence Estuary, Canada, *Environmental Science and Technology* **38**, 2971–2977.
- [18] Olsson, M. & Reutergårdh, L. (1986). DDT and PCB pollution trends in the Swedish aquatic environment, *Ambio* **15**, 103–109.
- [19] Kleivane, L., Skaare, J.U., Bjorge, A., De Ruiter, E. & Reijnders, P.J.H. (1995). Organochlorine pesticide residue and PCBs in harbour porpoise (*Phocoena phocoena*) incidentally caught in Scandinavian waters, *Environmental Pollution* **89**, 137–146.
- [20] Stapleton, H.M., Dodder, N.G., Kucklick, J.R., Reddy, C.M., Schantz, M.M., Becker, P.R., Gulland, F., Porter, B.J. & Wise, S.A. (2006). Determination of HBCD, PBDEs and MeO-BDEs in California sea lions (*Zalophus californianus*) stranded between 1993 and 2003, *Marine Pollution Bulletin* **52**, 522–531.
- [21] Kannan, K., Kajiwar, N., Le Boeuf, L.J.B. & Tanabe, S. (2004). Organochlorine pesticides and polychlorinated biphenyls in California sea lions, *Environmental Pollution* **131**, 425–434.
- [22] Pellettieri, M.B., Hallenbeck, W.H., Brenniman, G.R., Cailas, M. & Clark, M. (1996). PCB intake from sport fishing along the northern Illinois shore of Lake Michigan, *Bulletin of Environmental Contamination and Toxicology* **57**, 766–770.
- [23] Dewailly, E., Nantel, A., Weber, J.P. & Meyer, F. (1989). High levels of PCBs in breast milk of Inuit women from Arctic Québec, *Bulletin of Environmental Contamination and Toxicology* **43**, 641–646.
- [24] Mulvad, G., Pedersen, H.S., Hansen, J.C., Dewailly, E., Jul, E., Pedersen, M.B., Bjerregaard, P., Malcom, G.T.,

- Deguchi, Y. & Middaugh, J.P. (1996). Exposure of Greenlandic Inuit to organochlorines and heavy metals through the marine food-chain: an international study, *The Science of the total Environment* **186**, 137–139.
- [25] Mickalek, J.E., Pirkle, J.L., Needham, L.L., Patterson Jr, D.G., Caudill, S.P., Tripathi, R.C. & Mocarelli P. (2002). Pharmacokinetics of 2,3,7,8-tetrachlorodibenzo-p-dioxin in Seveso adults and veterans of operation Ranch Hand, *Journal of Exposure Analysis and Environmental Epidemiology* **12**, 44–53.
- [26] Ryan, J.J., Levesque, D., Panopio, L.G., Sun, W.F., Masuda, Y. & Kuroki, H. (1993). Elimination of polychlorinated dibenzofurans (PCDFs) and polychlorinated biphenyls (PCBs) from human blood in the Yusho and Yu-Cheng rice oil poisonings, *Archives of Environmental Contamination and Toxicology* **24**, 504–512.
- [27] Van Larebeke, N., Hens, L., Schepens, P., Covaci, A., Baeyens, J., Everaert, K., Bernheim, J.L., Vlietinck, R. & De Poorter, G. (2001). The Belgian PCB and dioxin incident of January-June 1999: exposure data and potential impact on health, *Environmental Health Perspectives* **109**, 265–273.
- [28] Koopman-Esseboom, C., Huisman, M., Touwen, B.C., Boersma, E.R., Brouwer, A., Sauer, P.J.J. & Weisglas-Kuperus, N. (1997). Newborn infants diagnosed as neurologically abnormal with relation to PCB and dioxin exposure and their thyroid-hormone status, *Developmental Medicine and Child Neurology* **39**, 785–785.
- [29] Koopman-Esseboom, C., Weisglas-Kuperus, N., De Ridder, M.A., Van der Pauw, C.G., Tuinstra, L.G.M.Th. & Sauer, P.J.J. (1996). Effects of polychlorinated biphenyl/dioxin exposure and feeding type on infants' mental and psychomotor development, *Pediatrics* **97**, 700–706.
- [30] Colborn, T., Saal, F.S.V. & Soto, A.M. (1993). Developmental effects of endocrine-disrupting chemicals in wildlife and humans, *Environmental Health Perspectives* **101**, 378–384.
- [31] Fox, G.A. (2001). Wildlife as sentinels of human health effects in the Great Lakes- St. Lawrence basin, *Environmental Health Perspectives* **109**(6), 853–861.
- [32] Van den Berg, M., Birnbaum, L., Bosveld, A.T.C., Brunstrom, B., Cook, P., Feeley, M., Giesy, J.P., Hanberg, A., Hasegawa, R., Kennedy, S.W., Kubiak, T., Larsen, J.C., Van Leeuwen, F.X.R., Liem, A.K., Nolt, C., Peterson, R.E., Poellinger, L., Safe, S.H., Schrenk, D., Tillitt, D.E., Tysklind, M., Younes, M., Waern, F. & Zacharewski, T.R. (1998). Toxic equivalency factors (TEFs) for PCBs, PCDDs, PCDFs for humans and wildlife, *Environmental Health Perspectives* **106**, 775–792.
- [33] Helle, E., Olsson, M. & Jensen, S. (1976). PCB levels correlated with pathological changes in seal uteri, *Ambio* **5**, 261–263.
- [34] De Guise, S., Martineau, D., Beland, P. & Fournier, M. (1995). Possible mechanisms of action of environmental contaminants on St. Lawrence beluga whales (*Delphinapterus leucas*), *Environmental Health Perspectives* **103**(4), 73–77.
- [35] Delong, R.L., Gilmartin, W.G. & Simpson, J.G. (1973). Premature births in California sea lions: association with high organochlorine pollutant residue levels, *Science* **181**, 1168–1170.
- [36] Mortensen, P., Bergman, A., Bignert, A., Hansen, H.-J., Härkönen, T. & Olsson, M. (1992). Prevalence of skull lesions in harbor seals (*Phoca vitulina*) in Swedish and Danish museum collections: 1835–1988, *Ambio* **21**, 520–524.
- [37] Gilbertson, M., Kubiak, T., Ludwig, J. & Fox, G.A. (1991). Great Lakes embryo mortality, edema, and deformities syndrome (GLEMEDS) in colonial fish-eating birds: similarity to chick-edema disease, *Journal of Toxicology and Environmental Health* **33**, 455–520.
- [38] Grasman, K.A., Fox, G.A., Scanlon, P.F. & Ludwig, J.P. (1996). Organochlorine-associated immunosuppression in prefledgling caspian terns and herring gulls from the Great Lakes: an ecoepidemiological study, *Environmental Health Perspectives* **104**, 829–842.
- [39] Ross, P.S., De Swart, R.L., Addison, R.F., Van Loveren, H., Vos, J.G. & Osterhaus, A.D.M.E. (1996). Contaminant-induced immunotoxicity in harbour seals: wildlife at risk? *Toxicology* **112**, 157–169.
- [40] Luebke, R.W., Hodson, P.V., Faisal, M., Ross, P.S., Grasman, K.A. & Zelikoff, J.T. (1997). Aquatic pollution-induced immunotoxicity in wildlife species, *Fundamental and Applied Toxicology* **37**, 1–15.
- [41] Reijnders, P.J.H. (1986). Reproductive failure in common seals feeding on fish from polluted coastal waters, *Nature* **324**, 456–457.
- [42] Brouwer, A., Reijnders, P.J.H. & Koeman, J.H. (1989). Polychlorinated biphenyl (PCB)-contaminated fish induces vitamin A and thyroid hormone deficiency in the common seal (*Phoca vitulina*), *Aquatic Toxicology* **15**, 99–106.
- [43] De Swart, R.L., Ross, P.S., Vos, J.G. & Osterhaus, A.D.M.E. (1996). Impaired immunity in harbour seals (*Phoca vitulina*) exposed to bioaccumulated environmental contaminants: review of a long-term study, *Environmental Health Perspectives* **104**(4), 823–828.
- [44] Fisk, A.T., Norstrom, R.J., Cymbalisty, C.D. & Muir, D.C.G. (1998). Dietary accumulation and depuration of hydrophobic organochlorines: bioaccumulation parameters and their relationship with the octanol/water partition coefficient, *Environmental Toxicology and Chemistry* **17**, 951–961.
- [45] Christensen, J.R., MacDuffee, M., Macdonald, R.W., Whitticar, M. & Ross, P.S. (2005). Persistent organic pollutants in British Columbia grizzly bears: consequence of divergent diets, *Environmental Science and Technology* **39**, 6952–6960.
- [46] Fisk, A.T., Hobson, K.A. & Norstrom, R.J. (2001). Influence of chemical and biological factors on trophic transfer of persistent organic pollutants in the Northwater Polynya marine food web, *Environmental Science & Technology* **35**, 732–738.

- [47] Muir, D.C.G., Omelchenko, A., Grift, N.P., Savoie, D.A., Lockhart, L., Wilkinson, P. & Brunskill, G.J. (1996). Spatial trends and historical deposition of polychlorinated biphenyls in Canadian midlatitude and Arctic lake sediments, *Environmental Science and Technology* **30**, 3609–3617.
- [48] Muir, D.C.G., Ford, C.A., Grift, N.P., Metner, D.A. & Lockhart, L. (1990). Geographic variation of chlorinated hydrocarbons in burbot (*Lota lota*) from remote lakes and rivers in Canada, *Archives of Environmental Contamination and Toxicology* **19**, 530–542.
- [49] Ter Schure, A.F.H., Larsson, P., Merilä, J. & Jönsson, K.I. (2002). Latitudinal fractionation of polybrominated diphenyl ethers and polychlorinated biphenyls in frogs (*Rana temporaria*), *Environmental Science and Technology* **36**, 5057–5061.
- [50] Ross, P.S., Jeffries, S.J., Yunker, M.B., Addison, R.F., Ikononou, M.G. & Calambokidis, J. (2004). Harbour seals (*Phoca vitulina*) in British Columbia, Canada, and Washington, USA, reveal a combination of local and global polychlorinated biphenyl, dioxin, and furan signals, *Environmental Toxicology and Chemistry* **23**, 157–165.
- [51] Blais, J.M., Schindler, D.W., Sharp, M., Braekevelt, E., Lafrenière, M., McDonald, K., Muir, D.C.G. & Strachan, W.M.J. (2001). Fluxes of semivolatile organochlorine compounds in Bow Lake, a high-altitude, glacier-fed, subalpine lake in the Canadian rocky mountains, *Limnology and Oceanography* **46**, 2019–2031.
- [52] Vives, I., Grimalt, J.O., Catalan, J., Rosseland, B.O. & Battarbee, R.W. (2004). Influence of age and altitude in the accumulation of organochlorine compounds in fish from high mountain lakes, *Environmental Science and Technology* **38**, 690–698.
- [53] Lichota, G.B., McAdie, M. & Ross, P.S. (2004). Endangered Vancouver Island marmots (*Marmota vancouverensis*): sentinels of atmospherically delivered contaminants to British Columbia, Canada, *Environmental Toxicology and Chemistry* **23**, 402–407.
- [54] Li, Y.-F., Macdonald, R.W., Jantunen, L.M.M., Harner, T., Bidleman, T.F. & Strachan, W.M.J. (2002). The transport of β -hexachlorocyclohexane to the western Arctic Ocean: a contrast to α -HCH, *Science of the total Environment* **291**, 229–246.
- [55] Blais, J.M., Kimpe, L.E., McMahon, D., Keatley, B.E., Mallory, M.L., Douglas, M.S.V. & Smol, J.P. (2005). Arctic seabirds transport marine-derived contaminants, *Science* **309**, 445–445.
- [56] Wania, F. & Mackay, D. (1995). A global distillation model for persistent organic chemicals, *Science of the total Environment* **160**, 211–232.
- [57] Sandala, G.M., Sonne-Hansen, C., Dietz, R., Muir, D.C.G., Valters, K., Bennett, E.R., Born, E.W. & Letcher, R.J. (2004). Hydroxylated and methyl sulfone PCB metabolites in adipose and whole blood of polar bear (*Ursus maritimus*) from East Greenland, *Science of the Total Environment* **331**, 125–141.
- [58] Boon, J.P., Eijgenraam, F. & Everaarts, J.M. (1989). A structure-activity relationship (SAR) approach towards metabolism of PCBs in marine animals from different trophic levels, *Marine Environment Research* **27**, 159–176.
- [59] Ross, P.S. (2000). Marine mammals as sentinels in ecological risk assessment, *Human and Ecological Risk Assessment* **6**, 29–46.
- [60] Fox, G.A. (1992). Epidemiological and pathobiological evidence of contaminant-induced alterations in sexual development in free-living wildlife, *The Wildlife/Human Connection*, Princeton Science Publishing, Princeton, pp. 147–158.
- [61] Dewailly, E., Ayotte, P., Bruneau, S., Laliberté, C., Muir, D.C.G. & Norstrom, R.J. (1993). Inuit exposure to organochlorines through the aquatic food chain in Arctic Québec, *Environmental Health Perspectives* **101**, 618–620.
- [62] Ross, P.S., Ellis, G.M., Ikononou, M.G., Barrett-Lennard, L.G. & Addison, R.F. (2000). High PCB concentrations in free-ranging Pacific killer whales, *Orcinus orca*: effects of age, sex and dietary preference, *Marine Pollution Bulletin* **40**, 504–515.
- [63] Braathen, M., Derocher, A.E., Wiig, Ø., Sørmo, E.G., Lie, E., Skaare, J.U. & Jenssen, B.M. (2004). Relationships between PCBs and thyroid hormones and retinol in female and male polar bears, *Environmental Health Perspectives* **112**, 826–833.
- [64] Dallaire, F., Dewailly, E., Muckle, G., Vézina, C., Jacobson, S.W., Jacobson, J.L. & Ayotte, P. (2004). Acute infections and environmental exposure to organochlorines in Inuit infants from Nunavik, *Environmental Health Perspectives* **112**, 1359–1364.
- [65] Fisk, A.T., de Wit, C.A., Wayland, M., Kuzyk, Z.Z., Burgess, N., Letcher, R., Braune, B., Norstrom, R., Polischuk Blum, S., Sandau, C., Lie, E., Larsen, H.J.S., Skaare, J.U. & Muir, D.C.G. (2005). An assessment of the toxicological significance of anthropogenic contaminants in Canadian arctic wildlife, *Science of the Total Environment* **351–352**, 57–93.
- [66] Jacobson, J.L. & Jacobson, S.W. (1996). Intellectual impairment in children exposed to polychlorinated biphenyls in utero, *New England Journal of Medicine* **335**, 783–789.
- [67] Ross, P.S. & Birnbaum, L.S. (2003). Integrated human and ecological risk assessment: a case study of persistent organic pollutants (POPs) in humans and wildlife, *Human and Ecological Risk Assessment* **9**(1), 303–324.
- [68] Suter II, G.W., Vermeire, T., Munns Jr, W.R. & Sekizawa, J. (2003). Framework for the integration of health and ecological risk assessment., *Human and Ecological Risk Assessment* **9**(1), 281–301.
- [69] de Bruijn, J., Hansen, B. & Munn, S. (2003). Role of the precautionary principle in the EU risk assessment process on industrial chemicals, *Pure and Applied Chemistry* **75**, 2535–2541.

- [70] Hathcock, J.N. (2000). The precautionary principle – an impossible burden of proof for new products, *AgBioForum* **3**, 255–258.
- [71] Kriebel, D., Tickner, J., Epstein, P., Lemons, J., Levins, R., Loechler, E.R., Quinn, M., Rudel, R., Schettler, T. & Stoto, M. (2001). The precautionary principle in environmental science, *Environmental Health Perspectives* **109**, 871–876.
- [72] Myers, N. (2002). The precautionary principle puts values first, *Bulletin of Science Technology and Society* **22**, 210–219.
- [73] United States Environmental Protection Agency (USEPA) (2002). *Hudson River PCBs Site Record of Decision*, EPA ID: NYD980763841, New York.
- [74] Weiss, B. (2000). Vulnerability of children and the developing brain to neurotoxic hazards, *Environmental Health Perspectives* **108**, 375–381.
- [75] Hagen, M.E., Colodey, A.G., Knapp, W.D. & Samis, S.C. (1997). Environmental response to decreased dioxin and furan loadings from British Columbia coastal pulp mills, *Chemosphere* **34**, 1221–1229.
- [76] Hansen, J.C. (2000). Environmental contaminants and human health in the Arctic, *Toxicology Letters* **112–113**, 119–125.
- [77] Thornton, J. (2000). Beyond risk: an ecological paradigm to prevent global chemical pollution, *International Journal of Occupational and Environmental Health* **6**, 318–330.

Related Articles

Environmental Health Risk

Environmental Monitoring

What are Hazardous Materials?

PETER S. ROSS AND PAUL B.C. GRANT

Pesticide Risk

Risk assessment of pesticides is intriguing for a number of reasons. Pesticides are substances that are made to deliberately kill or deter species that humans have designated as pests. And, these organism-killing substances are intentionally disseminated into the environment. Because of these rather unusual aspects, pesticide risk assessment is relatively straightforward, yet, at the same time, complex.

As with risk assessment for many technologies, pesticide risk assessment primarily has been, and will continue to be, driven by government regulatory frameworks. For pesticides, risk assessment throughout most of the world largely follows the “Red Book Paradigm” [1]. This paradigm represents a formalized basis for the objective evaluation of risk in which assumptions and uncertainties are clearly considered and presented. Risk assessment typically utilizes a tiered modeling approach extending from deterministic models (Tier 1) based on conservative assumptions to probabilistic models (Tier 4) using refined assumptions [2]. In risk assessment, conservative assumptions in lower-tier assessments represent overestimates of effect and exposure; therefore, the resulting quantitative risk values typically are conservative and err on the side of safety.

Like other types of environmental health risk assessment (*see* **Environmental Health Risk Assessment**), pesticide risk assessment typically relies on the first principle that risk is a function of effect and exposure. Effects assessment includes identifying the hazard (i.e., toxin) and understanding the dose–response relationships (*see* **Dose–Response Analysis**). Because of extensive global regulatory requirements, there is usually a large and varied amount of toxicity data for pesticidal active ingredients. For example, mammalian studies (to determine potential human health effects) include acute oral toxicity (rat), acute dermal toxicity, acute inhalation toxicity (rat), primary eye irritation (rabbit), primary dermal irritation, dermal sensitization, acute delayed neurotoxicity (hen), 90-day feeding (rodent), 90-day feeding (nonrodent), 21-day dermal toxicity, 90-day subchronic dermal toxicity, 90-day inhalation (rat), 90-day neurotoxicity (hen), 90-day neurotoxicity (rat), chronic feeding (rodent), chronic feeding (nonrodent), oncogenicity (rat), oncogenicity (mouse), teratogenicity (rat),

teratogenicity (rabbit), two-generation reproduction (rat), chronic feeding/oncogenicity (rat), gene mutation, structural chromosome aberration, and other genotoxic effects. Similarly, for ecological receptors, toxicity studies include those mentioned above for mammals, acute avian oral, acute avian dietary, avian reproductive, fish, freshwater invertebrate, estuarine invertebrate (six species), fish early life stage, honey bee, and plant. From these toxicity studies, acute, subchronic, and chronic no-effect and effect levels are established, which are then compared to exposure estimates and used to characterize risk (*see* **Cancer Risk Evaluation from Animal Studies**).

Data for estimating exposures are also extensive and include measured physicochemical parameters as well as environmental fate results. Examples include octanol–water partition, hydrolysis, photodegradation in water, soil, and air, aerobic and anaerobic soil and aquatic metabolism, leaching, volatility, foliar and soil residue dissipation, groundwater fate, spray drift, and accumulation in crops, livestock, fish, and processed food (*see* **Air Pollution Risk; Environmental Risk Assessment of Water Pollution**).

Exposure assessment for pesticides is unlike many other environmental risk assessments in that normally it is not based on exposure at a specific location, but rather it relies on generic (lower-tier) exposure assumptions representing many locations. This is especially true for exposure assessments conducted by regulatory agencies during registration and reregistration activities. The rationale for this is to account for nearly all of the scenarios in which the pesticide can be applied, especially in potentially sensitive environments. There are several environmental fate models that are used to predict exposures, such as PRZM–EXAMS, GENEEC, and AgDRIFT.

Finally, risks are characterized by integrating both effect and exposure data. This is typically done by determining a ratio (often termed the *risk quotient*) between the estimated environmental concentration (EEC) and the appropriate toxic endpoint. The ratio is then compared to preexisting regulatory or other levels of concern. For example, if the EEC of a pesticide is 3 ppm in surface water and the chronic no-observed-effect-concentration for the waterflea, *Daphnia magna*, is 5 ppm, the risk quotient is 0.6. The current US Environmental Protection Agency level of concern for chronic exposure to the waterflea

is 1.0, so the value of 0.6 is below the level of concern.

In addition to generic risk assessments designed for regulatory decision-making at the time of registration, many site-specific risk assessments for pesticides have been conducted. These assessments, which can be initiated based on regulatory action, usually involve actual measurements of environmental concentrations and may also involve species-specific toxicity data. Recent advances have involved probabilistic modeling of both toxicity and exposure data using techniques such as Monte Carlo simulations. These simulations allow for a random sampling technique in which each input variable in the risk model is sampled repeatedly from a range of possible values based on the probability distribution of each variable. Then, the variability for each input is propagated into the output of the model, so that the model output reflects the probability of values that could occur. This is advantageous because it allows the decision maker to evaluate probabilities of risk occurrences and sensitivities of input values to the overall output.

Regulatory policies around the world continue to evolve in response to changes in pesticide and risk assessment technologies. For example, in the United States, risk assessments for plant-incorporated pesticides (e.g., Bt cotton) derived through genetic engineering are based on many conventional pesticide requirements, but also include many other effect and exposure data specific to gene and pesticide protein expression in a plant. Data include identification of gene source, molecular characterization of insert, copy number, gene integrity, function, specificity, mode of action, protein expression levels, protein

digestibility, homology to known allergens, acute oral toxicity, ecotoxicity, crop morphology and yield, and food and feed composition analysis.

Another major change in pesticide risk assessment in the past 10 years has been the emergence of aggregate and cumulative risk concepts. This is most evident in the passage of the US Food Quality Protection Act (FQPA) in 1996. Aggregate risk is the total risk to receptors (such as humans) from all pathways and routes of exposure to a single pesticidal active ingredient. Cumulative risk is the total risk to receptors from all sources of all pesticides that have a common mechanism of toxicity. These concepts and FQPA, in general, are intended to ensure a comprehensive, integrated approach to risk assessment and management using a consistent, health-based standard.

References

- [1] U.S. National Research Council (1983). *Risk Assessment in the Federal Government: Managing the Process*, National Academy Press, Washington, DC.
- [2] Society for Environmental Toxicology and Chemistry (SETAC) (1994). *Aquatic Dialogue Group Pesticide Risk Assessment and Mitigation*, SETAC Foundation for Environmental Education, Pensacola.

Related Articles

Environmental Risks

What are Hazardous Materials?

ROBERT K.D. PETERSON

Pharmaceuticals in the Environment

Following their use in human and veterinary medicine, significant quantities of drugs end up in the environment. Many pharmaceuticals are not readily degraded in the environment and can affect biological processes in the soil, surface water, and groundwater, and thereby be passed back into the food chain for humans and animals. Even though intensive scientific research has been conducted in this area for the last 10 years, much is still unknown.

Entry into the Environment

Since 30–95% of drugs used to treat humans and animals are later excreted, drug residues end up in the environment [1]. There they pollute surface water, soil, groundwater, and air (*see Water Pollution Risk; Soil Contamination Risk; Air Pollution Risk*) [2–5].

Drugs excreted following human use end up with municipal wastewater in sewage treatment plants and then in surface water (Figure 1). Another source of surface water contamination results from aquaculture in which pharmaceuticals are added to the water along with fish food. Following the field application of untreated manure and slurry, pharmaceuticals excreted by livestock end up in the soil and water (Figure 1). Particularly, slurry applied in agricultural areas is a source of groundwater pollution. *Via* dust particles from animal housing (manure dust, feed dust) drugs can also be passed into the air. In addition to slurry and manure, sewage sludge is also used as fertilizer, and pharmaceuticals from sewage treatment could seep into the soil and groundwater. Likewise, *via* bank filtration of surface water, drugs in municipal wastewater can get into groundwater. Drugs that land in the soil can be taken up by plants, get into the groundwater, or drain into a nearby stream. The movement of pharmaceuticals into the groundwater depends on multiple factors, such as soil type, soil moisture, soil pore distribution, and how the soil is used [6]. The risk of pharmaceutical penetration into groundwater increases with increasing solubility of the substance, reduced adsorption, and decreased degradation rate.

Degradation in the Environment

The elimination of pharmaceuticals in the environment occurs mainly by microbial degradation but it also occurs by hydrolysis, photolysis (UV light), and adsorption. Whether the degradation occurs rapidly or slowly depends on several factors, especially the chemical structure of the substance and the properties of the medium (matrix) in which the pharmaceuticals occur. Most pharmaceuticals are lipophilic and difficult to degrade, which leads to potential bioaccumulation and persistence in the environment.

In sewage treatment plants, the elimination rate of pharmaceuticals is between 7 and 99%, the average being 60%. Pharmaceuticals end up in streams that receive treated water. Biodegradation occurs more slowly in the streams than in sewage treatment plants owing to a lower stream concentration of microorganisms. This leads to increased danger of wider contamination. To some degree, these drugs are adsorbed in the sewage sludge, which is used mainly as fertilizer. The biodegradation half-life of various antibiotics lies between 10 and 104 days in surface water [7, 8].

Most medical drugs are slow to degrade in soil, so that repeated spread of slurry, manure, or sludge can result in an accumulation of pharmaceutical residues [9, 10]. Over an 80-day period sarafloxacin (fluoroquinolone) is hardly mineralized. Virginiamycin has a half-life of 87–173 days in soil. Pharmaceuticals can be taken up by plants [11].

Occurrence in the Environment

Pharmaceuticals that may be detected in wastewater and surface water include lipid regulators, analgesics, anti-inflammatory agents, β blockers, sympathomimetics, antiepileptic drugs, bronchodilators, cytostatic chemotherapy agents, and ethinylestradiol [2]. In water, following sewage treatment, drugs such as antibiotics can be detected at the microgram per liter level [2, 12]. Antibiotic concentrations up to $6\text{ }\mu\text{g l}^{-1}$ have been reported [13].

In samples of soil in agricultural areas regularly fertilized with liquid manure, tetracycline and chlortetracycline residues up to $198.7\text{ }\mu\text{g kg}^{-1}$ and $7.3\text{ }\mu\text{g kg}^{-1}$, respectively, can be detected. Since only 4.0 mg kg^{-1} tetracycline and 0.1 mg kg^{-1} chlortetracycline could be detected in the liquid manure,

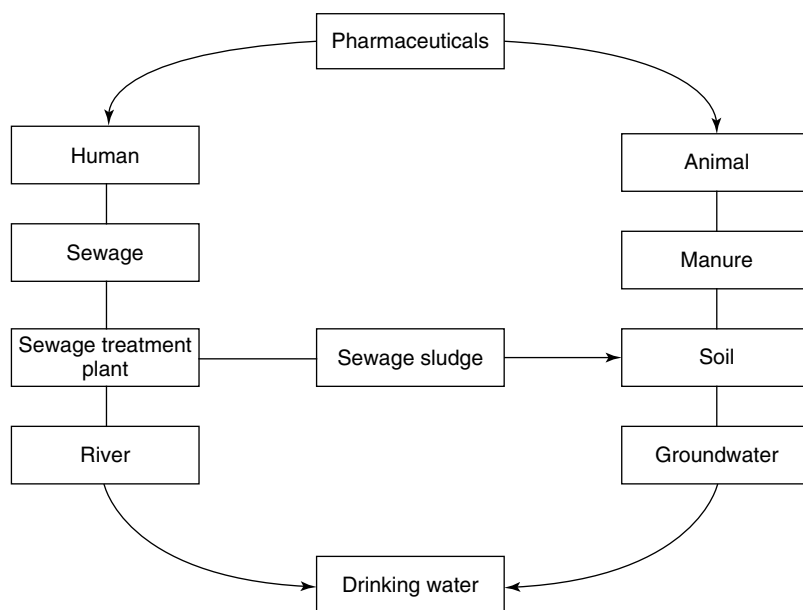


Figure 1 Passage of pharmaceuticals into the terrestrial and aquatic environment following their use in human and veterinary medicine

the higher concentration in the soil is regarded as an indication of cumulative effects [14]. Fluoroquinolones, often prescribed in human medicine, have been detected in soil samples of fields on which sewage sludge had been applied [15].

At present, knowledge of pharmaceuticals in groundwater and drinking water is quite limited, but there are indications of the infiltration of drugs into these resources [16, 17]. A few drug residues have been detected at the nanogram per liter level in drinking water [18]. Studies conducted near large swine and poultry facilities in the United States also detected antimicrobials (tetracyclines, sulfonamides, β -lactams, macrolides, and fluoroquinolones) in surface water and groundwater [19], with concentrations in some cases being over $100 \mu\text{g l}^{-1}$.

Risk for the Environment

Pharmaceuticals that land in the environment can be detrimental to microorganisms, plants, insects, fish, and other life forms (Figure 2). Of special interest are detrimental effects in the areas of toxicity, genotoxicity, and development of resistance [20–23]. The so-called aquatic and terrestrial

environment is particularly at risk. Concern should be directed to possible past environmental pollution from human drugs such as antibiotics, psychopharmaceuticals, chemotherapy cytostatic drugs, and pharmaceuticals with hormonal effects. In animal production, antibiotics and antiparasitic drugs, including coccidiostats, are of particular relevance to the environment.

Antibiotics, which land in the environment, may impact microbial biodegradation in the soil and surface water, including water treatment processes in sewage treatment plants. Likewise, decomposition of dung and production of biogas may be influenced. Antiparasitic drugs can heavily damage the biocenosis in manure, slurry, water, and soil, which may, for example, lead to delayed biodegradation. Following the spread of liquid manure on soil, antibiotics may be taken up by the roots of plants, and thereby arrive in the food chain.

A particular problem is the development of bacterial resistance in the environment. Resistance genes can be carried over to other bacteria. The route of the carryover of antibiotics or resistant bacteria from animal to humans *via* meat products has been well examined; other paths *via* water, soil, and plants are less well examined [24].

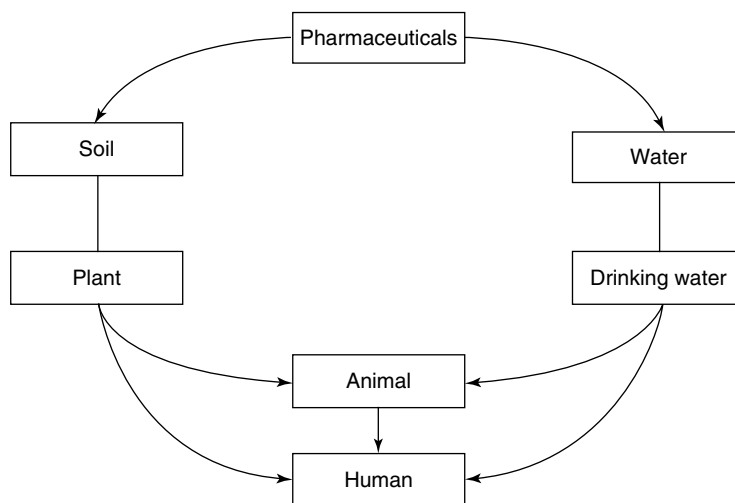


Figure 2 Environmental pollution by pharmaceuticals and routes of threat to humans and animals

Minimizing the Risk

In the United States, 60–80% of the antibiotics used in animal husbandry are used as growth promoters. To avoid the development of resistant and cross-reacting strains, as well as gene toxic and carcinogenic effects, a number of pharmacologically active feed additives have been withdrawn in the European Union since 1997. As of January 1, 2006, the use of antibiotic performance enhancers is no longer allowed throughout Europe according to Commission Regulation (EC) No 183/2003 Article 11 Paragraph 2. As a result, the use of antibiotics has decreased in food animals, but not in companion animals [25], which are often treated with antibiotics used in human medicine.

Since 1993/1995, environmental risk assessment must be established before a drug developed for human medicine or for veterinary medicine is allowed to be authorized [26, 27]. In an attempt to assess environmental risk during phase I of drug trials, the predicted environmental concentration (PEC) must be determined. If specified concentrations are exceeded ($100 \mu\text{g kg}^{-1}$ in the soil or $1 \mu\text{g l}^{-1}$ in water) then risk assessment phase II takes place. In this phase, the PEC is compared to the predicted no effects concentration (PNEC). If the ratio PEC to PNEC is greater than 1, additional evaluations must be conducted.

Conclusions

Effects of pharmaceuticals in the environment have been described; particularly those antibiotics, which have an inhibitory effect on environmental bacteria, can lead to development of resistance in bacteria. At present the damage to terrestrial and aquatic life is not yet critical. Concentrations of drug residues in water consumed by humans and animals are far below therapeutic doses. Possible long term effects resulting from permanent consumption are possible, however [28].

References

- [1] Halling-Sorensen, B., Nors Nielsen, S., Lanzky, P.F., Ingerslev, F., Holten Lutzhoft, H.C. & Jorgensen, S.E. (1998). Occurrence, fate and effects of pharmaceutical substances in the environment-a review, *Chemosphere* **36**, 357–393.
- [2] Ternes, T.A. (1998). Occurrence of drugs in German sewage treatment plants and rivers, *Water Research* **32**, 3245–3260.
- [3] Hektoen, H., Berge, J.A., Samuelsen, O. & Yndestad, M. (1995). Resistance of antibacterial agents in marine sediments, *Aquaculture* **133**, 175–184.
- [4] Boxall, A.B., Fogg, L.A., Blackwell, P.A., Kay, P., Pemberton, E.J. & Croxford, A. (2004). Veterinary medicines in the environment, *Reviews of Environmental Contamination and Toxicology* **180**, 1–91.
- [5] Hamscher, G., Pawelzick, H.T., Sczesny, S., Nau, H. & Hartung, J. (2003). Antibiotics in dust originating from

- a pig-fattening farm: a new source of health hazard for farmers? *Environmental Health Perspectives* **111**, 1590–1594.
- [6] Rabolle, M. & Spliid, N.H. (2000). Sorption and mobility of metronidazole, olaquinox, oxytetracycline and tylosin in soil, *Chemosphere* **40**, 715–722.
- [7] Ingerslev, F., Toräng, L., Loke, M.-L., Halling-Sørensen, B. & Nyholm, N. (2001). Primary biodegradation of veterinary antibiotics in aerobic and anaerobic surface water simulation systems, *Chemosphere* **44**, 865–872.
- [8] Alexy, R., Kumpel, T. & Kümmerer, K. (2004). Assessment of degradation of 18 antibiotics in the closed bottle test, *Chemosphere* **57**, 505–512.
- [9] Marengo, J.R., Kok, R.A., Velageleti, R. & Stamm, J.M. (1997). Aerobic degradation of 14C-sarafloxacin hydrochloride in soil, *Environmental Toxicology and Chemistry* **16**, 462–471.
- [10] Weerasinghe, C.A. & Towner, D. (1997). Aerobic biodegradation of virginiamycin in soil, *Environmental Toxicology and Chemistry* **16**, 1873–1876.
- [11] Ingerslev, F. & Halling-Sørensen, B. (2001). Biodegradability of metronidazole, olaquinox, and tylosin and formation of tylosin degradation products in aerobic soil-manure slurries, *Ecotoxicology and Environmental Safety* **48**, 311–320.
- [12] Mcardell, C.S., Molnar, E., Suter, M.J.F. & Giger, W. (2003). Occurrence and fate of macrolide antibiotics in wastewater treatment plants and in Glatt Valley watershed, Switzerland, *Environmental Science and Technology* **37**, 5479–5486.
- [13] Hirsch, R., Ternes, T., Haberer, K. & Kratz, K.L. (1999). Occurrence of antibiotics in the aquatic environment, *Science of the Total Environment* **225**, 109–118.
- [14] Hamscher, G., Sczesny, S., Hoper, H. & Nau, H. (2002). Determination of persistent tetracycline residues in soil fertilized with liquid manure by high-performance liquid chromatography with electrospray ionization tandem mass spectrometry, *Analytical Chemistry* **74**, 1509–1518.
- [15] Golet, E.M., Xifra, I., Siegrist, H., Alder, A.C. & Giger, W. (2003). Environmental exposure assessment of fluoroquinolone antibacterial agents from sewage to soil, *Environmental Science and Technology* **37**, 3243–3249.
- [16] Batt, A.L., Snow, D.D. & Aga, D.S. (2006). Occurrence of sulfonamide antimicrobials in private water wells in Washington County, Idaho, USA, *Chemosphere* **64**, 1963–1971.
- [17] Carlson, J.C. & Mabury, S.A. (2006). Dissipation kinetics and mobility of chlortetracycline, tylosin, and monensin in an agricultural soil in Northumberland County, Ontario, Canada, *Environmental Toxicology and Chemistry* **25**, 1–10.
- [18] Jones, O.A., Lester, J.N. & Voulvoulis, N. (2005). Pharmaceuticals: a threat to drinking water? *Trends in Biotechnology* **23**, 163–167.
- [19] Campagnolo, E.R., Johnson, K.R., Karpati, A., Rubin, C.S., Kolpin, D.W., Meyer, M.T., Esteban, J.E., Currier, R.W., Smith, K., Thu, K.M. & McGeehin, M. (2002). Antimicrobial residues in animal waste and water resources proximal to large-scale swine and poultry feeding operations, *Science of the Total Environment* **299**, 89–95.
- [20] Al-Ahmad, A., Daschner, F.D. & Kümmerer, K. (1999). Biodegradability of cefotiam, ciprofloxacin, meropenem, penicillin G, and sulfamethoxazole and inhibition of waste water bacteria, *Archives of Environmental Contamination and Toxicology* **37**, 158–163.
- [21] Baguer, A.J., Jensen, J. & Krogh, P.H. (2000). Effects of the antibiotics oxytetracycline and tylosin on soil fauna, *Chemosphere* **40**, 751–757.
- [22] Halling-Sørensen, B. (2000). Algal toxicity of antibacterial agents used in intensive farming, *Chemosphere* **40**, 731–739.
- [23] Jorgensen, S.E. & Halling-Sørensen, B. (2000). Drugs in the environment, *Chemosphere* **40**, 691–699.
- [24] Van den Bogaard, A.E. & Stobberingh, E.E. (2000). Epidemiology of resistance to antibiotics. Links between animals and humans, *International Journal of Antimicrobial Agents* **14**, 327–335.
- [25] Odensvik, K., Grave, K. & Greko, C. (2001). Antibacterial drugs prescribed for dogs and cats in Sweden and Norway 1990–1998, *Acta Veterinaria Scandinavica* **42**, 189–198.
- [26] Koschorreck, J., Koch, C. & Ronnefahrt, I. (2002). Environmental risk assessment of veterinary medicinal products in the EU-a regulatory perspective, *Toxicology Letters* **131**, 117–124.
- [27] Straub, J.O. (2002). Environmental risk assessment for new human pharmaceuticals in the European Union according to the draft guideline/discussion paper of January 2001, *Toxicology Letters* **135**, 231–237.
- [28] Webb, S., Ternes, T., Gibert, M. & Olejniczak, K. (2003). Indirect human exposure to pharmaceuticals via drinking water, *Toxicology Letters* **142**, 157–167.

Related Articles

Ecological Risk Assessment

Environmental Health Risk

Hazardous Waste Site(s)

GERD SCHLENKER

Players in a Decision

Decisions inevitably involve people; and they may have distinct roles. The simplest decisions involve just one person, the decision maker. He or she alone provides the necessary contextual knowledge, expresses all the judgments, performs the analyses, and makes the decision. In practice, this seldom happens. Rather a group of decision makers, e.g. a management board or government department, aware that they have a decision to make, ask an analyst or, maybe, a consulting company to work with accountants, scientists, engineers, and other subject experts to gain relevant information to formulate the problem, gather and analyze data, and advise on the decision. In doing so, they consult a range of stakeholders. Thus, many will contribute to the process that leads to a decision; indeed, many may be party to the decision making (*see Group Decision*).

Roles Played in a Decision Analysis

Decision makers “own the problem”, although that does not necessarily mean that they know exactly what the problem is. In many contexts, problems related to decision arise when the decision maker becomes aware of a raft of issues that need resolving, but may be unable even to describe what a resolution might look like. For instance, a planning authority may be aware that flood risk is increasing because of climate change, yet be unclear on what the options are, whether there are other potentially relevant trends, what their objectives are, etc. However, they do know that it is their task to explore and resolve the issues. By “owning a problem”, we mean that the decision makers have the responsibility for resolving it, they have the authority to assign the necessary resources to do so, and also have the accountability for the effects of its resolution. If they lack any one of these, their abilities and motivation to deal with the issues will then be severely limited.

Decision analysis assumes at the fundamental level of the underlying modelling that all uncertainty and value judgements are provided by the decision makers: indeed, in precise terms by one decision maker [1] (*see Subjective Expected Utility; Decision Analysis*). When groups of decision makers are

involved, the situation becomes more complex. There is a need to combine their views and conclusions, and reach a consensus. Moreover, in many decision-making groups, the relative power and importance of the members of the group can differ and be further moderated by external political pressures, making some members more like experts or stakeholders than decision makers.

Experts provide economic, marketing, scientific, and other professional advice that is used to assess the likelihood of the many eventualities. Experts also provide contextual knowledge. The information that experts impart is used in the modeling and forecasting of outcomes. In principle, their advice should inform the decision makers’ subjective probabilities; however, this is not an easy process in practice. Experts can differ in their advice, so a balance has to be made. Some experts are good at encoding their knowledge into the uncertainty judgments – subjective probabilities (*see Subjective Probability*) – needed in an analysis; others are less so. There is a large literature on expert judgment, which deals with these issues and offers practical approaches for incorporating their input into a decision or risk analysis [1–4] (*see Expert Judgment*).

Stakeholders share or, at the very least, perceive that they share in the impacts arising from a decision. They have a claim, therefore, that their perceptions and values should be taken into account. The decision makers are stakeholders, if only by virtue of their accountabilities; but stakeholders are not necessarily decision makers. Within a decision analysis, their views may be taken into account within the multiattribute utility modeling [5]. It is becoming common in some areas of decision making with impacts on a wide range of stakeholders to involve them at many points in the analytic process, even if the ultimate decision is reserved for a small group of decision makers [6–8] (*see Decision Conferencing/Facilitated Workshops; Public Participation*).

Whereas experts provide contextual knowledge, *decision (or risk) analysts* have expertise in the process of decision (or risk) analysis. Their role is to manage and conduct the analysis, and is increasing these days to smooth the communication and build a shared understanding between the decision makers, stakeholders, and experts [9]. In the latter role, they are often called *facilitators*.

Discussion

Of course, it should be realized that, in practice, the distinction between roles drawn above is seldom clear. Decision makers are stakeholders and may – indeed, usually will – have relevant expertise. Similar experts can be stakeholders and *vice versa*. Ideally, however, at least one decision analyst should be dissociated from the issues, being neither an expert nor stakeholder, and certainly not a decision maker. Such independence is necessary if he or she is to manage a decision analysis in an even-handed manner.

We have concentrated on differing roles in a decision analysis; but naturally similar roles exist within a risk analysis. *Risk owners* are people who are responsible and accountable for dealing with a risk and have the authority to do so, thus their roles correspond with that of decision makers. Experts, stakeholders, and risk analysts have the obvious correspondences with the roles described above. One additional role is that of *risk manager*, who is responsible for the process of monitoring and managing the risk, and thus the protection of risk owners. Essentially, he or she is responsible for the implementation of the risk-management policy. We have not discussed implementation within decision analysis in which there are similar roles, usually, relating to tasks assigned to “middle” or operational management. However, such roles occur after the analysis has concluded.

References

- [1] French, S. & Rios Insua, D. (2000). *Statistical Decision Theory*, Kendall's Library of Statistics, Arnold.
- [2] Bedford, T. & Cooke, R. (2001). *Probabilistic Risk Analysis: Foundations and Methods*, Cambridge University Press, Cambridge.

- [3] Cooke, R.M. (1991). *Experts in Uncertainty*, Oxford University Press, Oxford.
- [4] Cooke, R.M. (2007). Expert judgement studies, *Reliability Engineering and System Safety* (in press).
- [5] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Trade-offs*, John Wiley & Sons, New York.
- [6] Gregory, R.S., Fischhoff, B. & McDaniels, T. (2005). Acceptable input: using decision analysis to guide public policy deliberations, *Decision Analysis* **2**(1), 4–16.
- [7] Renn, O. (1999). A model for an analytic-deliberative process in risk management, *Environmental Science and Technology* **33**(18), 3049–3055.
- [8] French, S. (2003). The challenges in extending the MCDA paradigm to e-democracy, *Journal of Multi-Criteria Decision Analysis* **12**(2–3), 61–233.
- [9] French, S. (2007). Web-enabled strategic GDSS, e-democracy and Arrow's theorem: a Bayesian perspective, *Decision Support Systems* **43**, 1476–1484.

Related Articles

Behavioral Decision Studies

Risk Attitude

Supra Decision Maker

SIMON FRENCH

Point Source Modeling

The term *point-source modeling* arises in a number of different contexts. In this article, we consider the specific problem of investigating spatial variation in disease risk around one or more preidentified point sources of pollution (*see Spatial Risk Assessment*). A well-known UK example is the investigation of an apparent excess of childhood leukemia cases in the vicinity of the Sellafield nuclear reprocessing plant [1–3]. The qualifying phrase “preidentified” distinguishes our topic from statistical methods used in conducting an unrestricted search for possible disease clusters throughout a prespecified geographical region [4–7], and from methods intended to characterize the general tendency for cases of a disease to exhibit spatial clustering [8, 9] or spatial variation in risk [10, 11].

We first describe an idealized point process model for the spatial distribution of individual cases of disease, then show how this model leads to case–control methods of analysis that are robust to certain kinds of departure from the idealized model. This model-based approach encourages a focus on estimating interpretable parameters using principled statistical methods, and avoids the need for *ad hoc* tests for the existence of a point-source effect.

The Idealized Model

Suppose that the population at risk can be represented as a Poisson process [12] with spatially varying intensity $\lambda(x)$. Suppose also that the risk of disease varies spatially, and that cases occur independently. Then, the cases constitute a second Poisson process with intensity $\lambda_1(x) = c_1 \lambda(x) \rho(x)$, where c_1 is the absolute risk of the disease in question and $\rho(x)$ is the spatially varying relative risk, which is the quantity of scientific interest [13]. Now suppose that controls are selected at random from the population at risk. Then, the controls constitute a third Poisson process with intensity $\lambda_0(x) = c_0 \lambda(x)$, where c_0 is the sampling fraction for the controls. It follows from standard properties of the Poisson process that, given the locations x_i of n cases and m controls, the case–control labels constitute a set of independent Bernoulli trials with the probability that an event at

x is a case given by

$$p(x) = \frac{\lambda_1(x)}{\{\lambda_0(x) + \lambda_1(x)\}} = \frac{\alpha \rho(x)}{\{1 + \alpha \rho(x)\}} \quad (1)$$

where $\alpha = c_1/c_0$ is determined by the sampling fraction for the controls. Typically, neither α nor $\lambda(x)$ is of scientific interest. Equation (1) shows how the conditioning on the case and control locations eliminates the potentially awkward function $\lambda(x)$. Furthermore, the validity of equation (1) requires no assumptions about the population at risk; its only requirements are that the controls are sampled randomly, and that cases occur independently. Hence, the model is not valid for infectious diseases, but is a reasonable starting point for investigating spatial variation in environmentally determined, noninfectious diseases.

When the controls are defined specifically to exclude cases, the argument leading to equation (1) needs to be modified accordingly but the conclusion that the parameters of $\rho(\cdot)$ can be estimated using a binary regression model for the case–control labels, still holds.

Inference

We now consider data in the form of n case locations $x_i : i = 1, \dots, n$ and m control locations $x_i : i = n + 1, \dots, n + m$. Under the stated assumptions, the log-likelihood associated with equation (1) is

$$L(\alpha, \rho) = n \log(\alpha) + \sum_{i=1}^n \log \rho(x_i) - \sum_{i=1}^{n+m} \log \{1 + \alpha \rho(x_i)\} \quad (2)$$

Parametric Models

Given a parametric specification for $\rho(x)$, inference can be carried out using standard likelihood-based methods [14]. For a single point source at location x_0 , a typical parametric model would specify $\rho(x)$ to be a nonincreasing function of the distance, $u(x)$ say, between x and x_0 , with $\rho(u) \rightarrow 1$ as $u \rightarrow \infty$. Possible choices include the following:

Near and far [15]

$$\rho(u) = \begin{cases} 1 + \gamma & : u < \delta \\ 1 & : u \geq \delta \end{cases} \quad (3)$$

Gaussian plume [16, 17]

$$\rho(u) = 1 + \gamma \exp\{-(u/\phi)^2\} \quad (4)$$

Diggle *et al.* [18] embed both equations (3) and (4) into a single, three-parameter model that combines a constant elevated risk within distance δ of the source, with a smoothly decaying risk at larger distances.

Whatever functional form is chosen for $\rho(\cdot)$, three possible extensions to the basic model are the following:

1. Adjustments for measured risk factors

If the data include the values $z_{ik} : k = 1, \dots, r; i = 1, \dots, n + m$ of each of r known or hypothesized risk factors, these can be accommodated by replacing the single parameter α in equations (1) and (2) by a linear regression term, $\alpha_i = \sum_{k=1}^r z_{ik} \beta_k$ [17]. Note that the β_k are to be interpreted as relative risk parameters.

2. Multiple sources

If there are multiple point sources at locations $x_{0k} : k = 1, \dots, r$, we write u_{ik} for the distance between x_i and x_{0k} . Then, the single term $\rho(x_i)$ in equations (1) and (2) can be replaced by a product of r terms, $\rho(u_{i1})\rho(u_{i2})\dots\rho(u_{ir})$, or by any other combination of the $\rho(u_{ik})$, if this were thought more reasonable in a particular application; see, for example, Diggle and Rowlingson [17] or Dunn *et al.* [19].

3. Directional effects

Lawson [20] and Dunn *et al.* [19] allow point-source effects to depend on orientation as well as on distance. For example, a directional version of the Gaussian plume model used in Dunn *et al.* [19] specifies

$$\rho(u, t) = 1 + \gamma \exp\{-(u \exp(-\kappa \cos(t - \theta))/\phi)^2\} \quad (5)$$

where u and t denote the distance and orientation of x relative to x_0 . The additional parameters κ and θ represent the strength and orientation of the directional effect, respectively. Their combined effect is to generate a comet-shaped plume that distinguishes

between upstream and downstream effects along the axis of orientation.

Nonparametric Models

A nonparametric model specifies $\rho(x)$ in equations (1) and (2) as an arbitrary function of $u(x)$. The model then becomes a generalized additive logistic model [21]. Statistical methods for fitting models of this kind are described in Hastie and Tibshirani [21] and in Wood [22]. Standard likelihood-based methods fail in the nonparametric setting, but penalized versions that constrain $\rho(\cdot)$ to be a smoothly varying function have been shown to give satisfactory results.

References

- [1] Black, D. (1984). *Investigation of the Possible Increased Incidence of Cancer in West Cumbria*, Report of independent advisory group. HMSO, London.
- [2] Gardner, M.J. & Winter, P.D. (1984). Mortality in Cumberland during 1959–1978 with reference to cancer in young people around Windscale, *Lancet* **1**, 216–217.
- [3] Cook-Mozaffari, P., Darby, S., Doll, R., Forman, D., Hermon, C. & Pike, M.C. (1989). Geographical variation in mortality from leukaemia and other cancers in England and Wales in relation to proximity to nuclear installations, 1969–78, *British Journal of Cancer* **59**, 476–485.
- [4] Openshaw, S., Charlton, M., Wymer, C. & Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets, *International Journal of Geographical Information Systems* **1**, 335–358.
- [5] Openshaw, S., Craft, A.W., Charlton, H. & Birch, J.M. (1988). Investigation of leukaemia clusters by the use of a geographical analysis machine, *Lancet* **1**, 272–273.
- [6] Besag, J. & Newell, J. (1991). The detection of clusters in rare diseases, *Journal of the Royal Statistical Society, Series A* **154**, 143–155.
- [7] Kulldorff, M. (1999). Spatial scan statistics: models, calculations, and applications, in *Scan Statistics and Applications*, J. Glaz, ed, Birkhauser, Boston, pp. 303–322.
- [8] Cuzick, J. & Edwards, R. (1990). Spatial clustering for inhomogeneous populations (with discussion), *Journal of the Royal Statistical Society, Series B* **52**, 73–104.
- [9] Diggle, P.J. & Chetwynd, A.G. (1991). Second-order analysis of spatial clustering for inhomogeneous populations, *Biometrics* **47**, 1155–1163.
- [10] Bithell, J.F. (1990). An application of density estimation to geographical epidemiology, *Statistics in Medicine* **9**, 691–701.
- [11] Kelsall, J.E. & Diggle, P.J. (1998). Spatial variation in risk: a nonparametric binary regression approach, *Applied Statistics* **47**, 559–573.

- [12] Cox, D.R. & Isham, V. (1980). *Point Processes*, Chapman & Hall, London.
- [13] Woodward, M. (1999). *Epidemiology: Study Design and Data Analysis*, Chapman & Hall, London.
- [14] Pawitan, Y. (2001). In *All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, Oxford.
- [15] Stone, R.A. (1988). Investigations of excess environmental risks around putative sources: statistical problems and a proposed test, *Statistics in Medicine* **7**, 649–660.
- [16] Diggle, P.J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point, *Journal of the Royal Statistical Society, Series A* **153**, 349–362.
- [17] Diggle, P.J. & Rowlingson, B.S. (1994). A conditional approach to point process modelling of raised incidence, *Journal of the Royal Statistical Society, Series A* **157**, 433–440.
- [18] Diggle, P., Elliott, P., Morris, S. & Shaddick, G. (1997). Regression modelling of disease risk in relation to point sources, *Journal of the Royal Statistical Society, Series A* **160**, 491–505.
- [19] Dunn, C.E., Bhopal, R.S., Cockings, S., Walker, D., Rowlingson, B. & Diggle, P. (2006). Advancing insights into environment-health relationships: a multidisciplinary approach to understanding Legionnaires' disease, *Health and Place* **13**, 677–690.
- [20] Lawson, A. (1993). On the analysis of mortality events associated with a prespecified fixed point, *Journal of the Royal Statistical Society, A* **156**, 363–377.
- [21] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*, Chapman & Hall, London.
- [22] Wood, S. (2006). *Generalized Additive Models: An Introduction with R*, Chapman & Hall, London.

Related Articles

Spatiotemporal Risk Analysis

PETER J. DIGGLE

Polychlorinated Biphenyls

Polychlorinated Biphenyls: Synthesis, Applications, and Environmental Contamination

Polychlorinated biphenyls (PCBs) were widely used industrial chemicals that are readily synthesized in bulk quantities by the iron-catalyzed chlorination of biphenyl. Depending on the degree of chlorination, commercial PCBs vary from liquids to solids [1, 2]. These products have excellent industrial properties, which include chemical stability, dielectric properties, and inflammability, and were extensively used as lubricants, organic diluents, adhesives, heat transfer fluids, and dielectric fluids for capacitors and transformers. PCBs were marketed by several manufacturers worldwide and were graded according to their chlorine content (% by weight). For example, the major North American PCBs called *Aroclors* were manufactured by Monsanto Chemical Co., and Aroclors 1248 and 1260 contained 48 and 60% chlorine by weight. Another key property of commercial PCBs is that these compounds are complex mixtures of isomers and congeners (Figure 1). Since biphenyl contains five unsubstituted carbon atoms on each ring, there are a total of 209 individual PCBs (congeners) and multiple isomers within the mono and nona chlorobiphenyl group of compounds (Figure 1). High-resolution analysis of PCB mixtures has identified individual congeners in most commercial PCB formulations [3–5] and in environmental samples, which usually resemble an Aroclor 1260 pattern with 2,2', 4, 4', 5, 5'-hexachlorobiphenyl as the dominant individual PCB congener.

Environmental problems associated with PCBs were highlighted in the 1960s when these compounds were first identified in the gas chromatographic analysis for the pesticide dichlorodiphenyltrichloroethane (DDT) and its persistent metabolite dichlorodiphenyl-dichloroethylene (DDE) in various environmental media [6]. Jensen reported that PCBs were detected in environmental and human samples, and it was apparent that these compounds were present in almost every component of the global ecosystem. PCBs and DDE/DDT were among the first persistent organic pollutants (POPs) identified in the environment. The

use and manufacture of these compounds were restricted or banned in the 1970s, and there is a worldwide restriction on the production and application of all POPs. Thus, some of the major chemical properties such as lipophilicity and stability that contributed to the widespread use of PCBs ultimately lead to their downfall. Their use patterns coupled with careless disposal practices resulted in their introduction into the environment where many PCB congeners are highly persistent and bioaccumulate in fish, wildlife, and humans [7–9].

Environmental and Human Health Effects of PCBs

The effects of environmental mixtures of PCBs on fish and wildlife species has been primarily investigated by comparing and correlating levels of exposure with various adverse responses in these populations. The problem associated with these studies is that PCBs are found in the environment along with other POPs including organochlorine pesticides, DDT/DDE, polychlorinated dibenzo-*p*-dioxins (PCDDs), polychlorinated dibenzofurans (PCDFs), and other chlorinated aromatic hydrocarbons. PCBs along with organochlorine compounds have been associated with reproductive problems in fish and wildlife, and there is concern regarding their role in compromising immune responses among species that inhabit polar regions where PCB levels remain high [10, 11]. Occupational exposure to relatively high levels of PCBs results in several responses including increase in 17-hydroxycorticosteroid excretion; γ -glutamyl transpeptidase activity; lymphocyte levels; skin diseases such as chloracne, folliculitis, and dermatitis; and serum cholesterol. It also leads to hepatomegaly, elevated blood pressure, and decreased serum bilirubin. These effects are highly variable among different exposed groups, and similar variability has been observed for different tumor types among various occupationally exposed cohorts [12].

There is also considerable variability in correlative hypothesis-based as well as random studies correlating exposure levels of PCBs with various outcomes. Studies in Europe and North America have extensively investigated the potential effects of *in utero* exposure to PCBs on neurodevelopment and neurobehavioral deficits in the offspring, since these effects correlate with laboratory animal

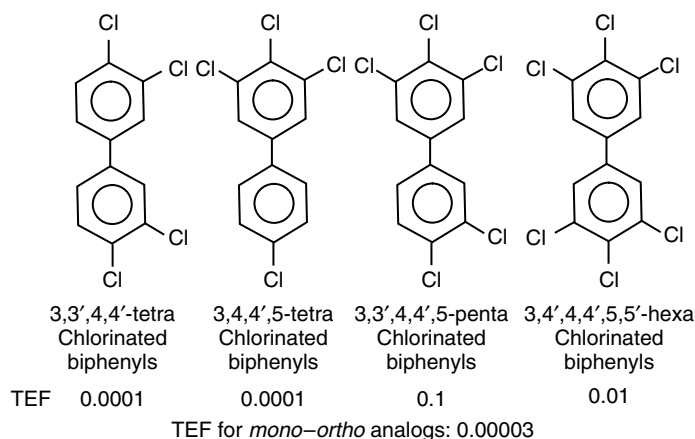


Figure 1 Structures of non-ortho substituted PCB congeners containing four or more chloro substituents and their corresponding TEF values

studies with rodent models. Jacobson and coworkers reported that, in offspring of women who consumed PCB-contaminated fish, there was a correlation between higher serum and cord blood PCB levels with decreased performance on Fagan's visual recognition memory test [13] and the Brazelton neonatal behavioral assessment scale [14]. At age 4, levels of umbilical cord serum PCBs were associated with decreased verbal and memory subtests of the McCarthy memory scales, and some deficits were also observed in children at 11 years of age [15]. Subsequent studies in the Netherlands and Germany also correlated *in utero* exposure to PCBs with neurodevelopment and behavioral deficits [16–20]. However, a recent report by the Centers for Disease Control [21] showed that relative levels of *in utero* exposure to PCBs were not correlated with a decreased intelligence quotient in offspring, even though background PCB levels were similar to those observed in the European cohorts. The inconsistencies may be related to several factors and suggest that adverse human health impacts associated with exposure to low background levels of PCBs are unresolved and require further validation.

Risk Assessment of PCB Mixtures

Risk assessment of PCB mixtures has been a formidable problem because the composition of the commercial mixtures is different from "environmental" PCB mixtures identified in different extracts.

Since the toxicity of mixtures is dependent on their individual components and concentrations as well as interactions, risk assessment of PCB mixtures requires analytical methods capable of separating and quantitating individual PCB congeners and information on the quantitative structure–toxicity relationships among PCB congeners. Mullin and coworkers first reported the high-resolution capillary gas chromatographic separation of all 209 synthetic PCB congeners [3], and subsequent improvements have made it possible to quantitate individual PCB congeners in various PCB mixtures [4, 5]. Structure–toxicity relationships have identified one class of PCB congeners that exhibit activities similar to those described for 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD), namely, binding to the aryl hydrocarbon receptor (AhR) and induction of cytochrome P4501A1 (CYP1A1) and other AhR-mediated biochemical and toxic responses. Non-ortho-substituted coplanar PCBs, such as 3,3',4,4'-tetrachlorobiphenyl (TCB), 3,3',4,4',5-pentachlorobiphenyl, 3,3',4,4',5,5'-hexachlorobiphenyl, and 3,4,4',5-TCB, are substituted in both para and at least two meta positions and exhibit TCDD-like activity (Figure 2). A second group of mono-ortho-substituted analogs of the coplanar also exhibit similar activity with reduced potency [22–24]. Risk assessment of complex mixtures of PCBs, PCDDs, and PCDFs has focused on individual compounds within these mixtures that bind and activate the AhR. It has been assumed that these compounds induce additive responses that are

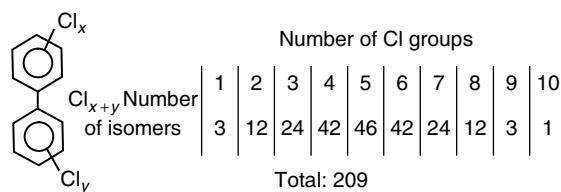


Figure 2 A summary of the number of PCB isomers and congeners

related to their relative potencies and concentrations. Quantitative structure–activity relationships among coplanar and mono-ortho PCBs, PCDDs, and PCDFs have provided potency or toxic equivalency factors (TEFs) of individual congeners relative to TCDD, and Figure 2 summarizes the latest World Health Organization TEFs for various PCB congeners [25]. TEF values have been selected by an expert panel from among a range of TEF values for each compound and the values can be used to calculate 2,3,7,8-tetrachlorodibenzo-*p*-dioxin equivalents (TEQs) from any mixture of PCBs (or PCDDs/PCDFs) based on the following equation:

$$\text{TEQ} = \sum [\text{PCB}]_i \times \text{TEF}_i \quad (1)$$

where $[\text{PCB}]_i$ and TEF_i represent the concentration and TEF value for individual PCB congeners in a mixture. The TEF/TEQ scheme has been developed for estimating potential exposures to or emissions of TEQs associated with mixtures of PCBs, PCDDs, and PCDFs. However, estimating health effect risks from the data is still problematic and other criticisms of this approach include nonadditive interactions with other compounds including PCBs.

Thus, risk assessment of the “TCDD-like” PCB congeners (Figure 2) has been developed and is routinely used for estimating the overall TEQs associated with PCBs in various mixtures of POPs. However, there is also evidence that other PCB congeners such as ortho-substituted analogs also contribute to the hazards and risks of PCB mixtures and methodologies for hazard and risk assessment of these compounds have not been developed.

Acknowledgments

The financial assistance of the National Institutes of Health (P42-ES04917, P42-E09106) and the Texas Agriculture

Experimental Station is gratefully acknowledged. S. Safe is a Sid Kyle Professor of Toxicology.

References

- [1] Hutzinger, O., Safe, S. & Zitko, V. (1974). *The Chemistry of PCBs*, CRC Press, Boca Raton.
- [2] De Voogt, P. & Brinkman, U.A.T. (1989). Production, properties and usage of polychlorinated biphenyls, in *Halogenated Biphenyls, Terphenyls, Naphthalenes, Dibenzodioxins and Related Products*, R.D. Kimbrough & A.A. Jensen, eds, Elsevier-North Holland, Amsterdam, pp. 3–45.
- [3] Mullin, M.D., Pochini, C.M., McCrindle, S., Romkes, M., Safe, S. & Safe, L. (1984). High-resolution PCB analysis: the synthesis and chromatographic properties of all 209 PCB congeners, *Environmental Science and Technology* **18**, 468–476.
- [4] Schulz, D.E., Petrick, G. & Duinker, J.C. (1989). Complete characterization of polychlorinated biphenyl congeners in commercial Aroclor and Clophen mixtures by multidimensional gas chromatography-electron capture detection, *Environmental Science and Technology* **23**, 852–859.
- [5] Frame, G.M., Wagner, R.E., Carnahan, J.C., Brown Jr, J.F., May, R.J., Smullen, L.A. & Bedard, D.L. (1996). Comprehensive, quantitative, congener-specific analyses of eight Aroclors and complete PCB congener assignments on DB-1 capillary GC columns, *Chemosphere* **33**, 603–623.
- [6] *New Scientist* (1966). Report of a new chemical hazard **32**, 621.
- [7] Risebrough, R.W., Rieche, P., Herman, S.G., Peakall, D.B. & Kirven, M.N. (1968). Polychlorinated biphenyls in the global ecosystem, *Nature* **220**, 1098–1102.
- [8] Kutz, F.W., Strassman, S.C. & Sperling, J.F. (1979). Survey of selected organochlorine pesticides in the general population of the United States: fiscal years 1970–1975, *Annals of the New York Academy of Sciences* **320**, 60–68.
- [9] Noren, K. & Meironyte, D. (2000). Certain organochlorine and organobromine contaminants in Swedish human milk in perspective of past 20–30 years, *Chemosphere* **40**, 1111–1123.
- [10] Dallaire, F., Dewailly, E., Muckle, G. & Ayotte, P. (2003). Time trends of persistent organic pollutants and heavy metals in umbilical cord blood of Inuit infants born in Nunavik (Quebec, Canada) between 1994 and 2001, *Environmental Health Perspectives* **111**, 1660–1664.
- [11] Weber, K. & Goerke, H. (2003). Persistent Organic Pollutants (POPs) in antarctic fish: levels, patterns, changes, *Chemosphere* **53**, 667–678.
- [12] Safe, S. (2007). Polychlorinated biphenyls, in *Environmental and Occupational Medicine*, W. Rom, ed, Lippincott, Williams & Wilkins, Philadelphia, pp. 1203–1212.

- [13] Jacobson, S.W., Fein, G.G., Jacobson, J.L., Schwartz, P.M. & Dowler, J.K. (1985). The effect of PCB exposure on visual recognition memory, *Child Development* **56**, 853–860.
- [14] Jacobson, J.L., Jacobson, S.W., Fein, G.G., Schwartz, P.M. & Dowler, J.K. (1984). Prenatal exposure to an environmental toxin: a test of the multiple effects model, *Developmental Psychology* **20**, 523–532.
- [15] Jacobson, J.L. & Jacobson, S.W. (1996). Intellectual impairment in children exposed to polychlorinated biphenyls in utero, *New England Journal of Medicine* **335**, 783–789.
- [16] Koopman-Esseboom, C., Morse, D.C., Weisglas-Kuperus, N., Lutkeschipholt, I.J., van der Paauw, C.G., Tuinstra, L.G., Brouwer, A. & Sauer, P.J. (1994). Effects of dioxins and polychlorinated biphenyls on thyroid hormone status of pregnant women and their infants, *Pediatric Research* **36**, 468–473.
- [17] Huisman, M., Koopman-Esseboom, C., Fidler, V., Hadders-Algra, M., van der Paauw, C.G., Tuinstra, L.G., Weisglas-Kuperus, N., Sauer, P.J., Touwen, B.C. & Boersma, E.R. (1995). Perinatal exposure to polychlorinated biphenyls and dioxins and its effect on neonatal neurological development, *Early Human Development* **41**, 111–127.
- [18] Patandin, S., Lanting, C.I., Mulder, P.G., Boersma, E.R., Sauer, P.J. & Weisglas-Kuperus, N. (1999). Effects of environmental exposure to polychlorinated biphenyls and dioxins on cognitive abilities in Dutch children at 42 months of age, *Journal of Pediatrics* **134**, 33–41.
- [19] Vreugdenhil, H.J., Lanting, C.I., Mulder, P.G., Boersma, E.R. & Weisglas-Kuperus, N. (2002). Effects of prenatal PCB and dioxin background exposure on cognitive and motor abilities in Dutch children at school age, *Journal of Pediatrics* **140**, 48–56.
- [20] Walkowiak, J., Wiener, J.A., Fastabend, A., Hein-zow, B., Kramer, U., Schmidt, E., Steingruber, H.J., Wundram, S. & Winneke, G. (2001). Environmental exposure to polychlorinated biphenyls and quality of the home environment: effects on psychodevelopment in early childhood, *Lancet* **358**, 1602–1607.
- [21] Needham, L.L., Barr, D.B., Caudill, S.P., Pirkle, J.L., Turner, W.E., Osterloh, J., Jones, R.L. & Sampson, E.J. (2005). Concentrations of environmental chemicals associated with neurodevelopmental effects in U.S. Population, *Neurotoxicology* **26**, 531–545.
- [22] Parkinson, A., Robertson, L., Safe, L. & Safe, S. (1981). Polychlorinated biphenyls as inducers of hepatic microsomal enzymes: effects of di ortho substitution, *Chemico-Biological Interactions* **31**, 1–12.
- [23] Bandiera, S., Safe, S. & Okey, A.B. (1982). Binding of polychlorinated biphenyls classified as either phenobarbitone-, 3-methylcholanthrene-, or mixed-type inducers to cytosolic Ah receptor, *Chemico-Biological Interactions* **39**, 259–277.
- [24] Safe, S. (1994). Polychlorinated biphenyls (PCBs): environmental impact, biochemical and toxic responses, and implications for risk assessment, *CRC Critical Reviews in Toxicology* **24**, 87–149.
- [25] Van den Berg, M., Birnbaum, L.S., Denison, M., DeVito, M., Farland, W., Feeley, M., Fiedler, H., Hakansson, H., Hanberg, A., Haws, L., Rose, M., Safe, S., Schrenk, D., Tohyama, C., Tritscher, A., Tuomisto, J., Tysklind, M., Walker, N. & Peterson, R.E. (2006). The 2005 World Health Organization reevaluation of human and mammalian toxic equivalency factors for dioxins and dioxin-like compounds, *Toxicological Sciences* **93**, 223–241.

Related Articles

Environmental Hazard

Environmental Risks

Persistent Organic Pollutants

What are Hazardous Materials?

STEPHEN SAFE AND FEI WU

Potency Estimation

A common goal in environmental risk assessment is description of a hazardous agent's toxic or otherwise-harmful effects, often for comparison with others agents of related structure or exposure. Summary measures are most useful in this regard, as these describe the potency of an agent or some product of an agent on the organism or system under study. A classic example of such a measure is the median effective dose (described below). Using such a measure, a risk assessor can make comparative statements about a series of hazardous agents as applied to a specific assay or system. When no significant response is evidenced after exposure to the agent, the potency is set equal to some boundary value that indicates no potent response. (Zero is common, although other boundary values are possible depending on the measure used to describe potency.) When a significant exposure-response relationship is present, however, some descriptive statistic that quantifies the impact of the exposure is calculated from the data.

Statistical issues dealing with estimation of potency have a long history, ranging back to bioassays of pesticides, herbicides, and other poisons [1, 2]. A number of quantitative summaries can be used to measure or define potency, each based on a slightly different feature of the observed dose-response. We consider a few of these in this article.

Median Effective Dose

A basic measure used to summarize dose-response is the *median effective stimulus* or *median effective dose*, denoted as ED_{50} . If the environmental agent is given as a concentration, we use the term *median effective concentration*, or EC_{50} . In either case, this is defined as the amount of agent required to produce a response at the median (50%) level [3]. This can be 50% of the maximum possible response in a quantal response experiment, or 50% of the control response in a decreasing-response experiment, etc. For example, suppose we measure quantal response

data as the proportion of times organisms respond (e.g., die) when exposed to a toxic stimulus. Then, since the proportions are bounded between 0 and 1, ED_{50} is the dose at which the response equals 0.50. Note here that ED_{50} is an inverse measure of potency: a higher ED_{50} suggests a weaker agent, since a greater dose is required to produce the same median result.

Obviously, ED_{50} depends upon the outcome variable under study, and the terminology has developed to indicate this. For instance, when simple lethality is the common outcome, ED_{50} is more specifically the *median lethal dose*, or LD_{50} . Note that the ED_{50} or LD_{50} is typically derived from experiments conducted over a range of dose levels for a fixed period of time. It is also possible to fix the dose level in an experiment and vary the exposure times. Then this could be used to derive a *median lethal time*, or LT_{50} , associated with the exposure. Issues of statistical estimation and testing for such quantities can be addressed using survival analytic methods [4].

An extensive literature has developed on the use of ED_{50} , EC_{50} , LD_{50} , etc.; see the early works by Irwin [5] and Fieller [6], and more contemporary discussions by Bhattacharya and Kong [7], Stukel [8], or our own work [9]. Our goal here is to distill features of the ED_{50} useful for potency estimation, particularly as they pertain to stimulus-response risk analysis.

For example, suppose we observe quantal response data where the parent distribution is binomial with sample size parameter n , and with response probability $\pi(x)$ varying over a dose x . Assume that an underlying form of the dose-response is the linear logistic form $\pi(x) = 1/(1 + \exp\{-\beta_0 - \beta_1 x\})$ or the linear probit form $\pi(x) = \Phi(\beta_0 + \beta_1 x)$, where $\Phi(z)$ is the standard normal cumulative density function. In these cases, ED_{50} is found by setting $\pi(x)$ equal to 1/2, and solving for x ; i.e., ED_{50} is the value of x that solves $1/(1 + \exp\{-\beta_0 - \beta_1 x\}) = 0.5$ or $\Phi(\beta_0 + \beta_1 x) = 0.5$, respectively. In either case, assuming $\beta_1 \neq 0$ one finds $ED_{50} = -\beta_0/\beta_1$. Next, suppose that from the observed data we calculate point estimates b_0 and b_1 for β_0 and β_1 , along with their attendant standard errors $se[b_0]$ and $se[b_1]$. Then, an estimate of ED_{50} is $\widehat{ED}_{50} = -b_0/b_1$, and large-sample $1 - \alpha$ confidence limits on the ratio $-\beta_0/\beta_1$ are based on Fieller's theorem [6, 10]. We find

$$\widehat{ED}_{50} + \frac{\gamma}{1-\gamma} \left\{ \widehat{ED}_{50} + \frac{\widehat{\sigma}_{01}}{se^2[b_1]} \right\} \pm \frac{z_{\alpha/2}}{(1-\gamma)|b_1|} \sqrt{se^2[b_0] + 2\widehat{\sigma}_{01}E\widehat{D}_{50} + se^2[b_1]E\widehat{D}_{50}^2 - \gamma \left\{ se^2[b_0] - \frac{c_{01}^2}{se^2[b_1]} \right\}} \quad (1)$$

where $\gamma = z_{\alpha/2}^2 se^2[b_1]/b_1^2$ measures departure from symmetry in the distribution of $\widehat{ED}_{50}$ (as γ approaches zero, it indicates greater symmetry), $z_{\alpha/2}$ is the upper- $\alpha/2$ critical point from a standard normal distribution, and $\widehat{\sigma}_{01}$ is the estimated covariance between the maximum likelihood (ML) estimators b_0 and b_1 . The estimated covariance is available from most logistic or probit regression computer outputs.

Example: Copper Toxicity in Fish

We illustrate estimation of median effective dose with a toxicity study in which fish were exposed to potentially toxic concentrations of copper (Cu). In this study [11], $n = 20$ fish were exposed to the metal at each of seven concentrations (in $\mu\text{g l}^{-1}$). The proportions of fish dying after 4 days were recorded: at Cu concentrations $d_1 = 0.1$, $d_2 = 0.2$, $d_3 = 0.3$, $d_4 = 0.5$, $d_5 = 1$, $d_6 = 2$, and $d_7 = 2.5$, the observed proportions were 2/20, 2/20, 5/20, 8/20, 15/20, 15/20, and 17/20, respectively. From a plot of fish mortality against the natural log of the Cu concentration, a sigmoidal response is evident. Here, we assume a logistic relationship between $\pi(x)$ and $x = \log(\text{Cu})$.

The ML estimates of the logistic regression parameters for these data can be found to be $b_0 = 0.5463$ and $b_1 = 1.3458$. The point estimate of LC_{50} is then based on $\log(\widehat{LC}_{50}) = -0.5463/1.3458 = -0.4059$, with large-sample 95% Fieller limits $-0.7199 \leq \log(LC_{50}) \leq -0.0734$. Converting to original Cu concentrations produces $\widehat{LC}_{50} = e^{-0.4059} = 0.9602 \mu\text{g l}^{-1}$, with 95% Fieller limits $e^{-0.7199} = 0.4868 \leq LC_{50} \leq e^{-0.0734} = 0.9292 \mu\text{g l}^{-1}$.

For data in the form of counts, a simple log-linear model may be employed for stimulus-response analysis: $\log\{\mu(x)\} = \beta_0 + \beta_1 x$. If the dose-response to the environmental agent is expected to produce a decrease in the mean rate of response, then we define the ED_{50} as that dose producing a 50% drop from the mean response at zero stimulus, $x = 0$. That is, under $\mu(x) = \exp\{\beta_0 + \beta_1 x\}$, a 50% reduction in

$\mu(0)$ is $\frac{1}{2} \exp\{\beta_0\}$, so we solve for x in $2 \exp\{\beta_0 + \beta_1 x\} = \exp\{\beta_0\}$. This yields $ED_{50} = -\log(2)/\beta_1$. Given stimulus-response count data under a Poisson parent distribution, we calculate the ML estimate, b_1 , of β_1 and then find $\widehat{ED}_{50} = -\log(2)/b_1$. If we also calculate the typical ("Wald") $1 - \alpha$ confidence limits $b_1 \pm z_{\alpha/2} se[b_1]$, a set of $1 - \alpha$ limits on $ED_{50} = -\log(2)/\beta_1$ is given by

$$\frac{-\log(2)}{b_1 - z_{\alpha/2} se[b_1]} < ED_{50} < \frac{-\log(2)}{b_1 + z_{\alpha/2} se[b_1]} \quad (2)$$

An informal assumption being made here is that $b_1 \leq 0$, since we expect the stimulus to decrease the mean response. Notice then that, if $z_{\alpha/2} se[b_1]$ is so large as to force $b_1 + z_{\alpha/2} se[b_1] > 0$, the interval calculation will lead to a nonsensical result. This is appropriate, however, since in this case an approximate $1 - \alpha$ confidence interval for β_1 contains zero. That is, there is no significant evidence for a dose effect, and hence no reason to calculate a potency estimate.

Bailer and Oris [12] consider the more general problem of estimating ED_{50} 's under any form of generalized linear model, of which the logistic, probit, and log-linear models are special cases. They develop core polynomial (in dose) predictor functions, and give examples with data in the form of binary survival rates, counts, and continuous measurements.

Other Levels of Effective Stimulus

Traditionally, ED_{50} has been used as a summary measure of the dose effect owing to its central location on the stimulus-response curve. However, many other response levels, such as ED_{01} , ED_{10} , ED_{90} , etc., are possible. Depending on the application, any of these may be appropriate for summary use. In general, the *effective dose* ρ is the dose that yields a $100\rho\%$ effect over the stimulus-response curve, for $0 < \rho < 1$. We denote this as $ED_{100\rho}$. For example, with quantal response under the simple linear logistic model

$\pi(x) = 1/(1 + \exp\{-\beta_0 - \beta_1 x\})$, $ED_{100\rho}$ is the value of x that solves $\rho = [1 + \exp\{-\beta_0 - \beta_1 x\}]^{-1}$. Hence,

$$ED_{100\rho} = \frac{1}{\beta_1} \left\{ \log \left\{ \frac{\rho}{1-\rho} \right\} - \beta_0 \right\} \quad (3)$$

At $\rho = 1/2$, this collapses to the ED_{50} seen above.

A similar construction is possible for any generalized linear model with quantal response data; e.g., under the simple linear probit model $\pi(x) = \Phi(\beta_0 + \beta_1 x)$ we find $ED_{100\rho} = \{z_{(1-\rho)} - \beta_0\}/\beta_1$. To estimate $ED_{100\rho}$, we substitute the ML estimates of the regression parameters, b_0 and b_1 , into the pertinent expression for $ED_{100\rho}$. Tamhane [13] reviews these and other estimation issues for measuring $ED_{100\rho}$.

We once again use Fieller's theorem to construct large-sample $1 - \alpha$ confidence limits on $ED_{100\rho}$. As long as $ED_{100\rho}$ is of the form $(\delta - \beta_0)/\beta_1$ where δ is some known constant, the form of the confidence limits from equation (1) remains unchanged. For example, under the simple linear logistic model, $ED_{100\rho}$ is of the form $(\delta - \beta_0)/\beta_1$, with $\delta = \log\{\rho/(1 - \rho)\}$. Hence, an associated set of $1 - \alpha$ confidence limits on $ED_{100\rho}$ is

$$\begin{aligned} \widehat{ED}_{100\rho} + \frac{\gamma}{1-\gamma} \left\{ \widehat{ED}_{100\rho} + \frac{\widehat{\sigma}_{01}}{se^2[b_1]} \right\} \\ \pm \frac{z_{\alpha/2}}{(1-\gamma)|b_1|} \sqrt{se^2[b_0] + 2\widehat{\sigma}_{01}E\widehat{D}_{100\rho} + se^2[b_1]E\widehat{D}_{100\rho}^2 - \gamma \left\{ se^2[b_0] - \frac{c_{01}^2}{se^2[b_1]} \right\}} \end{aligned} \quad (4)$$

where $\gamma = z_{\alpha/2}^2 se^2[b_1]/b_1^2$.

Other Potency Measures

Many other measures have been proposed to describe the potency of an environmental stimulus, taken within the context of risk estimation. Perhaps the simplest ones are known as *observed-effect levels*. Specifically, the *no-observed-adverse-effect level* (NOAEL, sometimes written simply as NOEL), is the largest level of the stimulus or dose where no adverse effect is observed in the experimental units. If the stimulus is recorded as a concentration, we often denote this as NOAEC. If at even the lowest level of the environmental stimulus a significant increase is observed over background, one calculates the *lowest-observed-adverse-effect level* (LOAEL or LOAEC).

LOAEL is the lowest level of the stimulus or dose above background where a significant adverse effect is observed. The LOAEL is also called the *least effective dose* (LED) or *minimal effective dose* (MED) by some authors; the latter is especially common in the pharmacological and medical literature.

For both NOAEL and LOAEL, estimates are determined statistically by performing a series of pairwise comparisons at each nonzero level of the stimulus against the response at the control or background level(s). Clearly, correction for multiple testing is required [14].

Numerous concerns have been raised regarding the NOAEL/LOAEL approach, since it suffers from a variety of instabilities. For instance, a NOAEL or LOAEL must by definition correspond to the actual levels of the stimulus under study. Thus, the spacing of stimulus levels is critical to these summary values, and ultimately to any risk assessments based upon them. Dose selection must be performed carefully, and include a dense enough grid of stimulus levels to estimate LOAEL and/or NOAEL with sufficient accuracy. If the experiment is poorly designed or controlled, excessively high variability can result. This

typically raises the values of NOAEL and LOAEL; thus, high variability in the response will translate into assertions of low potency (large NOAELs), regardless of the true response effect. Indeed, a design that allocates too few experimental units to the stimulus levels may overestimate or even fail to identify a NOAEL or LOAEL. These difficulties have led to concerns that observed-effect levels such as NOAEL are poor summary statistics for stimulus–response data [15, 16].

One feature that both $ED_{100\rho}$ and NOAEL share is that they are increasing measures of potency; i.e., as the potency of the stimulus decreases, the potency measure rises. Some authors find this sort of “inverse potency” to be counterintuitive, calling instead for transformations such as $1/ED_{100\rho}$ or $\log\{(1/ED_{100\rho}) + 1\}$ that recover direct proportionality, or to development of new, directly proportional

measures. For instance, suppose the form of the stimulus–response function is known and that an increasing response indicates a detrimental effect. Since in this case a sharper increase indicates a more potent stimulus, a direct measure of potency is the rate of increase in the stimulus–response curve. Mathematically, this is the slope or tangent line of the curve at each level of the stimulus. Unfortunately, for most models the slope varies with dose and so some specific dose must be chosen for measuring the curve’s slope. A common solution is to use the incremental rate of change in the dose–response just past $x = 0$. That is, given a stimulus–response function $\mu(x)$, define the potency of a hazardous agent as the stimulus–response slope at $x = 0$: this is typically the first derivative at zero, $\mu'(0)$. This “slope-at-zero” measure is motivated from low-dose estimation arguments: stimulus–response behavior at low doses of an environmental agent may approximate human responses to the agent, hence incremental change in dose–response is of greatest interest near $x = 0$. Estimates of the stimulus–response parameters and of $\mu'(0)$ can be calculated using some form of nonlinear regression.

The use of a slope-at-zero measure is not without controversy, since for some stimulus–response functions slope measures may not exist or be sensible. For example, suppose $\mu(x)$ is an unknown-order polynomial such as $\beta_0 + \beta_1 x^{\beta_2}$. For $\beta_2 > 1$, the slope at zero is $\mu'(0) = 0$ and hence will always give a zero potency for any set of data. Clearly, the choice of a stimulus–response function can influence the ultimate value of any functionally derived potency estimator, and so should be made carefully.

Measures such as NOAEL, LOAEL, $\mu'(0)$, etc., all refer to the background or control response, i.e., to organisms or subjects in clean, uncontaminated conditions. When stimulatory or protective effects are observed at low concentrations of the environmental agent, however, alternate definitions of potency may be desired. For instance, suppose that the detrimental effect is exhibited as a reduction in response, such as inhibition of reproductive capacity. Then, the potency of the environmental inhibitor might be referenced against the maximal observed reproductive response, in effect defining the reproductive impact relative to the optimal output of the organisms or subjects under study. In particular, suppose the mean reproductive response at inhibitor level x is given by the log-linear model $\mu(x) = \exp\{\beta_0 + \beta_1 x + \beta_2 x^2\}$. Then,

the response inhibitor (RI) level associated with a $100(1 - \rho)\%$ drop in capacity relative to maximal response is

$$RI_{100\rho} = -\frac{\beta_1}{2\beta_2} + \sqrt{\frac{\log(1 - \rho)}{\beta_2}} \quad (5)$$

where we assume $\beta_2 < 0$ and $\beta_1 \geq 0$. Notice that this measure is a form of $ED_{100\rho}$. The control response level and the maximal response level are the most common choices for the baseline from which an inhibition concentration is estimated. Bailer and Oris [17] discuss these and several other possibilities in more detail.

Carcinogenic Potency

Historically, in potency estimation for the special case of carcinogenic agents the Armitage and Doll [18] multistage model was used to describe how the probability of tumor/cancer presence varies as a function of exposure to the agent. A mathematically simplified form of this model for the quantal response data setting is $\pi(x) = 1 - \exp(-q_0 - q_1 - q_2 x^2 - \dots - q_k x^k)$, where $q_i \geq 0, i = 0, 1, \dots, k$, and k is a fixed integer no greater than the number of dose groups minus 1 [19]. The dose associated with a specified increase in the probability of tumor onset in individuals who would have been tumor free in the absence of exposure is often of interest. This translates to operating on an *extra risk* (ER) metric: the excess risk function is

$$ER(x) = \frac{\pi(x) - \pi(0)}{1 - \pi(0)} \quad (6)$$

We solve $ER(x) = R$ for x , where R is some small benchmark value of the ER, say 10^{-6} . For the multistage model, the excess risk function simplifies to $ER(x) = 1 - \exp(-q_1 - q_2 x^2 - \dots - q_k x^k)$. For small values of x approaching zero, this approximates further to $ER(x) \approx q_1 x$. (The requirement that x be very close to zero is critical here, since the approximation worsens dramatically as x grows large.) This motivates the definition of a common potency endpoint defined for carcinogenic responses: q_1 is often called the *unit cancer risk*, since a unit increase in dose x leads to a q_1 increase in extra (cancer) risk. In practice, we calculate a $1 - \alpha$ upper confidence limit on the q_1 parameter, say, $q_1 + z_{\alpha} se[q_1]$. This limit is often denoted as q_1^* in the literature.

Alternatively, an analog to the ED_{50} for carcinogenic responses is the *median tumorigenic dose*, or TD_{50} . This corresponds to the dose at which the proportion of tumor-free animals decreases by 50% [20–22], and can be a useful alternative for a carcinogenic potency value [23].

Benchmark Doses

The dose associated with a specified level of change relative to responses in unexposed individuals has been captured under the umbrella of *benchmark doses*, or BMDs. The specified level of change is called the *benchmark response*, or BMR (see **Benchmark Dose Estimation**). Originally, Crump [24] proposed BMD as a replacement for NOAEL in calculations of allowable daily intakes. This endpoint can be defined for responses on any scale of measurement. For example, with quantal responses we often employ the ER scale as seen in the carcinogenic potency setting discussed above. Continuous responses might be described in terms of the dose required to change the mean response in an exposed group by a specified amount relative to control group responses [25, 26]. BMD can also be defined in the context of epidemiological investigations [27] or ecotoxicological studies [28].

Relative Potency

When comparing two different chemicals, the ability of one chemical to induce an adverse response relative to the other may be of interest. As noted in Finney [2], if two chemicals (or a “preparation” versus a “standard”) are compared, then the relative potency can be defined as ρ in $F_0(x) = F_1(\rho x)$ where $F_i(\cdot)$ is the regression function relating dose to response for the i th chemical ($i = 0, 1$). Here, a dose x of chemical 0 elicits the same response as a dose ρx of chemical 1. If the $F_i(\cdot)$ functions are linear in $\log(x)$ with a common slope (as seen in *parallel line assay*) we write $F_i(x) = \alpha_0 + \beta \log(x)$. Then, the relative potency can be defined as $\rho = \exp[(\alpha_0 - \alpha_1)/\beta]$.

The concept of relative potency has been encountered in the risk assessment of mixtures, in particular, mixtures of chemicals that might be structurally related. In cases where a mixture exhibits *dose addition*, the risk of a mixture is determined by summing

the doses of the constituents of a mixture multiplied by the relative potency of each constituent [29]. *Dose addition* is linked to assumptions of a common mode of action for the mixture components, or of similarly shaped dose–response relationships for the different components [30]. The use of relative potencies to weight mixture components is also considered in the construction of toxic equivalence factors (TEFs); see, e.g. [31].

References

- [1] Bliss, C.I. (1935). The calculation of dosage-mortality curves, *Annals of Applied Biology* **22**, 134–167.
- [2] Finney, D.J. (1952). *Statistical Method in Biological Assay*, Charles Griffin, London.
- [3] Trevan, J.W. (1927). The error of determination of toxicity, *Proceedings of the Royal Society of London, Series B* **101**, 483–514.
- [4] Hosmer, D.W. & Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data*, John Wiley & Sons, New York.
- [5] Irwin, J.O. (1937). Statistical method applied to biological assays (with discussion), *Journal of the Royal Statistical Society* **4**(1), 1–60.
- [6] Fieller, E.C. (1940). The biological standardization of insulin, *Journal of the Royal Statistical Society, Series B* **7**, 1–53.
- [7] Bhattacharya, R. & Kong, M. (2006). Consistency and asymptotic normality of the estimated effective doses in bioassay, *Journal of Statistical Planning and Inference* **137**, 643–658.
- [8] Stukel, T.A. (2000). A general model for estimating ED_{100p} for binary response dose-response data, *American Statistician* **44**, 19–22.
- [9] Bailer, A.J. & Piegorsch, W.W. (2000). Quantitative potency estimation to measure risk with bio-environmental hazards, in *Handbook of Statistics Volume 18: Bioenvironmental and Public Health Statistics*, P.K. Sen & C.R. Rao, eds, Elsevier Publishing, Amsterdam, pp. 441–463.
- [10] Buonaccorsi, J.P. (2002). Fieller’s theorem, in *Encyclopedia of Environmetrics*, A.H. El-Shaarawi & W.W. Piegorsch, eds, John Wiley & Sons, Chichester, Vol. 2, pp. 773–775.
- [11] Wilhelm, M.S., Carter, E.M. & Hubert, J.J. (1998). Multivariate iteratively re-weighted least squares, with applications to dose-response data, *Environmetrics* **9**, 303–315.
- [12] Bailer, A.J. & Oris, J.T. (1997). Estimating inhibition concentrations for different response scales using generalized linear models, *Environmental Toxicology and Chemistry* **16**, 1554–1560.
- [13] Tamhane, A.C. (1986). A survey of literature on quantal response curves with a view towards application of

- the problem of selecting the curve with the smallest q -quantile (ED100 q), *Communications in Statistics: Theory and Methods* **15**, 2679–2718.
- [14] Yanagawa, T., Kikuchi, Y. & Brown, K.G. (1994). Statistical issues on the no-observed-adverse-effect level in categorical response, *Environmental Health Perspectives* **102**(1), 95–104.
 - [15] Brand, K.P., Rhomberg, L. & Evans, J.S. (1999). Estimating noncancer uncertainty factors: are ratios NOAELs informative? *Risk Analysis* **19**, 295–308.
 - [16] Chapman, P.M., Caldwell, R.S. & Chapman, P.F. (1996). A warning: NOECs are inappropriate for regulatory use, *Environmental Toxicology and Chemistry* **15**, 77–79.
 - [17] Bailer, A.J. & Oris, J.T. (2000). Defining the baseline for inhibition concentration calculations for hormetic hazards, *Journal of Applied Toxicology* **20**, 121–125.
 - [18] Armitage, P. & Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis, *British Journal of Cancer* **8**, 1–12.
 - [19] Crump, K.S. (1984). A new method for determining allowable daily intake, *Fundamental and Applied Toxicology* **4**, 854–871.
 - [20] Goddard, M., Krewski, D. & Zhu, Y. (1994). Measuring carcinogenic potency, in *Environmental Statistics, Assessment and Forecasting*, C.R. Cothorn & N.P. Ross, eds, Lewis Publications, Boca Raton, pp. 193–208.
 - [21] Krewski, D. (1990). Measuring carcinogenic potency, *Risk Analysis* **10**, 615–617.
 - [22] Sawyer, C., Peto, R., Bernstein, L. & Pike, M.C. (1984). Calculation of carcinogenic potency from long-term animal carcinogenesis experiments, *Biometrics* **40**, 27–40.
 - [23] Dewanji, A. (2000). Carcinogenic potency: statistical perspectives, in *Handbook of Statistics Volume 18: Bioenvironmental and Public Health Statistics*, P.K. Sen & C.R. Rao, eds, Elsevier Publishing, Amsterdam, pp. 895–911.
 - [24] Crump, K.S. (1984). An improved procedure for low-dose carcinogenic risk assessment from animal data, *Journal of Environmental Pathology, Toxicology, and Oncology* **5**, 339–348.
 - [25] Gaylor, D.W. & Slikker, W.L. (1990). Risk assessment for neurotoxic effects, *NeuroToxicology* **11**, 211–218.
 - [26] West, R.W. & Kodell, R.L. (1993). Statistical methods of risk assessment for continuous variables, *Communications in Statistics – Theory and Methods* **22**, 3363–3376.
 - [27] Budtz-Jørgensen, E., Keiding, N. & Grandjean, P. (2001). Benchmark dose calculation from epidemiological data, *Biometrics* **57**, 698–706.
 - [28] Piegorsch, W.W., West, R.W., Pan, W. & Kodell, R.L. (2005). Simultaneous confidence bounds for low-dose risk assessment with non-quantal data, *Journal of Biopharmaceutical Statistics* **15**, 17–31.
 - [29] Krewski, D. & Thomas, R.D. (1992). Carcinogenic mixtures, *Risk Analysis* **12**, 105–113.
 - [30] U.S. EPA (2003). Developing relative potency factors for pesticide mixtures: biostatistical analyses of joint dose-response, Technical report, number EPA/600/R-03/052, Office of Research and Development, National Center for Environmental Assessment, Cincinnati.
 - [31] Birnbaum, L.S. & DeVito, M.J. (1995). Use of toxic equivalency factors for risk assessment for dioxins and related compounds, *Toxicology* **105**, 391–401.

A. JOHN BAILER AND WALTER W. PIEGORSCH

Precautionary Principle

The *precautionary principle* (PP) is a term used about a set of statements all expressing the general intention that, in environmental risk policy, it is prudent to not always wait for definitive proof of harm before taking regulatory action. There are various formulations, such as the following from the third North Sea Conference in 1990 [1, p. 39]:

To take action to avoid potentially damaging impacts of substances that are persistent, toxic and liable to accumulate even when there is no scientific evidence to prove a causal link between emission and effects

and – perhaps one of the most widely known – in the Rio declaration from 1992 [2]:

Where there are threats of serious or irreversible damage, scientific uncertainty shall not be used to postpone cost-effective measures to prevent environmental degradation.

A stronger statement was formulated by a group of experts in the so-called Wingspread declaration (1998) [3]:

When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause and effect relationships are not fully established scientifically.

The European Union considers the PP [4]

a general rule of public policy action to be used in situations of potentially serious or irreversible threats to health or to the environment, where there is a need to act to reduce potential hazards before there is strong proof of harm, taking into account the likely costs and benefits of action and inaction

and the Maastricht Treaty states (Article 130r(2))

Community Policy on the environment shall aim at a high level of protection taking into account the diversity of situations in the various regions of the Community. It shall be based on the precautionary principle and on the principles that preventative action should be taken, that environmental damage should as a priority be rectified at source, and that the polluter should pay.

As developed more generally by Grandjean [2], the role of the PP in the policy process needs to be

understood in the wider context that there is always a conflict between unavoidable scientific uncertainty and the policy makers' need to make decisions. The PP can then imply focus on assessing of alternatives, full-cost assessment, and a participatory decision process. It may be coupled with the right-to-know aspect, empowering the citizen to take action. And employing the PP will often imply taking actions involving more research, preventive intervention, or both.

Barnett and O'Hagan [1, p. 40] formulated the consequence of the PP for the use of scientific knowledge in setting environmental standards:

In general, we have seen that weak knowledge must lead to more stringent standards, on precautionary grounds. The value of science is to allow us, where appropriate, to relax the standards.

Critique of the Precautionary Principle

The PP has been criticized for resulting in huge expenses to prevent risks that do not exist, and for generally hampering progress, because nobody can take the risks necessary were the PP applied in general. Here, it is important to realize, as alluded to above, that there is always a judgment phase involved, when transforming current scientific knowledge into policy decisions.

It has also been stated that the PP could be problematic in a democratic society by prescribing action over and above what the democratic process itself would decide. However, a similar critique would also be valid against blind technocratic application of conventional risk assessment.

For a broad discussion of the interface of the PP, science, and regulatory activity, see [2] and the proceedings [5] of a dedicated workshop.

The Precautionary Principle, Scientific Discovery, and Statistics

Much scientific activity is based on the Popper paradigm of formulating precise hypotheses, which experimental data may then falsify. Following falsification, the researcher must formulate new hypotheses generating new experimental testing. The statistical significance test as formulated by Fisher [6] is very parallel. He stated: "A test of significance contains

no criterion for “accepting” a hypothesis. According to circumstances it may or may not influence its acceptability”.

The PP points out that this paradigm for scientific discovery should not be interpreted in the risk assessment context to say that only if a claim of no hazard has been scientifically (statistically) falsified should there be regulatory action. This is because there may be two very different reasons for the available data not to contain evidence against the hypothesis of no hazard: either there may indeed be no effect or an effect of a size below biological interest, or the data may be too noisy (often because there are too few of them) to enable any positive conclusion, the latter being termed *type II error* in statistical jargon. This dilemma is well known in biostatistics; in bioequivalence testing [7] one turns the logic upside down, testing the null hypothesis that two drugs are different by more than a given threshold, and by rejecting that hypothesis, on obtaining positive evidence of equality.

Precautionary Properties of Environmental Standards under Statistical Uncertainty

Keiding and Budtz-Jørgensen [8] outlined how statistical uncertainty affects the no observed adverse effect level (NOAEL), the benchmark approach, and hockey-stick models.

NOAEL

The NOAEL is the highest dose (in a standard toxicological animal experiment) that does not show increased risk over that of dose 0. The standard hypothesis-testing framework discussed above implies that the more evidence, the higher the NOAEL, so that the NOAEL is *antiprecautionary*.

Benchmark Approach

The Benchmark approach was introduced by Crump [9] partly motivated by an aim of avoiding the unsatisfactory antiprecautionary property of NOAEL. Statistical uncertainty will indeed lead to smaller studies having lower benchmark bounds, so that in its original form designed for standard toxicological animal experiments, the Benchmark approach is indeed precautionary.

For application to epidemiological exposure-response data, interest widens to sources of variation other than just experiment size. Budtz-Jørgensen *et al.* [10, 11] showed that the larger the variance around the dose–response curve, the larger the lower bound: an antiprecautionary property, and that if there is measurement uncertainty in the exposure variable (and there often is!), then the lower bound is overestimated, again an antiprecautionary property of the Benchmark approach.

Hockey-Stick Model

A simple approach to setting standards is to postulate that the exposure–response relationship is zero up to a certain breakpoint and increases linearly thereafter. It may be shown that small experiments, and large variation around the exposure–response relationship, both lead to overestimation of the breakpoint (antiprecautionary), while uncertainty in the exposure measurements decreases the estimate of the breakpoint (precautionary).

Concluding Remarks

The key element of the PP is the justification for acting in the face of uncertain knowledge about risks from environmental exposures. Appropriate public health action should be taken in response to limited, but plausible and credible evidence of likely and substantial harm. The PP is thereby aimed at avoiding possible future harm associated with suspected, but not conclusive, environmental risks. In parallel, the burden of proof is shifted from demonstrating the presence of risk to demonstrating the absence of risk. While the PP is not a prescription for banning the use of industrial chemicals, it requires that alternative options be evaluated to minimize the costs of surprises and maximize the benefits of innovation. Although developed in regard to social and environmental policies, the idea that policy development should not await a “final” proof or overwhelming scientific documentation may be of wider relevance.

While the legal implications of the PP are being developed within the EU legal system, the PP may also inspire some change in scientific tradition. In particular, replication has been held as a key to hypothesis testing, but the PP could, in some

situations, limit the need for repetitive verification. Further, the PP does not require testing of the null hypothesis, but rather demands information whether a potentially serious hazard may be present. Thus, an upper confidence limit may be more useful than a probability calculated in regard to the null hypothesis. Accordingly, the PP may inspire new ways of planning, conducting, and reporting research.

References

- [1] Barnett, V. & O'Hagan, A. (1997). *Setting Environmental Standards: The Statistical Approach to Handling Uncertainty and Variation*, Chapman & Hall, London.
- [2] Grandjean, P. (2004). Implications of the precautionary principle for primary prevention and research, *Annual Review of Public Health* **25**, 199–223.
- [3] Raffenberger, C. & Tickner, J. (eds) (1999). Wingspread statement on the precautionary principle, in *Protecting Public Health and the Environment: Implementing the Precautionary Principle*, Island Press, Washington, DC, pp. 353–355.
- [4] European Environment Agency (2001). *Late Lessons from Early Warnings: The Precautionary Principle 1896–2000*, Environmental issues report no. 22, Official Publications of the European Communities, Luxembourg.
- [5] Grandjean, P., Soffritti, M., Minardi, F. & Brazier, J.V. (eds) (2003). *The Precautionary Principle*, Ramazzini Foundation Library, European Foundation of Oncology and Environmental Sciences, Bologna, Vol. 2.
- [6] Fisher, R.A. (1959). *Statistical Methods and Scientific Inference*, Oliver & Boyd, London.
- [7] Endrenyi, L. (2005). Bioequivalence, in *Encyclopedia of Biostatistics*, 2nd Edition, P. Armitage & T. Colton, eds, John Wiley & Sons, Chichester, pp. 455–460.
- [8] Keiding, N. & Budtz-Jørgensen, E. (2003). The precautionary principle and statistical approaches to uncertainty, *European Journal of Oncology Library* **2**, 185–191. Republished in *International Journal of Occupational Medicine and Environmental Health* **17**, 147–151 (2004). Republished in *Human and Ecological Risk Assessment* **11**, 201–207 (2005).
- [9] Crump, K. (1984). A new method for determining allowable daily intakes, *Fundamental and Applied Toxicology* **4**, 854–871.
- [10] Budtz-Jørgensen, E., Keiding, N. & Grandjean, P. (2001). Benchmark dose calculation from epidemiological data, *Biometrics* **57**, 698–706.
- [11] Budtz-Jørgensen, E., Keiding, N. & Grandjean, P. (2004). Effects of exposure imprecision for estimation of the benchmark dose, *Risk Analysis* **24**, 1689–1696.

Related Articles

Conflicts, Choices, and Solutions: Informing Risky Choices

Ecological Risk Assessment

Environmental Health Risk

Risk and the Media

PHILIPPE GRANDJEAN AND NIELS KEIDING

Precautionary Principles: Definitions, Issues, and Implications for Risky Choices

The several forms of the precautionary principle generally stand for the proposition that *when there are threats of serious or irreversible damage to the environment, scientific uncertainty should not prevent prudent actions to prevent potentially large or irreversible damage*. The main difference is in how precaution should be exercised in practice. This foundational basis for precautionary principle in many conventions and treaties, as well as pronouncements by a number of entities, requires differentiating them between those that are legally binding (e.g., Title XVI, Art. 130r of the Maastricht Treaty on European Union (EU) (1992); the Delaney Clause to the US Federal Food, Drugs, and Cosmetics Act (FFDCA)) and those that are not (as in a convention that has not been ratified by a nation that has signed it). Perhaps, the best-known enunciation of the precautionary principle is the Rio (de Janeiro) Declaration on Environment and Development of 1992.

In order to protect the environment the Precautionary Approach shall be widely applied by states according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation.

Precautionary principles should also be differentiated from other principles of precaution, such as *zero risk* or *significant risk*. All reflect varying levels of public policy, often resulting in secondary legislation that errs on the side of uniformity and ease of use by adopting defaults that are conjectural. And yet, the concern is that – because of the conflicting or incomplete results of science, or its inability to provide the answers needed at the required time and with sufficient certainty as to the potential magnitude of the adverse effects – anticipatory actions and erring on the side of caution are necessary in choices that affect occupational or public hazards. In some enunciations of those principles, the magnitude or

the severity of the potential consequences leads to disregarding their probabilities and the uncertainty in the causal connection between them and the hazard. Bans (either *all-or-none* or phased over time, and prohibitions) characterize these choices. Others believe in the proposition that precautionary decisions and choices must be guided by balancing risks, costs, and benefits, as well as being characterized by a specific level of legally demonstrable causation. In those enunciations, economic rationality guides ranking and selection of options: for instance, the use of a criterion of choice (such as the minimax, or the expected value of the magnitude of the consequences) may indicate that the *do nothing* action is superior, at least until some key uncertainties are resolved or reduced.

Regardless of the different forms of precautionary principles, it is their instrumental aspects that can create practical difficulties, not the principles themselves. The reason is that precautionary principles formulate a moral justification for acting when a hazard can *potentially* cause severe or irreversible consequences and its causation is uncertain. Yet, this moral justification can result in a suboptimal choice: *consequences that are more adverse* have been known to occur after implementation of a precautionary choice than before. This situation can occur when the choice of action, incorrect perceptions, media reiterations, ideology, political gain, and dread amplify the importance of the potential hazard and thus serve to justify that choice. Adding to these perceptual and cognitive issues, incomplete, heterogeneous, uncertain, variable, and complex scientific evidence also serves to justify a possibly scientifically unsound choice, such as one based on *political will* uninformed by a scientific risk assessment. Unless a sound and replicable analysis, based on the current state of knowledge, is coupled with the moral basis for the choice, such a choice may be (after more knowledge becomes available) incorrect and possibly illegal.

The repeated lesson, still only partially included in much political discussion, is that precautionary choices thought to prevent hazards when they are implemented, are later found to cause unanticipated harm. If that unanticipated harm could or should have been predicted, the failure to have done so is inexcusable in the sense that acting on a conjectured danger, justified by a legal maxim, is unfair – because it redirects scarce resources from more valuable societal investments. Yet, not acting by postponing preventive

action, because of large uncertainties surrounding the magnitude or severity of the hazard, can increase the magnitude of those consequences. Thus the dilemma, should society wait until the uncertainties are sufficiently resolved to trigger regulatory action, or act precipitously on a conjecture? The corollary question is, what is sufficient evidence?

Resolving this dilemma is beyond our scope, but its approach to a resolution is discussed throughout other articles in this encyclopedia (see **Axiomatic Measures of Risk and Risk-Value Models; Statistics for Environmental Justice**). For this essay, there is a simple rule that is specific to uncertain hazards: the assessment and selection of the preferred choice from the set of choices is based on *scientific analyses* of these choices, given the state of knowledge about the hazard. Each choice, consequence, probability of occurring, and so on is assessed through risk-cost-benefit balancing; the idea is to *inform* and *guide* decision makers and stakeholders, not command them to commit to a specific action. The justification of the final choice in the public interest, a matter of public policy, must account for factors that, for one reason or another, cannot be directly included in the risk-cost-benefit balancing. In the EU, for example, having to account for the *economic and social development of the community as a whole* as per Art. 134r(3) of the Treaty of Maastricht can require analyses that extend beyond the confines of a choice-specific balancing. The risk-cost-benefit analyses are essential because they account for uncertainty, causation, the value of new information and the cost of gathering it, and the utility of the payoffs associated with each option open to the decision maker. In terms of how to analyze those actions, balancing criteria such as the maximization of the expected discounted values of the payoffs, can guide the selection of the optimal (relative to the criterion adopted) act by identifying it. Those analyses are transparent, replicable, and testable *via* a number of methods, including simulations. Their extent can go beyond microeconomic consideration to include macroeconomic ones. There are costs of either action or inaction to society; nonetheless, the magnitude of the stakes justifies formal analysis of risky choices. However, conducting formal analyses is also costly. These costs include having to deal with: complex mathematical and statistical issues, data needs and their availability, and several forms of uncertainty and variability. In the

end, the results from formal analysis are constructive and meant to inform the policy-makers. Those analyses are not normative.

Precautionary Principles

An early example of legal precaution is Justinian's statement, in 527 A.D., that *the maxims of the law are to live honestly, to cause no harm unto others, and to give everyone his due*. Today, the forms of the precautionary principle generally require ranking of choices based on either a full risk-cost-benefit balancing or some accounting for the costs or risks, or both, but one should also consider, depending on the jurisdiction, factors (such as harmonization and subsidiarity) that go beyond the confines of cost-benefit analysis. Legal principles are generally unenforceable as such, although modern legislation accords them a positive and guiding force; preambles and principles are statements of general legal intent and thus lack the specificity required for enforcement. On the other hand, certain enunciations of the precautionary principle are a constitutional command, as in the EU.

Some Influential Precautionary Principles

Montreal Protocol on Substances that Deplete the Ozone Layer (Preamble, Para. 6), (Amended, 1990) ... protect the ozone layer by taking precautionary measures to control equitably total global emissions of substances that deplete it, with the ultimate objective of their elimination on the basis of developments in scientific knowledge, taking into account technical and economic considerations and bearing in mind the developmental needs of developing countries.

Convention for the Protection of the Marine Environment of the North-East Atlantic (Article 2 (2) (a)), 1992 ... The precautionary principle, by virtue of which preventive measures are to be taken when there are reasonable grounds for concern that substances or energy introduced, directly or indirectly, into the marine environment may bring about hazards to human health, harm living resources and marine ecosystems, damage amenities, or interfere with other legitimate uses of the sea, even when there is no conclusive evidence of a causal relationship between the inputs and the effects.

UN Framework Convention on Climate Change (Article 3 (3)), 1992 ... The Parties should take precautionary measures to anticipate, prevent or minimise the causes of climate change and mitigate its adverse effects. Where there are threats of serious or

irreversible damage, lack of full scientific certainty should not be used as a reason for postponing such measures, taking into account that policies and measures to deal with climate change should be cost effective so as to ensure global benefits at the lowest possible cost.

UK Strategy for Sustainable Development – DoE Consultation Paper, 1993 ... Where appropriate (for example, where there is uncertainty combined with the possibility of the irreversible loss of valued resources) actions should be based on the so-called ‘precautionary principle’ if the likely balance of costs and benefits justifies it. Even then the action taken should be in proportion to the risk.

Depending on the jurisdiction enunciating its form of precaution, it can be used by agencies or authorities to develop secondary legislation (emanating from the relevant federal statute, for example) that can range from a directive in the EU, and guidelines and standards in the United States (where standards are legally enforceable, but guidelines are not). Secondary legislation generally consists of an administrative process that result in numerical environmental or health standards (justified scientifically) designed to achieve the appropriate level of protection, after scientific peer review and public comments, as envisaged in the primary legislation (such as the US Clean Air Act (CAA)). The precautionary standard can be reviewed at the administrative levels (e.g., through the *exhaustion of remedies* doctrine), but its examination can end up in the courts or even be repealed – or made more stringent depending on the scientific information – by the legislature. For example, in American federal environmental law, some statutes have a clause that requires congressional reauthorization after some specified period, e.g., 5 years. In the EU, a new treaty can contain rewritten articles, or have articles deleted from the previous treaty, as is the case of Art. 130r (renumbered in a later treaty of the EU), although it was not changed and still has legal force within the EU. Those legal processes are lengthy and complex, with the potential for inaccurate and inefficient outcomes for society, and often result in protracted litigation. For instance, consider the statement of the Comptroller General of the United States in 1979:

Major constraints plague the Environmental Protection Agency’s ability to set standards and issue regulations. The most important factor is the inconclusive scientific evidence on which it must often base decisions. Numerous court suits result.

European Union

In the EU, the proposed but not ratified constitution’s (Charter) Part I, Title 1, Article 3 (The Union’s Objectives) states as follows:

The Union shall work for the sustainable development of Europe based on balanced economic growth, a social market economy, highly competitive and aiming at full employment and social progress, and with a high level of protection, and improvement of the quality of the environment. It shall promote scientific and technological advance.

Article II-37 (environmental protection) reports as given below:

A high level of environmental protection and the improvement of the quality of the environment must be integrated into the policies of the Union and ensured in accordance with the principle of sustainable development.

Part III (the policies and functioning of the Union) Section 5, Art. III-129 (2) states as follows:

Union policy on the environment shall aim at a high level of protection taking into account the diversity of situations in the various regions of the Union. It shall be based on the precautionary principle and on the principles that preventive action should be taken, that environmental damage should as a priority be rectified at source and that the polluter should pay.

It adds the following guidelines:

3. In preparing its policy on the environment, the Union shall take account of:
 - (a) available scientific and technical data;
 - (b) environmental conditions in the various regions of the Union;
 - (c) the potential benefits and costs of action or lack of action;
 - (d) the economic and social development of the Union as a whole and the balanced development of its regions.

Under the earlier European Treaty of Union (Maastricht, 1992), Article 130r (renumbered as Article 170 in a later Treaty) states the following legislative mandate for environmental protection policy:

2. Community policy on the environment shall aim at a high level of protection ... (and) shall be based on the precautionary principle ...

3. In preparing its policy on the environment, the Community shall take account of:
 - (i) available scientific and technical data;
 - (ii) environmental conditions in various regions of the Community;
 - (iii) the potential benefits and costs of action or lack of action;
 - (iv) the economic and social development of the Community as a whole and the balanced development of its regions.

Acting *notwithstanding* uncertainty is implicit in (2) and (i), *causal reasoning* is explicit in (i), (ii), and (iii), rational decision making is explicit in (iii), and an (unspecified) measure of risk aversion is implicit in (b). This article goes beyond cost-effectiveness and even cost–benefit analysis. Article 130r (1) also states the following European community environmental protection objectives:

- (i) preserving, protecting, and improving the quality of the environment;
- (ii) protecting human health;
- (iii) prudent and rational utilization of natural resources and
- (iv) promoting measures at international level to deal with regional or worldwide environmental problems.

The Maastricht Treaty's Titles II and IV, moreover, contains language such as "sustainable growth" "quality of life" environmental protection as well as "economic and social cohesion". Yet, what should be clear is that the EU treaties do call for the *precautionary principle* but fail to define it. Nonetheless, the sections of Art. 130r do provide the constraints to be placed on precautionary justifications of policy choices (made by the European commission, for example, as discussed below).

Precautionary reasoning creates a threshold above which a public institution must act to prevent harm, provided such actions are demonstrably cost-beneficial, even when allowing for uncertainty. For any given hazard-to-risk situation that falls under any of the forms of precaution, intuitively, and factually, there is a *tolerable window* of low risk where precautionary actions are dormant, but during which further data must be developed, analyzed, and used to determine if the hazard is of societal concern. Below the lower limit of that interval, no action would take place. Above the upper limit, actions can be taken,

provided that the risk-cost-benefit balancing is positive. The EU context of the precautionary principle points to decisions justified by a formal account of the changes in risk, cost, and benefit. Essentially, it commands making rational decisions based on the efficient social choices by balancing of *the potential benefits and costs of action or lack of action*.

The EU context for the precautionary principle is legal. The European Court of Justice (ECJ)'s review of the EU commission decision to ban export of beef from the United Kingdom to reduce the risk of bovine spongiform encephalopathy (BSE) transmission (Judgments of May 5, 1998, cases C-157/96 and C-180/96) issued the following suggestion:

Where there is uncertainty as to the existence or extent of risks to human health, the institutions may take protective measures without having to wait until the reality and seriousness of those risks become fully apparent.

The Court reasoned as follows:

approach is borne out by Article 130r(1) of the EC Treaty, according to which Community policy on the environment is to pursue the objective *inter alia* of protecting human health. Article 130r(2) provides that that policy is to aim at a high level of protection and is to be based, in particular, on the principles that preventive action should be taken and that environmental protection requirements must be integrated into the definition and implementation of other Community policies.

The ECJ, however, did not provide the framework explicitly to deal with the difficulties inherent to making these decisions. In particular, it did not characterize the formal (statistical) meaning of *fully apparent risks* nor how to deal with the complexities of causation that invariably arise in the situations that trigger precautionary reasoning based on the precautionary principle.

In case T-70/99, the president of the EU court of first instance (CFI) referred to this ECJ judgment, stating that *the protection of public health should undoubtedly be given greater weight than economic considerations*. Nonetheless, because of consistency with the treaty of Union, case law, and subsequent treaties, precautionary decisions must be based on sound analytical – and thus replicable – reasoning or else the assertion that *protection requirements must be integrated into the definition and implementation of other community policies* may not be met

with accuracy and consistency. That is to say, the characteristics (factual and conjectural) of a situation will be different from another, but the formalities inherent to Art. 130r precautionary principle apply to all equally.

Consider the European Council Regulation (December 17, 1998) that resulted in the withdrawal of four antibiotics, including virginiamycin, from the list of additives in animal feeds. Some of the scientific evidence was that pathogen bacteria, which become resistant to the antibiotics due to feeding of antibiotic additives to livestock, could get transferred from animals to humans and thus make them more resistant to treatment with antibiotics. Pfizer and Alpharma, directly affected by the ban, sued in the CFI (a lower court in the EU judicial system) to annul that regulation because of the danger that it would cause *serious and irreparable damage* to them. The CFI found against these two companies because: (i) they could sell their products on markets outside the Community, (ii) two of their European plants would probably not be closed, and (iii) that the ban issued was not final.

Pfizer (virginiamycin)

Although the CFI held that “preventive measures can be adopted without having to wait until the reality and seriousness of the risks being fully apparent”, nonetheless, preventive measures cannot be based on a scientific conjecture, but can be taken where there is a *real risk*. Such risk entails some probability that the negative effects, which the measure should prevent, will occur, but such probability cannot be *zero risk*. The CFI held that an agency must develop a risk assessment, based on both science (*risk assessment*) and policy (*risk management*), to decide on the appropriate action. The role of science is held to be essential; the ultimate outcome is based on policy, informed by science. For virginiamycin, the Scientific Committee for Animal Nutrition (SCAN), an expert committee specifically established to provide scientific advice to the EU to animal feedstuffs, concluded that there was insufficient scientific evidence that virginiamycin would cause the alleged hazard, and thus did not advise the EU to withdraw it from the market.

Alpharma (Bacitracin zinc)

Here, SCAN was not asked to review the scientific evidence, before the council wrote the regulation; nonetheless, the council concluded that it was risky to

human health. Not referring to SCAN is only appropriate in *exceptional circumstance and when there are adequate guarantees*, as the CFI stated, of *scientific objectivity*. The CFI upheld the regulation, stating that EU institutions could take a *horizontal approach* to an *entire class of antibiotics by systematically excluding the use as additives in feeding stuffs of products also used in human medicine*. The CFI found those “exceptional circumstances”, and held that SCAN’s advice is not binding, unless explicitly stated to be so in the legislation, and that it was not mandatory for the commission to seek such advice.

Peel (2004) concludes the following (footnotes omitted):

Despite the limited nature of the scientific evidence available to the Community institutions and the lack of anything indicating an immediate health threat, the CFI found that the Commission and Council had not committed any manifest errors in their review of scientific studies and assessment of the risks to health prior to adopting the measure.¹²⁹ Although the Court stressed that regulatory authorities must have at their disposal scientific information which is sufficiently reliable and cogent to allow them to understand the ramifications of the scientific questions raised and to make a decision on policy measures in full knowledge of the facts, the CFI displayed a strongly deferential attitude when reviewing the institutions’ interpretation of the scientific material and their judgments as to the existence of genuine scientific uncertainty.

The CFI found that the uncertainty in the use of antibiotics as additives to animal diets and human resistance to those antibiotics justified their ban by appeal to the Art. 130r Precautionary Principle. This ban is consistent (proportionate) with protecting human health according to the *high level of protection* of this article. Unfortunately, when virginiamycin, macrolides, and other animal antibiotics used to prevent animal illnesses and enhance performance were indeed withdrawn in several European countries, in keeping with this interpretation of the precautionary principle, bacterial illness rates in both chickens and humans increased dramatically.

Precautionary principles tend to place the burden of proof about who causes the hazardous situation, leading to risk on those who market or develop a product. An example of such a shift (Commission of the European Communities, 2001) is as follows: [*r*]responsibility to generate knowledge about chemicals should be placed on industry. Industry should

also ensure that only chemicals that are safe for the intended purposes are produced. However, partly to avoid spurious litigation, the European Commission (EC's) commentary also states the following:

A scientific evaluation of the potential adverse effects should be undertaken based on the available data ... [T]his requires reliable scientific data and logical reasoning, leading to a conclusion which expresses the possibility of occurrence and the severity of a hazard's impact on the environment, or health of a given population ...

The commentary then concludes that *precautionary measures must not be applied to address conjectured risks*. As the CFI held:

It is necessary, first, to define the 'risk' which must be assessed when the precautionary principle is applied ... A preventive measure cannot properly be based on a purely hypothetical approach to the risk, founded on mere conjecture which has not been scientifically verified ...

Views from Other Jurisdictions

The Canadian government proposed guiding principles for risk analysis that include the precautionary principles (Government of Canada, 2001, *The Need for a Federal Approach*, http://www.tbs-sct.gc.ca/pubs_pol/dcgpubs/RiskManagement/rmf-cgr_e.html 3), which are given below.

- A. Follow-up scientific activities, including further research and scientific monitoring, are a key part of the application of the precautionary approach. ... (a)
- B. "Sufficiently sound information base" should be interpreted as sound and reasonable scientific information, including uncertainties that, through evaluation ... [t]hat is, while scientific information would not need to demonstrate definitively the cause-and-effect relationship between risk and serious harm, it would demonstrate that such a risk exists. (And)
- C. Generally, the responsibility for providing the scientific information base (the burden of proof) should rest with the party who is taking an action associated with potential or serious harm. ...

An important issue that often goes unstated is addressed by Canada (Gov of Canada, 2001, http://www.tbs-sct.gc.ca/pubs_pol/dcgpubs/RiskManagement/rmf-cgr_e.html 3), which points out the following:

... it is impossible to prove a negative (e.g., to prove categorically that something will cause no harm, or to prove with absolute certainty that something bad might not happen or to prove that something is not harmful), but possible to demonstrate that "reasonable testing" was done with no evidence of harm. ... Precautionary measures should be cost-effective, with the goal of generating (i) an overall net benefit for society at least cost, and (ii) efficiency in the choice of measures. ... The real and potential impacts of making a precautionary decision (whether to act or not to act), including social, economic, and other relevant factors, should be assessed. Moreover, consideration of risk-risk trade-offs or comparative assessments of different risks would generally be appropriate (although this may not be possible in circumstances where urgent action is needed). This can ensure that society receives net benefits from decision making, and that the precautionary approach is not used as an unnecessary or unintentional barrier to innovation or technological change. ... As the precautionary approach is, by definition, an evolutionary process, precautionary measures should be monitored on an ongoing basis so that new scientific data that alters cost-effectiveness considerations can be incorporated (including performance monitoring results).

In India, its Supreme Court defined the precautionary in the case *A. P. Pollution Control Board v. Nayudu*, as follows:

The principle of precaution involves the anticipation of environmental harm and taking measures to avoid it or to choose the least environmentally harmful activity. It is based on scientific uncertainty. Environmental protection should not only aim at protecting health, property, and economic interest but also protect the environment for its own sake. Precautionary duties must not only be triggered by the suspicion of concrete danger but also by concern or risk potential.

Importantly, because of the complex nature of uncertain causation, that court also shifted the burden of proof stating that "*the party ... maintaining a less-polluted state should not carry the burden of proof and the party who wants to alter it, must bear this burden.*"

Although this shift of the burden of proof is established in some "common law" countries, such as the United States in both tort and environmental law, bearing that burden requires the analysis of uncertain scientific evidence and meeting both scientific and legal causation. In particular, UK courts have rejected shifting the burden of proof in *Wilsher v. Essex Area*

Health Authority (1988), which reconfirms the tenet that the burden of proof remains on the plaintiff, although Lord Wilberforce made an unsuccessful attempt to reintroduce it in the *McGhee* case, decided in the House of Lords. It is yet unclear as to how the primacy of European law, as well as an eventual constitution for the EU, squares with the UK's rejection (of reversing the burden of proof) because it raises significant conflict of laws with the EU's laws (e.g., with regulations and directives) regarding precautionary measures.

US law contains several variants of the precautionary principle. The precautionary language of the federal legislation that direct the regulatory role of US federal agencies has many facets, unlike the EU's Art. 130r Precautionary Principle. A fundamental difference is the US specificity, at least in some instances of the statutory language:

- In the case of threshold effects an additional ten-fold margin of safety for the pesticide chemical residue shall be applied for infants and children ... (Federal Food, Drugs, and Cosmetics Act, § 408 (b)(2)(C)(b)).
- The Administrator shall, specify, to the extent practicable: 1) Each population addressed by any estimate of public health effects; 2) The expected risk or central estimate of risk for the specific populations; 3) Each appropriate upper-bound or lower-bound estimate of risk ... (Safe Drinking Water Act, § 300g-1 (b)(3)).
- The Administrator shall ... [add] pollutants which present, or may present, through inhalation or other routes of exposure, a threat of adverse human health effects ... or adverse environmental effects through ambient concentrations, bioaccumulation, deposition, or otherwise but not including releases subject to regulation under subsection (r) of this section as a result of emissions to air... (Clean Air Act, § 112(b)(2)).
- Provide an ample margin of safety to protect public health or to prevent an adverse environmental effect (Clean Air Act, § 112(f)).
- To assure chemical substances and mixtures do not present an unreasonable risk of injury to health or the environment (Toxic Substances Control Act, TSCA § 2(b)(3)).
- Function without unreasonable and adverse effects on human health and the environment (FIFRA § 3).
- Necessary to protect human health and the environment (Resources Conservation and Recovery Act, RCRA § 3005 as amended).
- Provide the basis for the development of protective exposure levels (National Contingency Plan) NCP § 300.430(d).
- Adequate to protect public health and the environment from any reasonably anticipated adverse effects (Clean Water Act CWA § 405(d)(2)(D)).

However, primary legislation can still be vague and thus open the road for litigation. Moreover, even the language of a single statute, the CAA, which directs the US environmental protection agency (EPA's) Office of Air and Radiation (OAR), shows that there are differences in the level of precaution to be adopted by this agency. The range is remarkable:

- a) Protect public health with an adequate margin of safety (Clean Air Act, CAA § 109).
- b) Provide an ample margin of safety to protect public health or to prevent an adverse environmental effect (OAR; CAA § 112(f)).
- c) Protect the public welfare from any known or anticipated adverse effects (OAR; CAA § 109).
- d) [Not] cause or contribute to an unreasonable risk to public health, welfare, or safety (OAR; CAA § 202(a)(4)).

The role of risk analysis, under Section 112(f)(2) (B) of the CAA, is that the US EPA uses a two-step risk assessment for setting residual risk standards (articulated in the US EPA's 1989 benzene National Standards for Hazardous Air Pollutants (NESHAP) (Federal Register, 54: 38044)) as given below (US EPA, 2004):

- 1) A first step, in which EPA ensures that risks are 'acceptable.' As explained in the Benzene NESHAP, in this step EPA generally limits the maximum individual risk, or 'MIR,' to no higher than approximately 1 in 10 thousand. The benzene NESHAP defines the MIR as the estimated risk that a person living near a plant would have if he or she were continuously exposed to the maximum pollutant concentrations for a lifetime (70 years).
- 2) A second step, in which EPA establishes an 'ample margin of safety.' In this step, EPA strives to protect the greatest number of persons possible to an estimated individual excess lifetime cancer risk level no higher than approximately 1 in 1 million. In judging whether risks are acceptable and whether an ample margin of safety is provided, the benzene NESHAP (under the Clean Air Act) states that EPA will consider not only the magnitude of individual risk, but 'the distribution of risks in the exposed population, incidence,

the science policy assumptions and uncertainties associated with the risk measures, and the weight of evidence that a pollutant is harmful to health.' . . . Note that the ample margin of safety analysis . . . also includes consideration of additional factors relating to the appropriate level of control, 'including costs and economic impacts of controls, technological feasibility, uncertainties, and any other relevant factors.'

Another US federal statute, comprehensive environmental response, compensation and liability act (CERCLA) (42 U.S.C. 9605 (CERCLA, 1980)) Superfund Amendments and Reauthorization (SARA) Act of 1986, requires that actions selected to clean up hazardous waste sites be protective of human health and the environment. The National Oil and Hazardous Substances Pollution Contingency Plan, NCP, implements CERCLA (US EPA, 2004). The NCP establishes the overall approach for determining appropriate remedial action at superfund sites with the national goal to select remedies that *are protective of human health and the environment, and that maintain protection over time*. One of the policy goals of the program is to protect against a high end, as distinguishable from the worst-case, individual exposure. It is based on the reasonable maximum exposure (RME), which is the highest exposure that is reasonably expected to occur at a superfund site. It is conservative, but within a realistic range of exposure (US EPA, 2004). In addition to RME individual risk, EPA evaluates risks for the *central tendency exposure* (CTE) estimate, or average exposed individual. EPA received comments implying that the RME represents a *worst-case or overly conservative exposure estimate*. Thus, the RME is not a worst-case estimate – *the latter would be based on an assumption that the person is exposed for his or her entire lifetime at the site* (US EPA, 2004).

The reasonable maximum exposure scenario is 'reasonable' because it is a product of factors, such as concentration and exposure frequency and duration, that are an appropriate mix of values that reflect averages and 95th percentile distribution . . . The RME represents an exposure scenario within the realistic range of exposure, since the goal of the Superfund program is to protect against high-end, not average, exposures. The "high end" is defined as that part of the exposure distribution that is above the 90th percentile, but below the 99.9th percentile.

Discussion

Assessing and implementing actions to limit exposure to the hazardous situation and thus reduce risk and monitoring public risky choices under a precautionary principle are the *explicit* results of *legislative* fiat that are inherently costly to society, because budgets are not infinite and resources are constrained. That is, it is possible that the improperly justified allocation under a precautionary principle would divert some scarce resources to pay for an action that is much less protective than its next best alternative, as measured by the opportunity cost of the action not taken.

Clearly, when the magnitude of the consequences is large and the probabilities are also large, quick action becomes imperative. Precautionary principles argue for both circumspection and prevention when the magnitude of the potential adverse event is large or severe, but its probability is relatively small. Arguing for costly action when the causes of the adverse event are either poorly understood or conjectural is unlikely to result in an equitable distribution of risks and benefits, if an action developed on such basis is taken. When the *expected loss* of a choice is small (e.g., a fractional average cancer), some can argue *against* taking immediate precautionary action if the magnitude of the expected value is smaller than some legal minimum (in legal terms, it is no longer *de minimis*). Others look at the magnitude and severity of the probable outcome, but keep probability and magnitude separate and do not multiply them, thus arguing for action even when the probability is small but the magnitude of the adverse consequences is large. These alternative situations reinforce the suggestions that, regardless of their wording, the enunciation of the precautionary principle should contain specific guidance on the strength of the scientific evidence and the explicit recognition of time-dependent changes in the scientific information.

Further Reading

- Cameron, J. (1994). The status of the precautionary principle in international law, in *Interpreting the Precautionary Principle*, T. O'Riordan & J. Cameron, eds, Earthscan Publication, pp. 262–289.
- Chalmers, A. (1988). *What Is This Thing Called Science?* Open University Press.
- Enquete Kommission (1994). *Preventive Measures to Protect the Earth's Atmosphere*, German Bunderstag, Bonn.

- Gerrard, S. (1995). Environmental risk management, in *Environmental Science for Environmental Management*, T. O'Riordan, ed, Longman, pp. 296–316.
- Haigh, N. (1994). The introduction of the precautionary principle into the UK, in *Interpreting the Precautionary Principle*, T. O'Riordan & J. Cameron, eds, Earthscan Publications, pp. 229–251.
- Johnston, P. & Simmonds, M. (1990). Precautionary principle (letter), *Marine Pollution Bulletin* **21**(8), 402.
- Kriebel, D., Tickner, J., Epstein, P., Lemons, J., Levins, R., Loechler, E.L., Quinn, M., Rudel, R. & Schetter, M.S. (2001). The precautionary principle in environmental science, *Environmental Health Perspectives* **109**, 871–876.
- Kuhn, T. (1962). *The Structure of Scientific Revolutions*, University of Chicago Press.
- Kuhn, T. (1963). Scientific paradigms, in *The Sociology of Science*, B. Barnes, ed, Penguin, pp. 80–103.
- MacGarvin, M. (1994). Precaution, science and the sin of hubris, in *Interpreting the Precautionary Principle*, T. O'Riordan & J. Cameron, eds, Earthscan Publications, pp. 69–101.
- Milne, A. (1993). The perils of green pessimism, *New Scientist* **138**(2400), 31–37.
- O'Riordan, T. (1995). Environmental science on the move, in *Environmental Science for Environmental Management*, T. O'Riordan, ed, Longman, pp. 1–11.
- O'Riordan, T. & Cameron, J. (eds) (1994). *Interpreting the Precautionary Principle*, Earthscan Publications.
- O'Riordan, T. & Jordan, A. (1995). The precautionary principle in contemporary environmental politics, *Environmental Values* **4**(3), 191–212.
- Peirce, C.S. (1932). *Collected Papers of Charles Sanders*, C.H. Peirce & P. Weiss, eds, Harvard University Press.
- Smith, F. (1995). Assessing the political approach to risk management, *Economic Affairs* **16**(1), 11–14.
- Wynne, B. (1992). Uncertainty and environmental learning: reconceiving science in the preventive paradigm, *Global Environmental Change* **2**, 111–127.
- Wynne, B. & Mayer, S. (1993). How science fails the environment, *New Scientist* **138**, 33–35.

Related Articles

Causality/Causation

Cost-Effectiveness Analysis

Risk and the Media

Scientific Uncertainty in Social Debates Around Risk

PAOLO F. RICCI

Premium Calculation and Insurance Pricing

The price of insurance is the monetary value for which two parties agree to exchange risk and “certainty”. There are two commonly encountered situations in which the price of insurance is subject of consideration: when an individual agent (for example, a household), bearing an insurable risk, buys insurance from an insurer at an agreed periodic premium; and when insurance portfolios (that is, a collection of insurance contracts) are traded in the financial industry (e.g., being transferred from an insurer to another insurer or from an insurer to the financial market (securitization)).

Pricing in the former situation is usually referred to as *premium calculation* (see **Insurance Applications of Life Tables**), while pricing in the latter situation is usually referred to as *insurance pricing* (see **Insurance Pricing/Nonlife; Risk Measures and Economic Capital for (Re)insurers; Actuary; Individual Risk Models; Multistate Models for Life Insurance Mathematics; Risk Attitude**), although such a distinction is not unambiguous. In this article, we refrain from such an explicit distinction, although the reader may verify that some of the methods discussed are more applicable to the former situation and other methods are more applicable to the latter situation.

To fix our framework, we state the following definitions and remarks:

Definition 1 (Fundamentals) We fix a measurable space $(\Omega, \mathcal{F})$ where Ω is the outcome space and $\mathcal{F}$ is a $(\sigma\text{-})$ algebra defined on it. A risk is a random variable defined on $(\Omega, \mathcal{F})$; that is, $X: \Omega \rightarrow \mathbb{R}$ is a risk if $X^{-1}((-\infty, x]) \in \mathcal{F}$ for all $x \in \mathbb{R}$. A risk represents the final net loss of a position (contingency) currently held. When $X > 0$, we call it a loss, whereas when $X \leq 0$, we call it a gain. The class of all random variables on $(\Omega, \mathcal{F})$ is denoted by $\mathcal{X}$.

Definition 2 (Premium calculation principle (pricing principle)) A premium (calculation) principle or pricing principle π is a functional assigning a real number to any random variable defined on $(\Omega, \mathcal{F})$; that is, π is a mapping from $\mathcal{X}$ to $\mathbb{R}$.

Remark 1 In general, no integrability conditions need to be imposed on the elements of $\mathcal{X}$. In the absence of integrability conditions, some of the premium principles studied below will be infinite for subclasses of $\mathcal{X}$. Instead of imposing integrability conditions, we may extend the range in the definition of π to $\mathbb{R} \cup \{-\infty, +\infty\}$. In case $\pi[X] = +\infty$, we say that the risk is unacceptable or noninsurable.

Remark 2 Statements and definitions provided below hold for all $X, Y \in \mathcal{X}$ unless mentioned otherwise.

No attempt is made to survey all of the many aspects of premium calculation and insurance pricing. For example, we largely ignore contract theoretical issues, and largely neglect insurer expenses and profitability, as well as solvency margins.

In the recent literature, premium principles are often studied under the guise of risk measures; see Föllmer and Schied [1] and Denuit *et al.* [2] for recent reviews of the risk measures literature.

Classical Theories and Some Generalizations

Classical actuarial pricing of insurance risks mainly relies on the economic theories of decision under uncertainty (see **Mathematics of Risk and Reliability: A Select History; Uncertainty Analysis and Dependence Modeling**), in particular on Von Neumann and Morgenstern’s [3] expected utility theory and Savage’s [4] subjective expected utility theory; the interested reader is referred to the early monographs of Borch [5, 6], Bühlmann [7], Gerber [8], and Goovaerts *et al.* [9] for expositions and applications of this theory in insurance economics and actuarial science. Using the principle of equivalent utility (see **Axiomatic Models of Perceived Risk; Axiomatic Measures of Risk and Risk-Value Models; Risk Measures and Economic Capital for (Re)insurers**) and specifying a utility function, various well-known premium principles can be derived. An important example is the exponential premium principle, obtained by using a (negative) exponential utility function, having a constant rate of risk aversion.

The Principle of Equivalent Utility

Consider an agent who is carrying a certain risk. Henceforth, he will be referred to as the “insured”

even though he might decide not to purchase the insurance policy he is offered. Furthermore, consider an insurance company offering an insurance policy for the risk.

We consider below a simple one-period-one-contract model, in which the insurer's wealth at the beginning of the period is w and the loss incurred due to the insured risk during the period is X , so that the insurer's wealth at the end of the period is $w + \pi[X] - X$. The loss X is uncertain and is here represented by a nonnegative random variable with some distribution $\mathbb{P}[X \leq x]$. We assume that both the insurer and the insured have preferences that comply with the expected utility hypothesis.

Let us first consider the viewpoint of the insurer. For a given utility function $u : \mathbb{R} \rightarrow \mathbb{R}$, we consider the expected utility $\mathbb{E}[u(w + \pi[X] - X)]$. In the Von Neumann–Morgenstern framework (see **Mathematics of Risk and Reliability: A Select History**), the utility function u is subjective, whereas the probability measure is assumed to be objective, known, and given. In the more general framework of Savage [4] (see **Subjective Probability; Clinical Dose–Response Assessment**), the probability measure also can be subjective. In the latter framework, the probability measure need not be based on objective statistical information, but may be based on subjective judgments of the decision situation under consideration, just as u is.

From the expected utility expression, an *equivalent utility principle* (also known under the slight misnomer *zero utility principle* (see **Subjective Probability; Clinical Dose–Response Assessment**)) can be established as follows: for a given random variable X and a given real number w , the equivalent utility premium $\pi^- [X]$ of the insurer is derived by solving

$$\mathbb{E}[u(w + \pi[X] - X)] = u(w) \quad (1)$$

Under the assumption that u is continuous (which is implied by the usual set of axioms characterizing expected utility preferences) and provided that the expectation is finite, a solution exists. If the insurer is risk-averse (which for expected utility preferences is equivalent to the utility function being concave), one easily prove that $\pi^- [X] \geq \mathbb{E}[X]$. The insurer will sell the insurance if and only if he can charge a premium $\pi[X]$ that satisfies $\pi[X] \geq \pi^- [X]$.

Next, we consider the viewpoint of the insured.

will be preferred to full self insurance, which leaves the insured with final wealth $\bar{w} - X$, if and only if $\bar{u}(\bar{w} - \pi[X]) \geq \mathbb{E}[\bar{u}(\bar{w} - X)]$, where $\bar{u}$ denotes the utility function of the insured. The equivalent utility premium $\pi^+ [X]$ is derived by solving

$$\bar{u}(\bar{w} - \pi[X]) = \mathbb{E}[\bar{u}(\bar{w} - X)] \quad (2)$$

The insured will buy the insurance if and only if $\pi[X] \leq \pi^+ [X]$. One easily verifies that a risk-averse insured is willing to pay more than the pure net premium $\mathbb{E}[X]$. An insurance treaty can be signed both by the insurer and by the insured only if the premium satisfies $\pi^- [X] \leq \pi[X] \leq \pi^+ [X]$.

Example 1 Consider an agent whose preferences can be described by an exponential utility function

$$u(x) = \frac{1}{\alpha} (1 - e^{-\alpha x}), \quad \alpha > 0 \quad (3)$$

where α is the coefficient of absolute risk aversion. We find that

$$\pi^+ [X] = \frac{1}{\alpha} \log(m_X(\alpha)) \quad (4)$$

where $m_X(\alpha)$ is the moment generating function of X evaluated in α . We note that for this specific choice of the utility function, $\pi^+ [X]$ does not depend on w . Furthermore, notice that the expression for $\pi^- [X]$ is similar. However, now α corresponds to the risk aversion of the insurer. When $\alpha \downarrow 0$ (i.e., for a risk-neutral agent), the exponential premium reduces to $\mathbb{E}[X]$.

Different specifications of the utility function yield different expressions for $\pi^+ [X]$ and $\pi^- [X]$. It is not guaranteed that an explicit expression for $\pi^+ [X]$ or $\pi^- [X]$ is obtained. In a competing insurance market, the insurer may want to charge the lowest “feasible” premium, in which case π will be set equal to π^- .

Rank-Dependent Expected Utility and Choquet Expected Utility. Since the axiomatization of expected utility theory by Von Neumann and Morgenstern [3] and Savage [4] numerous objections have been raised against it; see e.g., Allais [10] (see **Bayes’ Theorem and Updating of Belief; Utility Function**). Most of these relate to the descriptive power of the theory, that is, to empirical evidence of the extent to which agents’ behavior under risk and uncertainty coincides with expected utility theory.

Motivated by empirical evidence that individuals often violate expected utility, various alternative theories have emerged, usually termed as *nonexpected utility theories*; see Sugden [11] for a review.

A prominent example is the *rank-dependent expected utility* (RDEU) theory (see **Risk Attitude**) introduced by Quiggin [12] under the guise of *anticipated utility theory*, and by Yaari [13] for the special case of linear utility. A more general theory for decision under uncertainty (rather than decision under risk), which is known as *Choquet expected utility* (CEU) theory, is due to Schmeidler [14]; see also Choquet [15], Greco [16], Schmeidler [17], and Denneberg [18].

Consider an economic agent who is to decide whether or not to hold a lottery (prospect) denoted by the random variable V . Under a similar set of axioms as used to axiomatize expected utility theory, but with a modified additivity axiom, Yaari [13] showed that the amount of money c at which the decision maker is indifferent between holding the lottery V or receiving the amount c with certainty can be represented as follows (in his original work, Yaari [13] restricts attention to random variables supported on the unit interval; also, Yaari [13] imposes a strong continuity axiom to ensure that g is continuous; we present here a more general version):

$$c = - \int_{-\infty}^0 (1 - g(\bar{F}_V(v))) dv + \int_0^{+\infty} g(\bar{F}_V(v)) dv \quad (5)$$

with $\bar{F}_V(v) := 1 - \mathbb{P}[V \leq v]$. Here, the function $g : [0, 1] \rightarrow [0, 1]$ is nondecreasing and satisfies $g(0) = 0$ and $g(1) = 1$. It is known as a *distortion function*. Provided that g is right continuous, that $\lim_{x \rightarrow +\infty} xg(\bar{F}(x)) = 0$ and that $\lim_{x \rightarrow -\infty} x(1 - g(\bar{F}(x))) = 0$, expression (5) is an expectation calculated using the distorted distribution function $F_g^*(x) := 1 - g(\bar{F}(x))$, that is, $c = \int_{-\infty}^{+\infty} x dF_g^*(x)$. In the economics literature, c is usually referred to as the *certainty equivalent*. Applying the equivalent utility principle within the RDEU framework amounts to solving

$$\int_{-\infty}^{+\infty} (w + \pi[X] - X) dF_g^*(x) = w \quad (6)$$

from which we obtain the *distortion premium principle* (see **Insurance Pricing/Nonlife; Risk Measures and Economic Capital for (Re)insurers**):

$$\pi^-[X] = \int_{-\infty}^{+\infty} x dF_g^*(x) \quad (7)$$

with $\bar{g}(p) = 1 - g(1 - p)$, $0 \leq p \leq 1$, being the *dual distortion function*. The premium principle (7) and its appealing properties were studied by Wang [19] and Wang *et al.* [20].

Example 2 (PH distortion) Consider, as an example, the proportional hazard (PH) distortion function given by $\bar{g}(x) = x^{1/\alpha}$, $\alpha \geq 1$, advocated by Wang [19] and Wang *et al.* [20]. The value of the parameter α determines the degree of risk aversion: the larger the value of α , the larger the risk aversion with $\alpha = 1$ corresponding to the nondistorted (base) case. In this case,

$$\pi^-[X] = - \int_{-\infty}^0 (1 - (\bar{F}_X(x))^{1/\alpha}) dx + \int_0^{+\infty} (\bar{F}_X(x))^{1/\alpha} dx \quad (8)$$

The decumulative distribution function (or tail distribution function) $\bar{F}_X$ provides a natural insight into how to distribute insurance premiums between different layers of insurance, since the net premium for a layer $(a, a + b]$ can be computed as

$$\int_a^{a+b} \bar{F}_X(x) dx \quad (9)$$

An important concept in the axiomatization of RDEU as well as CEU is *comonotonicity* (see **Comonotonicity**). We state the following definition:

Definition 3 (Comonotonicity) A pair of random variables $X, Y : \Omega \rightarrow (-\infty, +\infty)$ is comonotonic if

- (a) there is no pair $\omega_1, \omega_2 \in \Omega$ such that $X(\omega_1) < X(\omega_2)$ while $Y(\omega_1) > Y(\omega_2)$;
- (b) there exists a random variable $Z : \Omega \rightarrow (-\infty, +\infty)$ and nondecreasing functions f, g such that

$$\begin{aligned} X(\omega) &= f(Z(\omega)), \\ Y(\omega) &= g(Z(\omega)), \quad \text{for all } \omega \in \Omega \end{aligned} \quad (10)$$

As we will see below, the premium principle (7) is additive for sums of comonotonic risks.

Another prominent example of a nonexpected utility theory is *maximin expected utility theory* introduced by Gilboa and Schmeidler [21]. It involves the evaluation of worst case scenarios. Laeven [22] proposes a theory that unifies CEU and maximin expected utility theory for linear utility.

Equilibrium Pricing and Pareto Optimality

Bühlmann [23] argued that in contrast to classical premium principles that depend on characteristics of the risk on its own, it is more realistic to consider premium principles that take into account general market conditions (e.g., aggregate risk, aggregate wealth, dependencies between the individual risk and general market risk). Bühlmann derived an *economic premium principle* (see **Optimal Risk-Sharing and Deductibles in Insurance**) in which a functional $\pi[\cdot, \cdot]$ assigns to a random variable X representing the risk to be insured and a random variable Z representing general market risk, a real number π ; that is, $\pi : (\mathcal{X}, \mathcal{Z}) \rightarrow \mathbb{R}$. We will briefly review Bühlmann's economic premium principle; it is based on general equilibrium theory, a main branch in microeconomics.

Consider a pure exchange market with n agents. Think of the agents as being buyers or sellers of insurance policies. Each agent is characterized by

1. a utility function $u_i(x)$, with $u'_i(x) > 0$ and $u''_i(x) \leq 0$, $i = 1, \dots, n$;
2. an initial wealth w_i , $i = 1, \dots, n$.

Each agent faces an exogenous potential loss according to an individual risk function $X_i(\omega)$, representing the payment due by agent i if state ω occurs. In terms of insurance, $X_i(\omega)$ represents the risk of agent i before (re)insurance. Furthermore, each agent buys a *risk exchange* (see **Default Correlation**) denoted by an individual risk exchange function $Y_i(\omega)$, representing the payment received by agent i if ω occurs. The market price of Y_i is denoted by $\pi[Y_i, \cdot]$. Bühlmann used the concept of a pricing density (pricing kernel, state price deflator) $\varphi : \Omega \rightarrow \mathbb{R}$ so that

$$\pi[Y_i, \cdot] = \int_{\Omega} Y_i(\omega) \varphi(\omega) d\mathbb{P}(\omega) \quad (11)$$

for a fixed probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$. An equilibrium is then defined as the pair $(\varphi^*, \mathbf{Y}^*)$ for which:

1.

$$\forall i, \int_{\Omega} u_i \left(w_i - X_i(\omega) + Y_i^*(\omega) - \int_{\Omega} Y_i^*(\omega') \varphi^*(\omega') d\mathbb{P}(\omega') \right) d\mathbb{P}(\omega) \quad (12)$$

maximal for all possible choices of the exchange function $Y_i(\omega)$;

2.

$$\sum_{i=1}^n Y_i^*(\omega) = 0, \quad \forall \omega \in \Omega \quad (13)$$

φ^* is called the *equilibrium pricing density* and $\mathbf{Y}^*$ is called the *equilibrium risk exchange*.

Assume that the utility function of agent i is given by

$$\frac{1}{\alpha_i} (1 - e^{-\alpha_i x}) \quad (14)$$

where $\alpha_i > 0$ denotes the coefficient of absolute risk aversion of agent i and $(1/\alpha_i)$ is his risk tolerance. Solving the equilibrium conditions, Bühlmann obtained the following expression for the equilibrium pricing density:

$$\varphi^*(\omega) = \frac{e^{\alpha Z(\omega)}}{\mathbb{E}[e^{\alpha Z}]}; \quad Z(\omega) = \sum_{i=1}^n X_i(\omega) \quad (15)$$

Equation (15) determines an economic premium principle for any random loss X , namely

$$\pi[X, Z] = \frac{\mathbb{E}[X e^{\alpha Z}]}{\mathbb{E}[e^{\alpha Z}]} \quad (16)$$

Notice that if X and $Z - X$ are independent, a *standard* premium principle is obtained:

$$\begin{aligned} \pi[X, Z] &= \frac{\mathbb{E}[X e^{\alpha X}] \mathbb{E}[e^{\alpha(Z-X)}]}{\mathbb{E}[e^{\alpha X}] \mathbb{E}[e^{\alpha(Z-X)}]} \\ &= \frac{\mathbb{E}[X e^{\alpha X}]}{\mathbb{E}[e^{\alpha X}]} = \pi[X] \end{aligned} \quad (17)$$

This is the time-honored *Esscher premium* (see **Insurance Pricing/Nonlife**). Here, $\frac{1}{\alpha} = \sum_{i=1}^n \frac{1}{\alpha_i}$ can

be interpreted as the risk tolerance of the market as a whole. Note that if the insurance market would contain a large number of agents, the value of $\frac{1}{\alpha} = \sum_{i=1}^n \frac{1}{\alpha_i}$ would be large, or equivalently, the value of α would be close to zero; hence applying the Esscher principle would yield the net premium $\mathbb{E}[X]$.

Remark 3 In equilibrium, only systematic risk requires a risk loading. To see this, suppose that a risk X can be decomposed as follows:

$$X = X_s + X_n \quad (18)$$

where X_s is the systematic risk that is comonotonic with Z , and X_n is the nonsystematic or idiosyncratic risk that is independent of Z (and X_s). Substituting this decomposition in (16), we obtain

$$\pi[X, Z] = \mathbb{E}[X_n] + \frac{\mathbb{E}[X_s e^{\alpha Z}]}{\mathbb{E}[e^{\alpha Z}]} \quad (19)$$

The derived equilibrium coincides with a *Pareto optimal risk exchange* (see **Individual Risk Models**) (Borch [5, 24]).

Bühlmann [25] extended his economic premium principle, allowing general utility functions for market participants. He derived the result that under the weak assumption of concavity of utility functions, equilibrium prices are locally equal to the equilibrium prices derived under the assumption of exponential utility functions. Globally, equilibrium prices are similar, the only difference being that the coefficient of absolute risk aversion is no longer constant, but depends on the agent's individual wealth. The interested reader is referred to Iwaki *et al.* [26] for an extension of the above model to a multiperiod setting. Connections between equations (17) and (7) are studied in Wang [27].

Arbitrage-Free Pricing

In the past decade, we have seen the emergence of a range of financial instruments with underlying risks that are traditionally considered to be “insurance risks”. Examples are catastrophe derivatives and weather derivatives. Such developments on the interface between insurance and finance have put forward questions on appropriate pricing principles for these instruments.

As explained above, traditional actuarial pricing of insurance relies on economic theories of decision

under uncertainty. Financial pricing of contingent claims mainly relies on risk-neutral valuation (see **Equity-Linked Life Insurance; Options and Guarantees in Life Insurance; Risk-Neutral Pricing; Importance and Relevance; Volatility Smile**) and applies an equivalent martingale measure as a risk-adjusted operator; see among many others Duffie [28], Bingham and Kiesel [29], Björk [30], and Protter [31] for textbook treatments. Risk-neutral valuation can be justified within a no-arbitrage setting. For the original work on the relation between the condition of no arbitrage and the existence of an equivalent martingale measure, we refer to Harrison and Kreps [32] and Harrison and Pliska [33]. The basic idea of valuation by adjusting the primary asset process is from Cox and Ross [34]. As is well-known, arbitrage-free pricing may well fit in an equilibrium pricing framework; see e.g., Iwaki *et al.* [26] in an insurance context.

While the no-arbitrage assumption is questionable for traditional insurance markets (see a.o. the discussion in Venter [35, 36] and Albrecht [37]), one may argue that it does hold for insurance products that are traded on the financial markets (securitized insurance risks). Also, especially in life insurance, a significant part of the risk borne is financial risk (interest rate risk, equity risk) rather than pure insurance risk; see e.g., Brennan and Schwartz [38] and Pelsser and Laeven [39]. Arbitrage-free valuation is the natural device for the valuation of the financial portion of the insurance risk portfolio. We focus attention on pure insurance risk.

Several problems arise in applying financial no-arbitrage pricing methodology to securitized insurance risks. The principal problem is that of market incompleteness when the underlying assets are not traded. Indeed, most securitized insurance risks are *index-based* rather than (traded) asset-based. Market incompleteness implies that the contingent claim process cannot be hedged, and therefore no-arbitrage arguments do not, in general, supply a unique equivalent martingale measure for risk-neutral valuation. The problem of market incompleteness proves to be even more serious when insurance processes are assumed to be stochastic jump processes. In the incomplete financial market, there exist an infinite number of equivalent martingale measures and hence, an infinite collection of prices.

The Esscher transform has proven to be a natural candidate and a valuable tool for the pricing of

insurance and financial products. In their seminal work, Gerber and Shiu [40, 41] show that the Esscher transform can be employed to price derivative securities if the logarithms of the underlying asset process follow a stochastic process with stationary and independent increments (Lévy processes). Whereas the Esscher transform of the corresponding actuarial premium principle establishes a change of measure for a random variable, here the Esscher transform is defined more generally as a change of measure for certain stochastic processes. The idea is to choose the Esscher parameter (or in the case of multi assets, the parameter vector) such that the discounted price process of each underlying asset becomes a martingale under the Esscher transformed probability measure. In the case where the equivalent martingale measure is unique, it is obtained by the Esscher transform. In the case where the equivalent martingale measure is not unique (the usual case in insurance), the Esscher transformed probability measure can be justified if there is a representative agent maximizing his expected utility with respect to a power utility function. Inspired by this, Bühlmann *et al.* [42] more generally use *conditional Esscher transforms* to construct equivalent martingale measures for classes of semimartingales; see also Kalssen and Shyriaev [43] and Jacod and Shyriaev [44].

In Goovaerts and Laeven [45], the approach of establishing risk evaluation mechanisms by means of an axiomatic characterization is used to characterize a pricing mechanism that can generate approximate-arbitrage-free financial derivative prices. The price representation derived, involves a probability measure transform that is closely related to the Esscher transform; it is called the *Esscher–Girsanov transform*. In a financial market in which the primary asset price is represented by a general stochastic differential equation, the price mechanism based on the Esscher–Girsanov transform can generate approximate-arbitrage-free financial derivative prices.

We briefly outline below the simple model of Gerber and Shiu [40, 41] for a single primary risk. Assume that there is a stochastic process $\{X(t)\}_{t \geq 0}$ with independent and stationary increments and $X(0) = 0$, such that

$$S(t) = S(0)e^{X(t)}, \quad t \geq 0 \quad (20)$$

Assume, furthermore, that the moment generating function of $X(t)$, denoted by $m_X(\alpha, t)$, exists for all

$\alpha > 0$ and all $t > 0$. Notice that the stochastic process

$$\{e^{\alpha X(t)} m_X(\alpha, 1)^{-t}\}_{t \geq 0} \quad (21)$$

is a positive martingale and can be used to establish a change of probability measure. Gerber and Shiu [40, 41] then define the risk-neutral Esscher measure of parameter α^* such that the process

$$\{e^{-rt} S(t)\}_{t \geq 0} \quad (22)$$

is a martingale with respect to the Esscher measure with parameter α^* ; here r denotes the deterministic risk-free rate of interest. Let us denote the Esscher measure with parameter α by $\mathbb{Q}(\alpha)$ and the true probability measure by $\mathbb{P}$. From the martingale condition

$$\begin{aligned} S(0) &= \mathbb{E}^{\mathbb{Q}(\alpha^*)} [e^{-rt} S(t)] \\ &= \mathbb{E}^{\mathbb{P}} [e^{-rt} e^{\alpha^* X(t)} m_X(\alpha^*, 1)^{-t} S(t)] \end{aligned} \quad (23)$$

we obtain

$$\begin{aligned} e^{rt} &= \mathbb{E}^{\mathbb{P}} [e^{(1+\alpha^*)X(t)} m_X(\alpha^*, 1)^{-t}] \\ &= \left(\frac{m_X(1 + \alpha^*, 1)}{m_X(\alpha^*, 1)} \right)^t \end{aligned} \quad (24)$$

or equivalently

$$\mathbb{E} [e^{(\alpha^*+1)X(1)}] = \mathbb{E} [e^{\alpha^* X(1)+r}] \quad (25)$$

Notice that the parameter α^* is unique since

$$\frac{m_X(1 + \alpha, 1)}{m_X(\alpha, 1)} \quad (26)$$

is strictly increasing in α .

A martingale approach to premium calculation for the special case of compound Poisson processes, the classical risk process in insurance, was studied by Delbaen and Haezendonck [46].

Premium Principles and Their Properties

Many of the (families of) premium principles that emerge from the theories discussed in the section “Classical Theories and Some Generalizations” can be (directly) characterized axiomatically. The general purpose of an axiomatic characterization is to demonstrate what are the essential assumptions to be imposed and what are relevant parameters or concepts to be determined. A premium principle is appropriate if and only if its characterizing axioms are. Axiomatizations can be used not only to justify a premium principle, but also to criticize it.

A systematic study of the properties of premium calculation principles and their axiomatic characterizations was pioneered by Goovaerts *et al.* [9]. We list below various properties that premium principles may (or may not) satisfy.

Properties of Premium Principles

Definition 4 (Law invariance (independence, objectivity)) π is law invariant if $\pi[X] = \pi[Y]$ when $\mathbb{P}[X \leq x] = \mathbb{P}[Y \leq x]$ for all real x .

Law invariance means that for a given risk X , the premium $\pi[X]$ depends on X only *via* the distribution function $\mathbb{P}[X \leq x]$. We note that, to have equal distributions, the random variables X and Y need not, in general, be defined on the same measurable space.

Definition 5 (Monotonicity) π is monotonic if $\pi[X] \leq \pi[Y]$ when $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$. π is $\mathbb{P}$ -monotonic if $\pi[X] \leq \pi[Y]$ when $X \leq Y$ $\mathbb{P}$ -almost surely.

Definition 6 (Preserving first-order stochastic dominance (FOSD)) π preserves FOSD if $\pi[X] \leq \pi[Y]$ when $\mathbb{P}[X \leq x] \geq \mathbb{P}[Y \leq x]$ for all $x \in \mathbb{R}$.

Law invariance together with monotonicity implies preserving FOSD.

Definition 7 (Preserving stop-loss (SL) order) π preserves SL order if $\pi[X] \leq \pi[Y]$ when $\mathbb{E}[(X - d)_+] \leq \mathbb{E}[(Y - d)_+]$ for all $d \in \mathbb{R}$.

We note that preserving SL implies preserving FOSD.

It is well known that if the utility function is nondecreasing, which is implied by the monotonicity axiom that is imposed to axiomatize expected utility, then $X \leq_{\text{FOSD}} Y$ implies ordered equivalent utility premiums. If, moreover, the expected utility maximizer is risk-averse, which is equivalent to concavity of the utility function, then $X \leq_{\text{SL}} Y$ implies ordered equivalent utility premiums; see Hadar and Russell [47] and Rothschild and Stiglitz [48] for seminal work on the ordering of risks within the framework of expected utility theory. For an extensive account on stochastic ordering in actuarial science, see Denuit *et al.* [49].

Definition 8 (Risk loading) π induces a risk loading if $\pi[X] \geq \mathbb{E}[X]$.

Definition 9 (Not unjustified) π is not unjustified if $\pi[c] = c$ for all real c .

Definition 10 (No rip off (Nonexcessive loading)) $\pi[X] \leq \text{esssup}[X]$.

Notice that $\mathbb{P}$ -monotonicity implies no rip-off.

Definition 11 (Additivity) π is additive if $\pi[X + Y] = \pi[X] + \pi[Y]$.

The assumption of no-arbitrage implies that the pricing rule is additive.

Definition 12 (Translation invariance (translation equivariance, translativity, consistency)) π is translation invariant if $\pi[X + c] = \pi[X] + c$ for all real c .

Definition 13 (Positive homogeneity (scale invariance, scale equivariance)) π is positively homogeneous if $\pi[aX] = a\pi[X]$ for all $a \geq 0$.

Definition 14 (Subadditivity respectively Super-additivity) π is subadditive (respectively super-additive) if $\pi[X + Y] \leq \pi[X] + \pi[Y]$ (respectively $\pi[X + Y] \geq \pi[X] + \pi[Y]$).

Definition 15 (Convexity) π is convex if $\pi[\alpha X + (1 - \alpha)Y] \leq \alpha\pi[X] + (1 - \alpha)\pi[Y]$ for all $\alpha \in (0, 1)$.

Notice that subadditivity implies convexity. See Deprez and Gerber [50] for an early account of convex premium principles.

Definition 16 (Independent additivity) π is independent additive if $\pi[X + Y] = \pi[X] + \pi[Y]$ when X and Y are independent.

See Gerber [51] and Goovaerts *et al.* [52] for axiomatizations of independent additive premium principles.

Definition 17 (Comonotonic additivity) π is comonotonic additive if $\pi[X + Y] = \pi[X] + \pi[Y]$ when X and Y are comonotonic.

Definition 18 (Iterativity) π is iterative if $\pi[X] = \pi[\pi[X|Y]]$.

In addition to the above-mentioned properties, axiomatic characterizations of premium principles often impose technical (mainly continuity) conditions

required for obtaining mathematical proofs. Such conditions appear in various forms (e.g., if X_n converges weakly to X then $\pi[X_n] \rightarrow \pi[X]$), and are usually difficult to interpret (justify) economically.

A Plethora of Premium Principles

In this section, we list many well-known premium principles and tabulate which of the properties from the section “Properties of Premium Principles” above are satisfied by them (*see Insurance Pricing/Nonlife; Securitization/Life*). The theories presented in the section “Classical Theories and Some Generalizations” are not mutually exclusive. Some premium principles (Esscher) arise from more than one theory. Other premium principles (Dutch) instead are not directly based on any of the above-mentioned theories nor are they based on an axiomatic characterization, but rather on the nice properties that they exhibit.

Definition 19 (Expected value principle) The expected value principle is given by

$$\pi[X] = (1 + \lambda)\mathbb{E}[X], \quad \lambda \geq 0 \quad (27)$$

If $\lambda = 0$, the net premium is obtained. The net premium can be justified as follows: for the risk $X := \sum_{i=1}^n X_i$ of a large homogeneous portfolio, with $X_i \sim \text{iid}$ and $\mathbb{E}[X_i] = \mu$, we have, on the basis of the law of large numbers,

$$\lim_{n \rightarrow \infty} \mathbb{P}[|X - n\mu| < \varepsilon] = 1 \quad (28)$$

for all $\varepsilon > 0$. Hence, in case the net premium μ is charged for each risk transfer X_i , the aggregate premium income $n\mu$ is sufficient to cover X with probability 1.

However, in the expected utility framework, the net premium suffices only for a risk-neutral insurer. Moreover, it follows from ruin theory that if no loading is applied, ruin will occur with certainty. In this spirit, the loading factor λ can be determined by setting sufficiently protective solvency margins. Such solvency margins may be derived from ruin estimates of the underlying risk process over a given period of time; see e.g., Gerber [8] or Kaas *et al.* [53]. This applies in a similar way to some of the premium principles discussed below. For $\lambda > 0$, the loading margin increases with the expected value.

Definition 20 (Equivalent utility (zero utility) principle) See section titled “The Principle of Equivalent Utility”.

Definition 21 (Distortion principle) See section titled “The Principle of Equivalent Utility”.

Definition 22 (Mean value principle) For a given nondecreasing and nonnegative function f on $\mathbb{R}$ the mean value principle is the root of

$$f(\pi) = \mathbb{E}[f(X)] \quad (29)$$

Definition 23 (Variance principle) The variance principle is given by

$$\pi[X] = \mathbb{E}[X] + \lambda \text{Var}[X], \quad \lambda > 0 \quad (30)$$

Definition 24 (Standard deviation principle) The standard deviation principle is given by

$$\pi[X] = \mathbb{E}[X] + \lambda \sqrt{\text{Var}[X]}, \quad \lambda > 0 \quad (31)$$

See Denneberg [54] for critical comments on the standard deviation principle, advocating the use of the absolute deviation rather than the standard deviation. Dynamic versions of the variance and standard deviation principles in an economic environment are studied by Schweizer [55] and Møller [56].

Definition 25 (Exponential principle) The exponential principle is given by

$$\pi[X] = \frac{1}{\alpha} \log \mathbb{E}[\exp(\alpha X)], \quad \alpha > 0 \quad (32)$$

See Gerber [51], Bühlmann [57], Goovaerts *et al.* [52], and Young and Zariphopoulou [58] for axiomatic characterizations and applications of the exponential principle.

Consider a random variable X and a real number $\alpha > 0$, for which $\mathbb{E}[e^{\alpha X}]$ exists. The positive random variable

$$\frac{e^{\alpha X}}{\mathbb{E}[e^{\alpha X}]} \quad (33)$$

with expectation equal to one, can be used as a Radon–Nikodym derivative to establish a change of probability measure. The probability measure obtained thus is called the *Esscher* measure and is *equivalent* to the original probability measure (i.e.,

the probability measures are mutually absolutely continuous). The Esscher transform was first introduced by the Swedish actuary F. Esscher; see Esscher [59].

Definition 26 (Esscher Principle) The Esscher premium is given by

$$\pi[X] = \frac{\mathbb{E}[Xe^{\alpha X}]}{\mathbb{E}[e^{\alpha X}]}, \quad \alpha > 0 \quad (34)$$

As an alternative representation, consider the cumulant generating function (cgf) of a random variable X :

$$\kappa_X(\alpha) := \log(\mathbb{E}[e^{\alpha X}]) \quad (35)$$

The first and second order derivatives of the cgf with respect to α are

$$\kappa'_X(\alpha) = \frac{\mathbb{E}[Xe^{\alpha X}]}{\mathbb{E}[e^{\alpha X}]} \quad (36)$$

and

$$\kappa''_X(\alpha) = \frac{\mathbb{E}[X^2 e^{\alpha X}]}{\mathbb{E}[e^{\alpha X}]} - \left(\frac{\mathbb{E}[Xe^{\alpha X}]}{\mathbb{E}[e^{\alpha X}]} \right)^2 \quad (37)$$

One can interpret $\kappa'_X(\alpha)$ (or the Esscher premium) and $\kappa''_X(\alpha)$ as the mean and the variance of the random variable X under the Esscher transformed probability measure.

Gerber and Goovaerts [60] establish an axiomatic characterization of an independent additive premium principle that involves a mixture of Esscher transforms. A drawback of both the mixed and nonmixed Esscher premium is that it is not monotonic; see Gerber [61] and Van Heerwaarden *et al.* [62]. Goovaerts *et al.* [52] establish an axiomatic characterization of risk measures that are additive for independent random variables, including an axiom that guarantees monotonicity. The premium principle obtained is a mixture of exponential premiums.

Definition 27 (Swiss principle) For a given non-negative and nondecreasing function w on $\mathbb{R}$ and a given parameter $0 \leq p \leq 1$ the Swiss premium is the root of

$$\mathbb{E}[w(X - p\pi)] = w((1 - p)\pi) \quad (38)$$

Notice that the Swiss principle includes both the equivalent utility principle and the mean value

principle as special cases; see Gerber [51], Bühlmann *et al.* [63] and Goovaerts *et al.* [9].

Definition 28 (Dutch principle) The Dutch principle is given by

$$\begin{aligned} \pi[X] &= \mathbb{E}[X] + \theta \mathbb{E}[(X - \alpha \mathbb{E}[X])_+], \quad \alpha \geq 1, \\ 0 &< \theta \leq 1 \end{aligned} \quad (39)$$

The Dutch premium principle was introduced by Van Heerwaarden and Kaas [64].

Definition 29 (Orlicz principle) Let $X \geq 0$. For a given normalized Young function ψ on $\mathbb{R}_+ \cup \{0\}$, the Orlicz premium is the root of

$$\mathbb{E}[\psi(X/\pi)] = 1 \quad (40)$$

The Orlicz principle is a multiplicative alternative to the equivalent utility principle; see Haezendonck and Goovaerts [65].

All premium principles listed in Table 1 are law invariant. All principles are not unjustified, except for the expected value principle with $\lambda > 0$.

A Generalized Markov Inequality

Many of the premium principles discussed above can be derived in a unified approach by minimizing a generalized Markov inequality; see Goovaerts *et al.* [68]. This approach involves two exogenous functions $v(\cdot)$ and $\phi(\cdot, \cdot)$ and an exogenous parameter $\alpha \leq 1$. Assuming that v is nonnegative and nondecreasing and that ϕ satisfies $\phi(x, \pi) \geq 1_{\{x > \pi\}}$, one easily proves the following generalized Markov inequality:

$$\mathbb{P}[X > \pi] \leq \frac{\mathbb{E}[\phi(X, \pi)v(X)]}{\mathbb{E}[v(X)]} \quad (41)$$

We now consider the case where the upper bound in the above inequality reaches a given confidence level $\alpha \leq 1$:

$$\mathbb{P}[X > \pi] \leq \frac{\mathbb{E}[\phi(X, \pi)v(X)]}{\mathbb{E}[v(X)]} = \alpha \leq 1 \quad (42)$$

When $\alpha = 1$, the equation

$$\frac{\mathbb{E}[\phi(X, \pi)v(X)]}{\mathbb{E}[v(X)]} = 1 \quad (43)$$

Table 1 Properties of premium principles

<i>Principle</i>	$M^{(a)}$	$\Pi^{(b)}$	$PH^{(c)}$	$A^{(d)}$	$SA^{(e)}$	$C^{(f)}$	$IA^{(g)}$	$CA^{(h)}$	$SL^{(i)}$	$NR^{(j)}$	$RL^{(k)}$	$I^{(l)}$
Expected value	+	$-(+ \text{ for } \lambda = 0)$	+	+	+	+	+	+	+	$-(+ \text{ for } \lambda = 0)$	+	$-(+ \text{ for } \lambda = 0)$
Equivalent utility	+	+	-	-	-	-	-	-	+	+	+	-
Distortion	+	+	+	-	+	+	-	+	+	+	+	-
Mean value	+	-	-	-	-	-	-	-	+	+	+	+
Variance	-	+	-	-	-	-	+	-	-	-	+	-
Standard deviation	-	+	+	-	+	+	-	-	-	-	+	-
Exponential	+	+	-	-	-	+	+	-	+	+	+	+
Esscher	-	+	-	-	-	-	+	-	-	+	+	-
Swiss	+	-	-	-	-	-	-	-	+	+	+	-
Dutch	+	-	+	-	+	+	-	-	+	+	+	-
Orlicz	+	$-(m)$	+	-	+	+	-	-	+	+	+	-

^(a) M: monotonicity;
^(b) Π : translation invariance;
^(c) PH: positive homogeneity;
^(d) A: additivity;
^(e) SA: subadditivity;
^(f) C: convexity;
^(g) IA: independent additivity;
^(h) CA: comonotonic additivity;
⁽ⁱ⁾ SL: stop-loss order;
^(j) NR: no rip-off;
^(k) RL: nonnegative risk loading;
^(l) I: iterativity.
^(m) The so-called Haezendonck risk measure, which is an extension of the Orlicz premium (see Goovaerts *et al.* [66] and Bellini and Rosazza Gianin [67]), is translation invariant.

Table 2 Premium principles derived from equation (43)

Premium principle	$v(x)$	$\phi(x, \pi)$
Mean value	1	$\frac{f(x)}{f(\pi)}$
Zero utility	1	$\frac{u(\pi - x)}{u(0)}$
Swiss	1	$\frac{w(x - p\pi)}{w((1 - p)\pi)}$
Orlicz	1	$\psi(x/\pi)$

can generate many well-known premium principles; see Table 2.

For all ϕ , v , by equation (42),

$$\pi \geq \inf\{x \in \mathbb{R} : \mathbb{P}[X \leq x] \geq 1 - \alpha\} \quad (44)$$

which sheds light on the relation between premium principles and quantile-based solvency margins.

Acknowledgments

Roger Laeven acknowledges the financial support of the Netherlands Organization for Scientific Research (NWO Grant No. 42511013 and NWO VENI Grant 2006). Marc Goovaerts acknowledges the financial support of the GOA/02 Grant (“Actuariële, financiële en statistische aspecten van afhankelijkheden in verzekerings- en financiële portefeuilles”).

References

- [1] Föllmer, H. & Schied, A. (2004). *Stochastic Finance*, 2nd Edition, De Gruyter, Berlin.
- [2] Denuit, M., Dhaene, J., Goovaerts, M.J., Kaas, R. & Laeven, R.J.A. (2006). Risk measurement with equivalent utility principles, in *Risk Measures: General Aspects and Applications (special issue)*, *Statistics and Decisions* 24, L. Rüschendorf, ed, John Wiley & Sons, 1–26.
- [3] Von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, 2nd Edition, Princeton University Press, Princeton.
- [4] Savage, L.J. (1954). *The Foundations of Statistics*, John Wiley & Sons, New York.
- [5] Borch, K.H. (1968). *The Economics of Uncertainty*, Princeton University Press, Princeton.
- [6] Borch, K.H. (1974). *The Mathematical Theory of Insurance*, Lexington Books, Toronto.
- [7] Bühlmann, H. (1970). *Mathematical Methods in Risk Theory*, Springer-Verlag, Berlin.
- [8] Gerber, H.U. (1979). *An Introduction to Mathematical Risk Theory*, Huebner Foundation Monograph 8, Homewood, Illinois.
- [9] Goovaerts, M.J., De Vylder, F.E.C. & Haezendonck, J. (1984). *Insurance Premiums*, North Holland, Amsterdam.
- [10] Allais, M. (1953). Le comportement de l’homme rationnel devant le risque: critique des postulats et axiomes de l’école Américaine, *Econometrica* 21, 503–546.
- [11] Sugden, R. (1997). Alternatives to expected utility theory: foundations and concepts, in *Handbook of Expected Utility Theory*, eds, S. Barberà, P.J. Hammond & C. Seidl, Kluwer Academic Publishers, Boston.
- [12] Quiggin, J. (1982). A theory of anticipated utility, *Journal of Economic Behavior and Organization* 3, 323–343.
- [13] Yaari, M.E. (1987). The dual theory of choice under risk, *Econometrica* 55, 95–115.
- [14] Schmeidler, D. (1989). Subjective probability and expected utility without additivity, *Econometrica* 57, 571–587.
- [15] Choquet, G. (1953–1954). Theory of capacities, *Annales de l’Institut Fourier (Grenoble)* 5, 131–295.
- [16] Greco, G. (1982). Sulla rappresentazione di funzionali mediante integrali, *Rendiconti del Seminario Matematico dell’Università di Padova* 66, 21–42.
- [17] Schmeidler, D. (1986). Integral representation without additivity, *Proceedings of the American Mathematical Society* 97, 255–261.
- [18] Denneberg, D. (1994). *Non-Additive Measure and Integral*, Kluwer Academic Publishers, Boston.
- [19] Wang, S.S. (1996). Premium calculation by transforming the layer premium density, *ASTIN Bulletin* 26, 71–92.
- [20] Wang, S.S., Young, V.R. & Panjer, H.H. (1997). Axiomatic characterization of insurance prices, *Insurance: Mathematics and Economics* 21, 173–183.
- [21] Gilboa, I. & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior, *Journal of Mathematical Economics* 18, 141–153.
- [22] Laeven, R.J.A. (2005). *Essays on Risk Measures and Stochastic Dependence*, Tinbergen Institute Research Series 360, Thela Thesis, Amsterdam, Chapter 3.
- [23] Bühlmann, H. (1980). An economic premium principle, *ASTIN Bulletin* 11, 52–60.
- [24] Borch, K.H. (1962). Equilibrium in a reinsurance market, *Econometrica* 3, 424–444.
- [25] Bühlmann, H. (1984). The general economic premium principle, *ASTIN Bulletin* 14, 13–22.
- [26] Iwaki, H., Kijima, M. & Morimoto, Y. (2001). An economic premium principle in a multiperiod economy, *Insurance: Mathematics and Economics* 28, 325–339.
- [27] Wang, S.S. (2003). Equilibrium pricing transforms: new results using Bühlmann’s 1980 economic model, *ASTIN Bulletin* 33, 57–73.
- [28] Duffie, D. (1996). *Dynamic Asset Pricing Theory*, 2nd Edition, Princeton University Press, Princeton.

- [29] Bingham, N.H. & Kiesel, R. (2004). *Risk-Neutral Valuation*, 2nd Edition, Springer, Berlin.
- [30] Björk, T. (2004). *Arbitrage Theory in Continuous Time*, 2nd Edition, Oxford University Press, Oxford.
- [31] Protter, P. (2004). *Stochastic Integration and Differential Equations*, 2nd Edition, Springer, Berlin.
- [32] Harrison, J.M. & Kreps, D.M. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- [33] Harrison, J. & Pliska, S.R. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and Their Applications* **11**, 215–260.
- [34] Cox, J. & Ross, S. (1976). The valuation of options for alternative stochastic processes, *Journal of Financial Economics* **3**, 145–166.
- [35] Venter, G.G. (1991). Premium calculation implications of reinsurance without arbitrage, *ASTIN Bulletin* **21**, 223–230.
- [36] Venter, G.G. (1992). Premium calculation without arbitrage – author’s reply on the note by P. Albrecht, *ASTIN Bulletin* **22**, 255–256.
- [37] Albrecht, P. (1992). Premium calculation without arbitrage? – a note on a contribution by G. Venter, *ASTIN Bulletin* **22**, 247–254.
- [38] Brennan, M.J. & Schwartz, E.S. (1976). The pricing of equity-linked life insurance policies with an asset value guarantee, *Journal of Financial Economics* **3**, 195–213.
- [39] Pelsser, A.A.J. & Laeven, R.J.A. (2006). *Optimal Dividends and Alm Under Unhedgeable Risk*, University of Amsterdam, Working Paper.
- [40] Gerber, H.U. & Shiu, E.S.W. (1994). Option pricing by Esscher transforms, *Transactions of the Society of Actuaries* **46**, 99–191.
- [41] Gerber, H.U. & Shiu, E.S.W. (1996). Actuarial bridges to dynamic hedging and option pricing, *Insurance: Mathematics and Economics* **18**, 183–218.
- [42] Bühlmann, H., Delbaen, F., Embrechts, P. & Shiryaev, A.N. (1996). No-arbitrage, change of measure and conditional Esscher transforms, *Centrum voor Wiskunde en Informatica (CWI) Quarterly* **9**, 291–317.
- [43] Kallsen, J. & Shiryaev, A.N. (2002). The cumulant process and Esscher’s change of measure, *Finance and Stochastics* **6**, 397–428.
- [44] Jacod, J. & Shiryaev, A.N. (2003). *Limit Theorems for Stochastic Processes*, 2nd Edition, Section III.7.b, Springer, New York.
- [45] Goovaerts, M.J. & Laeven, R.J.A. (2007). Actuarial risk measures for financial derivative pricing, *Insurance: Mathematics and Economics*, in press.
- [46] Delbaen, F. & Haezendonck, J. (1989). A martingale approach to premium calculation principles in an arbitrage free market, *Insurance: Mathematics and Economics* **8**, 269–277.
- [47] Hadar, J. & Russell, W.R. (1969). Rules for ordering uncertain prospects, *American Economic Review* **59**, 25–34.
- [48] Rothschild, M. & Stiglitz, J.E. (1970). Increasing risk I: a definition, *Journal of Economic Theory* **2**, 225–243.
- [49] Denuit, M., Dhaene, J., Goovaerts, M.J. & Kaas, R. (2005). *Actuarial Theory for Dependent Risks*, John Wiley & Sons, New York.
- [50] Deprez, O. & Gerber, H.U. (1985). On convex principles of premium calculation, *Insurance: Mathematics and Economics* **4**, 179–189.
- [51] Gerber, H.U. (1974). On additive premium calculation principles, *ASTIN Bulletin* **7**, 215–222.
- [52] Goovaerts, M.J., Kaas, R., Laeven, R.J.A. & Tang, Q. (2004). A comonotonic image of independence for additive risk measures, *Insurance: Mathematics and Economics* **35**, 581–594.
- [53] Kaas, R., Goovaerts, M.J., Dhaene, J. & Denuit, M. (2001). *Modern Actuarial Risk Theory*, Kluwer Academic Publishers, Dordrecht.
- [54] Denneberg, D. (1990). Premium calculation: why standard deviation should be replaced by absolute deviation, *ASTIN Bulletin* **20**, 181–190.
- [55] Schweizer, M. (2001). From actuarial to financial valuation principles, *Insurance: Mathematics and Economics* **28**, 31–47.
- [56] Möller, T. (2001). On transformations of actuarial valuation principles, *Insurance: Mathematics and Economics* **28**, 281–303.
- [57] Bühlmann, H. (1985). Premium calculation from top down, *ASTIN Bulletin* **15**, 89–101.
- [58] Young, V.R. & Zariphopoulou, T. (2002). Pricing dynamic insurance risks using the principle of equivalent utility, *Scandinavian Actuarial Journal* **2002**, 246–279.
- [59] Esscher, F. (1932). On the probability function in the collective theory of risk, *Scandinavian Actuarial Journal* **15**, 175–195.
- [60] Gerber, H.U. & Goovaerts, M.J. (1981). On the representation of additive principles of premium calculation, *Scandinavian Actuarial Journal* **4**, 221–227.
- [61] Gerber, H.U. (1981). The Esscher premium principle: a criticism, comment, *ASTIN Bulletin* **12**, 139–140.
- [62] Van Heerwaarden, A.E., Kaas, R. & Goovaerts, M.J. (1989). Properties of the Esscher premium calculation principle, *Insurance: Mathematics and Economics* **8**, 261–267.
- [63] Bühlmann, H., Gagliardi, B., Gerber, H.U. & Straub, E. (1977). Some inequalities for stop-loss premiums, *ASTIN Bulletin* **9**, 75–83.
- [64] Van Heerwaarden, A.E. & Kaas, R. (1992). The Dutch premium principle, *Insurance: Mathematics and Economics* **11**, 129–133.
- [65] Haezendonck, J. & Goovaerts, M.J. (1982). A new premium calculation principle based on Orlicz norms, *Insurance: Mathematics and Economics* **1**, 41–53.
- [66] Goovaerts, M.J., Kaas, R., Dhaene, J. & Tang, Q. (2004). Some new classes of consistent risk measures, *Insurance: Mathematics and Economics* **34**, 505–516.
- [67] Bellini, F. & Rosazza Gianin, E. (2006). *On Haezendonck Risk Measures*, Università di Milano, Bicocca, Working Paper.

- [68] Goovaerts, M.J., Kaas, R., Dhaene, J. & Tang, Q. (2003). A unified approach to generate risk measures, *ASTIN Bulletin* **33**, 173–191.
- Kaas, R., Van Heerwaarden, A.E. & Goovaerts, M.J. (1994). *Ordering of Actuarial Risks, Education Series 1*, CAIRE, Brussels.
- Karatzas, I. & Shreve, S.E. (1988). *Brownian Motion and Stochastic Calculus*, Springer, New York.

Further Reading

- Dionne, G. (ed) (2000). *Handbook of Insurance*, Springer, New York.
- Embrechts, P. (2000). Actuarial versus financial pricing of insurance, *Risk Finance* **1**, 17–26.

ROGER J.A. LAEVEN AND
MARC J. GOOVAERTS

Pricing of Life Insurance Liabilities

Life insurance policies provide benefits contingent on policyholders' lives. Payouts can be nominally fixed or linked to financial variables, they can embed financial guarantees or allow policyholders to participate in the insurer's profits. We will restrict ourselves to the simplest case of nominally fixed benefits to present the main issues caused by randomness in interest rates and mortality rates. In the section titled "Systematic and Unsystematic Mortality Risk", we begin by distinguishing between systematic and unsystematic mortality risk. We explain why and under what assumptions the number of policies has an impact on the riskiness of an insurer's portfolio. The section titled "Risk Associated with Randomness in Interest Rates" extends the previous section by taking into account the time value of money. The section titled "Risk Associated with Size of Benefits" describes the impact of heterogeneity of sums assured on the riskiness of a portfolio. In the remaining sections, we gradually introduce dynamic considerations into our analysis: in particular, the section titled "Annuities" links randomness in lifetimes with liability duration, while the section titled "Benefits Financed through Regular Premiums" examines the interplay between premiums, accumulation of reserves, and benefit payments. In the section "Random Hazard Rates", we briefly describe how randomness in mortality rates can be modeled dynamically. Finally, we show in the section titled "Pricing of Life Insurance Liabilities" how the risks described in the previous sections can be accounted for when pricing insurance contracts or when placing a market-consistent value on a portfolio of life insurance liabilities.

Systematic and Unsystematic Mortality Risk

We first focus on uncertainty in survival and death rates. Let us consider a portfolio of n insureds at the reference time 0. We are interested in the number of survivors, $N_T(n)$, and in the number of cumulated deaths, $D_T(n) \doteq n - N_T(n)$, up to a fixed future date T (see **Estimation of Mortality Rates**

from Insurance Data). Denoting the random residual lifetimes of the insureds by $\tau^1, \dots, \tau^n$, we need to study the random variable

$$N_T(n) = \sum_{i=1}^n 1_{\{\tau^i > T\}} \quad (1)$$

The symbol 1_A denotes the indicator function of the event A : this means $1_{\{\tau^i > T\}}$ is equal to 1 in case the i th insured is alive at time T and equal to 0 in case of death before T .

If each insured belongs to the same risk class (same age, sex, health status, and type of contract), an insurer can reasonably assume that the random times $\tau^1, \dots, \tau^n$ are *identically distributed*. In particular, the probability of being alive at time T can be expressed as

$$\mathbb{P}(\tau^i > T) = p \quad \text{for all } i \quad (2)$$

with p being a number between 0 and 1.

If the lifetimes $\tau^1, \dots, \tau^n$ can also be assumed to be *independent*, then our random variables $N_T(n)$ and $D_T(n)$ are distributed as binomial random variables with parameters, (n, p) and $(n, 1 - p)$, respectively. When the dimension of the portfolio is large, the law of large numbers applies (see [1]), yielding

$$\lim_{n \rightarrow \infty} \frac{N_T(n)}{n} = E\left(\frac{N_T(n)}{n}\right) = \frac{np}{n} = p \quad (3)$$

where $E(\cdot)$ denotes the expectation operator. By the very same assumptions, we can apply the central limit theorem and obtain

$$\frac{N_T(n)}{n} \sim \text{Normal}\left(p, \frac{p(1-p)}{n}\right) \quad (4)$$

for large enough n . Similar results can be written for the proportion of deaths, $D_T(n)/n$, as n grows larger.

The expressions above have a very intuitive interpretation. They imply that a large enough portfolio of insureds leads to

- higher concentration of actual survivors or deaths around their expected values;
- lower probability of extreme outcomes;
- more bell-shaped distribution of survivors and deaths.

If we denote the actual number of survivors at time T by n_T , the risk of n_T being different from $E(N_T(n)) = np$ is called "risk of random

fluctuations” around the expected value, or “process risk” (see [2, 3]). According to equations (3, 4) such risk can be reduced by pooling together a large enough number of individuals, and hence the name *pooling risk* (see **Model Risk**). This risk reduction is based on the fact that fluctuations of outcomes around the mean tend to offset each other for a large enough portfolio: such risk is then an example of *unsystematic risk*, in the sense that it affects different individuals in no systematic way. If we measure risk in terms of standard deviation ($\sigma(\cdot)$), we see from equation (4) that risk reduces at a lower pace than the increase in portfolio dimension, since

$$\lim_{n \rightarrow \infty} \sigma \left(\frac{N_T(n)}{n} \right) = \lim_{n \rightarrow \infty} \sqrt{\frac{p(1-p)}{n}} = 0 \quad (5)$$

As a consequence, the reduction in risk is only of order $\sqrt{n}$. Things become worse if we relax some of the assumptions made so far.

For example, we have assumed that the probability of surviving at time T is well known at contract inception and equal to a deterministic number p . In practice, this approach may be inadequate for two reasons. First, the insurer has no complete knowledge of the risk factors affecting policyholders’ lifetimes, so that it may desirable to have a model allowing for some unobservable variables. Second, the evolution of mortality is a complex nonstationary phenomenon, whose long-term modeling is heavily subject to model/parameter uncertainty. Depending on the specific types of contracts considered and on the length of the time horizon T , the previous caveats can become extremely relevant. If we explicitly allow for randomness in the key parameter p , then things change radically. As an example, assume that $N_T(n)$ depends on a random variable X that cannot be fully known before time T . Keeping the assumption of identical distribution of $\tau^1, \dots, \tau^n$, expression (2) can now be rewritten as

$$\mathbb{P}(\tau^i > T|X) = p(X) \quad \text{for all } i \quad (6)$$

where $p(\cdot)$ is a function of X taking values between 0 and 1. In this case, the lifetimes $\tau^1, \dots, \tau^n$ are no longer independent: the common factor X affects the probability of survival at time T of each insured in the portfolio. We thus say that X represents a *systematic risk*.

The question now is, in the present setup, can we still benefit from the limiting results in equations

(3, 4) we saw in the case of fully known p ? Only to a very limited extent, as we now show. First, if X can be truly thought of as the only risk factor commonly affecting the likelihood of death of the n individuals, we can assume that $\tau^1, \dots, \tau^n$ are *conditionally independent*, given knowledge of X . Second, if we take as fixed a realization x of X , we can exploit the independence and identical distribution of $\tau^1, \dots, \tau^n$ to repeat the reasoning developed in the deterministic case and obtain expression (4) with p replaced by $p(x)$. Thus, the nice properties associated with the law of large numbers and the central limit theorem can now hold only conditional on the value taken by X . Let us see what happens unconditionally instead. From equation (6), by iterating expectations we obtain

$$\mathbb{P}(\tau^i > T) = E(\mathbb{P}(\tau^i > T|X)) = E(p(X)) \quad (7)$$

$$E(N_T(n)) = E(E(N_T(n)|X)) = nE(p(X)) \quad (8)$$

and thus what used to be a given probability is now replaced by the estimate $E(p(X))$. Furthermore, the variance of $N_T(n)$ can now be expressed as follows:

$$\begin{aligned} V(N_T(n)) &= V(E(N_T(n)|X)) + E(V(N_T(n)|X)) \\ &= n^2 V(p(X)) + nE(p(X)(1-p(X))) \end{aligned} \quad (9)$$

where we have used the representation property of variances (see [1]) and the fact that conditional on X , $N_T(n)$ is binomially distributed with parameters $(n, p(X))$. From equation (9), it is easily seen that the standard deviation of $N_T(n)/n$ does not go to zero as n grows larger, since

$$\begin{aligned} \lim_{n \rightarrow \infty} \sigma \left(\frac{N_T(n)}{n} \right) &= \lim_{n \rightarrow \infty} \sqrt{V(p(X)) + \frac{E(p(X)(1-p(X)))}{n}} \\ &= \sigma(p(X)) \end{aligned} \quad (10)$$

As a consequence, no matter how large the portfolio, there always remains an undiversifiable risk component induced by X .

Risk Associated with Randomness in Interest Rates

In the simple portfolio model of the previous section, we simply considered the distribution of the number of survivors (or cumulated deaths) at some future date T . We now introduce monetary benefits and focus on the simple case in which each survivor is entitled to a payment equal to one monetary unit upon survival at T . This means the portfolio of n individuals entails a random liability equal to $N_T(n)$ monetary units at time T . To express it in current money terms, we need to take into account the time value of money.

Suppose that it is possible to deposit cash in a bank account obtaining an instantaneous return equal to a constant $r > 0$. We can thus generate e^{rT} units of money at the final date T by investing one unit at time 0 and continuously reinvesting the proceeds up to time T . If we go back to our portfolio, we can now express the survival benefits in current money terms as

$$e^{-rT} 1_{\{\tau^1 > T\}}, \dots, e^{-rT} 1_{\{\tau^n > T\}} \quad (11)$$

Since each random variable in equation (11) is scaled by a deterministic factor e^{-rT} , the analysis of risk made in the previous section carries over without modifications.

The assumption of a constant interest rate may be quite unrealistic, particularly when the time horizon T is long or if the setup is extended to benefits embedding interest rate guarantees or linked to other financial variables. Let us assume that at each time $t \geq 0$ the risk-free rate is a random variable r_t . The amount of money we can get at time T from depositing 1 in the bank account at 0 and rolling over the proceeds up to T is now a random variable, $e^{\int_0^T r_s ds}$. The present value of the survival benefits as in equation (11) must now be changed accordingly as follows:

$$e^{-\int_0^T r_s ds} 1_{\{\tau^1 > T\}}, \dots, e^{-\int_0^T r_s ds} 1_{\{\tau^n > T\}} \quad (12)$$

The risk exposure is now different from equation (11), because the quantity $e^{-\int_0^T r_s ds}$ represents a common risk factor simultaneously affecting the present value of the benefits payable to each policyholder. For example, if the realization of interest rates from 0 to T takes values lower than the constant r considered in equation (11), then each liability in equation (12)

is higher than the corresponding one in equation (11), independently of the specific policyholder considered.

Risk Associated with Size of Benefits

We have seen that the independence assumption plays a crucial role in allowing the insurer to reduce risk by pooling together several contracts. The hypothesis of identical distribution is also relevant and should not be taken for granted. To show its impact on the portfolio risk exposure, we examine the case of policyholders with similar biometric characteristics taking out the same type of contract but with different levels of benefits.

As a simple example, consider a one-period term assurance providing a benefit C^i at time T upon death of policyholder i between time 0 and T . All residual life times $\tau^1, \dots, \tau^n$ are assumed independent and identically distributed. The random present values of the benefits,

$$e^{-\int_0^T r_s ds} C^1 1_{\{\tau^1 \leq T\}}, \dots, e^{-\int_0^T r_s ds} C^n 1_{\{\tau^n \leq T\}} \quad (13)$$

give rise to the aggregate liability

$$H = e^{-\int_0^T r_s ds} \sum_{i=1}^n C^i 1_{\tau^i \leq T} \quad (14)$$

Let $\bar{C}$ and $\bar{C}^{(2)}$ represent the mean and the second moment of the sums assured:

$$\bar{C} \doteq \frac{\sum_{i=1}^n C_i}{n} \quad \bar{C}^{(2)} \doteq \frac{\sum_{i=1}^n C_i^2}{n} \quad (15)$$

Without loss of generality, we can assume zero interest rates and compute the expected value and variance of equation (14) as follows:

$$E(H) = n(1-p)\bar{C} \quad (16)$$

$$\begin{aligned} V(H) &= n(1-p)p\bar{C}^{(2)} \\ &= n(1-p)p(\bar{C}^2 + V(C)) \end{aligned} \quad (17)$$

with p defined by equation (2). From the second equality in equation (17), we can see that for fixed $\bar{C}$, the variance of the random liability H is minimal when $V(C) = 0$, i.e., when the insureds are homogeneous not only in terms of mortality risk but also

in terms of sums assured. Additional insights can be gathered by computing the coefficient of variation of H , $\sigma(H)/E(H)$, a quantity measuring the degree of dispersion of the distribution of H . Exploiting the first equality in equation (17), we obtain

$$\frac{\sigma(H)}{E(H)} = \sqrt{\frac{p}{1-p}} \frac{\sqrt{\bar{C}^{(2)}}}{\sqrt{n} \bar{C}} \quad (18)$$

which is increasing in $\bar{C}^{(2)}$ for fixed average sum assured $\bar{C}$ and portfolio dimension n . When $C_i = C$ for all i , the expression above reduces to

$$\frac{\sigma(H)}{E(H)} = \sqrt{\frac{p}{1-p}} \frac{1}{\sqrt{n}} \xrightarrow{n \rightarrow \infty} 0 \quad (19)$$

showing that risk can be reduced by increasing the dimension of the portfolio and that the convergence to 0 is of order $\sqrt{n}$ as observed in the section named “Systematic and Unsystematic Mortality Risk”. On the other hand, if one of the n sums assured is large enough, we have $\sqrt{\sum_{i=1}^n C_i^2} / \sum_{i=1}^n C_i \cong 1$ and thus equation (18) becomes

$$\frac{\sigma(H)}{E(H)} = \sqrt{\frac{p}{1-p}} \frac{\sqrt{\sum_{i=1}^n C_i^2}}{\sum_{i=1}^n C_i} \cong \sqrt{\frac{p}{1-p}} \quad (20)$$

showing that the pooling effect is completely prevented by the presence of a single outlier.

Risk Associated with Timing of Benefit Payments

So far we have considered liabilities involving survival benefits payable at a given fixed time. We extend this basic case in two directions: survival benefits payable until death (annuities) and benefits financed by premiums payable over time. In the following text, τ will represent the random residual life time of a generic policyholder (*see Risk Classification/Life*).

Annuities

An annuity is a contract providing a guaranteed stream of regular (e.g., monthly) payments contingent on survival of the policyholder. As an example, consider a contract paying a unitary rate continuously

until death. In current money terms, the stream of benefits is given by the following random variable:

$$\int_0^\tau e^{-\int_0^t r_s ds} dt \quad (21)$$

If the short rate is equal to zero, we can see immediately that equation (21) coincides with the insured’s life expectancy, $E(\tau)$. As a consequence, the longer the life expectancy of the policyholder, the higher the expected liabilities faced by the insurer. In the more general case of nonzero interest rates, an increase in life expectancy leads to an increase in the *duration* of liabilities and hence in the exposure to interest rate movements.

Benefits Financed through Regular Premiums

As in the section titled “Risk Associated with Randomness in Interest Rates”, let us consider a term assurance providing a benefit C in case of death between 0 and T , but suppose now that the benefit is paid at τ rather than at T . The random present value of the benefit is given by

$$e^{-\int_0^\tau r_s ds} C 1_{\{\tau \leq T\}} \quad (22)$$

In exchange for this cover, the insurer could require the policyholder to pay a premium rate $\pi(\cdot)$ continuously over time. Pricing issues are examined in detail in the section titled “Pricing of Life Insurance Liabilities”: for the moment, we can just suppose that the insurer is risk-neutral and that $\pi(\cdot)$ is a function leading to *equivalence*, in current money terms, between the cumulated cashflows from the insurer to the policyholder and those from the policyholder to the insurer. In other words, the following must hold at time 0:

$$\begin{aligned} E \left(\int_0^\tau e^{-\int_0^s r_u du} \pi(s) 1_{\{\tau > s\}} ds \right) \\ = E \left(e^{-\int_0^\tau r_u du} C 1_{\{\tau \leq T\}} \right) \end{aligned} \quad (23)$$

Besides equation (23), additional requirements (see below) are imposed to determine an optimal rate $\pi(\cdot)$. Only in the case of a constant premium rate $\pi(\cdot) \equiv \bar{\pi}$, equation (23) is enough to determine $\bar{\pi}$ uniquely.

Once a $\pi(\cdot)$ satisfying equation (23) is fixed, the equivalence equation does not necessarily hold at any generic time t following 0. Indeed, we have

$$\begin{aligned} V_t + E_t \left(\int_t^T e^{-\int_t^s r_u du} \pi(s) 1_{\{\tau > s\}} ds \right) \\ = E_t \left(e^{-\int_t^\tau r_u du} C 1_{\{t < \tau \leq T\}} \right) \end{aligned} \quad (24)$$

where $E_t(\cdot)$ denotes conditional expectation, given the information available at time t . The balancing term V_t makes sure that equality holds in equation (24) and is better understood if we rearrange equation (24) to obtain

$$\begin{aligned} V_t = E_t \left(e^{-\int_t^\tau r_u du} C 1_{\{t < \tau \leq T\}} \right. \\ \left. - \int_t^T e^{-\int_t^s r_u du} \pi(s) 1_{\{\tau > s\}} ds \right) \end{aligned} \quad (25)$$

so that V_t represents the time- t expected present value of the net outstanding liabilities. V_t is called *prospective reserve* of the contract, because it quantifies the funds needed to meet the insurer's liabilities from time t onward. Usually $\pi(\cdot)$ is chosen so as to make sure that at each $t \in [0, T]$ V_t is nonnegative with a high level of confidence. In other words, $\pi(\cdot)$ ensures that at each future date the expected present value of future benefits is larger than the remaining premiums to be paid. This guarantees that the policyholder is locked into a debt position and has no obvious incentives to walk away from the contract. Another important consequence is that premiums “anticipate” benefits (in expected present value terms) and an adequate investment return can be earned on the reserves before benefits are paid out.

The mechanism through which individual premiums lead to the creation of reserves to support benefit payments relies once again on the pooling of several independent homogeneous risks. Indeed, equation (24) simply relies on expected values: only a large portfolio of policyholders can ensure that, on average, the reserves “released” from those who die plus the premiums from those who survive are enough to pay off benefits and to set aside the reserves needed for the insureds remaining in the portfolio. Again, if the portfolio is not large enough, random fluctuations can lead to insufficient aggregate reserves

and force the insurer to inject capital in the portfolio. But now, contrary to the cases studied before, it is not only the level of deaths (higher or lower than expected) that matters, but also their timing. Indeed, benefit payouts occurring earlier than expected entail more risk than those occurring later, because they deplete reserves and reduce investment income to be earned on them (see [2, 4, 5]).

Random Hazard Rates

The previous sections have gradually introduced dynamic considerations into the analysis of risks affecting life insurance policies. In the same way that interest rates can randomly evolve over time, it is now natural to allow for uncertainty in the evolution of survival probabilities and death rates. One way of doing so is by modeling the likelihood of survival at each time t as a random variable (here we follow [6, 7]; see [8] for a survey of dynamic mortality models). Specifically, for a generic τ we can set

$$\mathbb{P}(\tau > t | \mathcal{F}_t) = e^{-\Gamma_t} \quad (26)$$

where Γ is a nondecreasing process starting from $\Gamma_0 = 0$ and $\mathcal{F}_t$ represents information about the evolution of Γ (but not about the actual occurrence of τ) up to time t . Under these assumptions, the process in equation (26) takes values between 0 and 1 and is decreasing with t , which is exactly what we need to model each point of the survival function of τ as a random variable.

Going back to the portfolio of the section titled “Systematic and Unsystematic Mortality Risk”, suppose that the policyholder's life times $\tau^1, \dots, \tau^n$ are conditionally independent, given Γ_T , and all with the same likelihood of survival as in equation (26). It then follows that $N_T(n)$ is conditionally binomial with parameters $(n, e^{-\Gamma_T})$ and we have, for example,

$$E(N_T(n) | \mathcal{F}_T) = \sum_{i=1}^n \mathbb{P}(\tau^i > T | \mathcal{F}_T) = ne^{-\Gamma_T} \quad (27)$$

If we set $X \doteq \Gamma_T$, from equation (26) we obtain exactly equation (6) with $p(X) = e^{-X}$, so that Γ represents a source of undiversifiable risk as explained in the section titled “Systematic and Unsystematic Mortality Risk”. The advantage of the present setup is that we need not fix our attention on a single date T nor on a “static” risk factor X . Also, we can easily obtain

stochastic counterparts of key actuarial quantities. For example, assume that Γ can be expressed as

$$\Gamma_t = \int_0^t \mu_s ds \quad (28)$$

for some nonnegative process μ . Under suitable assumptions (e.g., when $1_{\{\tau \leq t\}}$ is the first jump of a Cox process), μ can be identified with the random intensity of mortality of the insured (see [7] for details). This means that μ_t can be seen as the time- t instantaneous conditional death probability of an individual. Also, by using equations (26, 28) and iterating expectations we can write

$$\mathbb{P}(\tau > T) = E(\mathbb{P}(\tau > T | \mathcal{F}_T)) = E\left(e^{-\int_0^T \mu_s ds}\right) \quad (29)$$

where we recognize (apart from the expectation) the usual expression for survival probabilities one finds in standard actuarial textbooks (e.g., [4]). More useful implications become apparent when it comes to analyzing financial and demographic risks in an integrated and dynamic fashion. As an example, consider the benefits paid by the term assurance of section titled “Benefits Financed through Regular Premiums”. Under equation (28), the expected value of equation (22) takes the form

$$E\left(\int_0^T e^{-\int_0^t (r_s + \mu_s) ds} C \mu_t dt\right) \quad (30)$$

which can be easily computed via simulation, or even analytically, if r and μ are chosen within a suitable class of processes (see [6, 7, 9, 10]). The probabilistic framework described in this section can be used to extend any static, traditional demographic model to a dynamic setting (e.g., [11]). An extension to multiple cohorts of insureds is studied in [12]. Dealing with expressions like equation (30) is what is most needed when pricing life insurance liabilities, as we show in the next section.

Pricing of Life Insurance Liabilities

In the section titled “Benefits Financed through Regular Premiums”, we have seen how premiums could be computed through the equivalence principle in the case of a risk-neutral insurer. In practice, the equivalence principle is used under a probability measure $\tilde{\mathbb{P}}$ different from $\mathbb{P}$, in order to reflect risk

aversion: $\tilde{\mathbb{P}}$ is suitably *favorable* to the insurer, in that it embeds a profit loading providing compensation for the different types of risks taken on by the insurer.

The choice of $\tilde{\mathbb{P}}$ is generically said to be driven by *prudence*: the latter is defined in different ways, e.g., as the ability to generate profits over time or to maintain the company solvent with a sufficiently high probability. Since we have already introduced prospective reserves in the section titled “Benefits Financed through Regular Premiums”, it is natural to define prudence in terms of reserves computed under different probability measures. We say that a measure $\tilde{\mathbb{P}}$ is prudent if the valuation of the net liabilities under $\tilde{\mathbb{P}}$ is greater than what we would obtain by employing the objective (realistic) measure $\mathbb{P}$ (e.g., [4, 5, 13, 14]). In other words, this is true if

$$\tilde{V}_t \geq V_t \quad \text{for all } t \quad (31)$$

where we have extended the notation introduced in the section titled “Benefits Financed through Regular Premiums” to reserves computed under different probability measures. In particular, $\tilde{V}_t$ is obtained by taking expectation under $\tilde{\mathbb{P}}$ in equation (25).

Condition (31) shows that $\tilde{\mathbb{P}}$ is pessimistic enough to overestimate the insurer’s net liabilities at each point in time. This involves setting aside a larger amount of funds upon which to earn an adequate return to pay out benefits more comfortably. Of course, there is a trade-off, in that premiums are in the end determined by the interaction of the risk preferences of both the insurer and the policyholder.

From a capital markets perspective, insurance securities should be priced exactly as any other financial security. In a frictionless competitive market, prices should satisfy the principle of no-arbitrage, meaning that no sure profit can be realized at zero cost (see [15] for the general theory and detailed references). If the market is complete, any random payoff can be replicated by suitably trading in the available assets, and the no-arbitrage price of a security is uniquely determined by the cost of setting up the strategy that hedges the security’s payoffs. When markets are incomplete, prices may not uniquely determined anymore. Indeed, there may exist an entire interval of prices consistent with no-arbitrage and the exact price chosen depends on the risk preferences of market participants.

Life insurance markets are an example of incomplete markets for the following reasons: they involve

exposures to risk factors (mortality) that are typically not spanned by financial securities; they are sometimes very long term, so that some financial risks (e.g., long-term interest rates) are not fully spanned by traded securities; finally, they are subject to trading constraints, in that insureds only take long positions on insurance contracts and there is no liquid secondary market for insurance contracts.

From a technical point of view, no-arbitrage ensures essentially the existence of a probability measure $\mathbb{P}^*$ *equivalent* (i.e., agreeing on the sets of null probability) to $\mathbb{P}$ under which security prices grow on average at the risk-free rate. Stated another way, security prices can be computed by employing the discounted cashflow approach, with discount rate r and probability $\mathbb{P}^*$. Market completeness translates into uniqueness of $\mathbb{P}^*$, while incomplete markets entail the existence of infinitely many such equivalent probabilities, each of them reflecting different risk preferences. This perspective can be fruitfully employed in the context of insurance pricing, which we have already discussed, and of market consistent valuation of insurance liabilities. The latter is required by some accounting standards and is based on the idea that the fair value of a portfolio of insurance liabilities should be equal to the hypothetical price that another insurer would pay to take them over if a liquid secondary market existed. Under the setup introduced in the section “Random Hazard Rates”, all probability measures equivalent to $\mathbb{P}$ can be parameterized by a couple of processes (η, ϕ) (with ϕ strictly positive) playing different roles in reflecting an agent’s risk aversion (see [7]). As a result, we can recast insurance pricing and market-consistent valuation in a unified framework by setting $\tilde{\mathbb{P}} \doteq \mathbb{P}^{(\tilde{\eta}, \tilde{\phi})}$ and $\mathbb{P}^* \doteq \mathbb{P}^{(\eta^*, \phi^*)}$. The two probabilities may coincide in special cases: we will now focus on $\tilde{\mathbb{P}}$ for ease of notation.

The processes $\tilde{\eta}$ and $\tilde{\phi}$ arise in the following way: $\tilde{\eta}$ contributes to the change of probability measure by entering the dynamics of the systematic risk sources (for example, interest rates and intensity of mortality); $\tilde{\phi}$ enters the change of measure through the survival indicators $1_{\{\tau^1 > t\}}, \dots, 1_{\{\tau^n > t\}}$ of the policyholders in the portfolio. In terms of risk adjustments, $\tilde{\eta}$ introduces a conservative adjustment in the random evolution of the systematic sources of risk, while $\tilde{\phi}$ may induce a change in the intensity of mortality itself, in the sense that the intensity of mortality

under $\tilde{\mathbb{P}}$ is given by $\tilde{\mu} = \tilde{\phi}\mu$ (see [7] for details). The meaning of the risk adjustments is then the following: $\tilde{\eta}$ adjusts for systematic sources of risk, and $\tilde{\phi}$ for the unsystematic risk of random fluctuations in the jumps of $1_{\{\tau^1 > t\}}, \dots, 1_{\{\tau^n > t\}}$, once the systematic risk factors have been taken into account (see discussion in the section titled “Systematic and Unsystematic Mortality Risk” and expressions (6, 7, 9)). For all this to make sense, the crucial assumption of conditional independence of $\tau^1, \dots, \tau^n$ must be satisfied. As a consequence, we may see $\tilde{\phi}$ as accounting for the risk that the systematic risk factors may not encompass all sources of risk commonly affecting the random lifetimes $\tau^1, \dots, \tau^n$.

As a concrete example, let us consider the case of a single premium term assurance providing the benefit as in equation (22). The single premium, Π , can be expressed as

$$\Pi = \tilde{E} \left(e^{-\int_0^T r_s ds} C 1_{\{\tau \leq T\}} \right) \quad (32)$$

for suitable $\tilde{\mathbb{P}}$. In the random hazard rate setup of the previous section and under some technical conditions, the previous expression can be rewritten as

$$\begin{aligned} \Pi &= \tilde{E} \left(\int_0^T e^{-\int_0^t (r_s + \tilde{\mu}_s) ds} C \tilde{\mu}_t dt \right) \\ &= \tilde{E} \left(\int_0^T e^{-\int_0^t (r_s + \tilde{\phi}_s \mu_s) ds} C \tilde{\phi}_s \mu_t dt \right) \end{aligned} \quad (33)$$

which shows the way in which risk aversion is encapsulated in the pricing mechanism: by taking expectation under a probability measure different from $\mathbb{P}$ and by adjusting the intensity of mortality (unless $\phi^* \equiv 1$). From equation (33), we see that $\tilde{\phi}$ adjusts the sum assured C as well: it can thus capture aversion to the risks described in the section titled “Risk Associated with Size of Benefits”.

What we have described so far is a coherent pricing structure, but the key point is to specify $(\tilde{\eta}, \tilde{\phi})$. A common approach is to fix some optimality criterion (e.g., expected utility maximization) and determine the risk premiums associated with the trading strategies solving the optimization problem. This can be done fairly explicitly in a few cases. For example, in [10, 16] liabilities are priced by identifying the initial price of the so-called risk-minimizing strategies. These are trading strategies that allow to perfectly replicate a liability payoff, while minimizing

the fluctuations in costs to be incurred over time to sustain the strategy. They typically embed risk aversion to systematic risk ($\tilde{\eta} \neq 0$), but risk neutrality to other risks ($\tilde{\phi} \equiv 1$). In practice, insurers have a range of requirements to fulfill when optimally designing and pricing complex contracts. Simulation and iterative optimization procedures are therefore utilized to come up with adequate risk margins.

When more emphasis is laid on market-consistent accounting of liabilities, it is natural to investigate how the above framework translates the risk preferences inherent in prices and assumptions used in the insurance market. Some results are provided in [7, 17], where the authors calibrate $(\tilde{\eta}, \tilde{\phi})$ and recover the risk premiums implied by different annuity markets.

To conclude, we mention two additional recent contributions. In [18], liabilities are priced by adopting the principle of equivalent utility, i.e., by setting a price that leaves the insurer indifferent between accepting or refusing a liability according to a given utility function. In [19], liabilities are priced by setting a target instantaneous Sharpe ratio. In both cases, under suitable assumptions nonzero risk premiums for systematic and unsystematic risk can be identified.

References

- [1] Williams, D. (2001). *Weighing the Odds. A Course in Probability and Statistics*, Cambridge University Press.
- [2] Daykin, C.D., Pentikäinen, T. & Pesonen, M. (1994). *Practical Risk Theory for Actuaries*, Chapman & Hall.
- [3] Cairns, A.J.G. (2000). A discussion of parameter and model uncertainty in insurance, *Insurance: Mathematics & Economics* **27**(3), 313–330.
- [4] Gerber, H.U. (1991). *Life Insurance Mathematics*, Springer.
- [5] Kalashnikov, V. & Norberg, R. (2003). On the sensitivity of premiums and reserves to changes in valuation elements, *Scandinavian Actuarial Journal* **2003**(3), 238–256.
- [6] Biffis, E. (2005). Affine processes for dynamic mortality and actuarial valuations, *Insurance: Mathematics & Economics* **37**(3), 443–468.
- [7] Biffis, E., Denuit, M. & Devolder, P. (2005). *Stochastic Mortality Under Measure Changes*, Pensions Institute, London.
- [8] Pitacco, E. (2004). Survival models in a dynamic context: a survey, *Insurance: Mathematics & Economics* **35**(2), 279–298.
- [9] Milevsky, M. & Promislow, D. (2001). Mortality derivatives and the option to annuitize, *Insurance: Mathematics & Economics* **29**(3), 299–318.
- [10] Dahl, M. & Møller, T. (2006). Valuation and hedging of life insurance liabilities with systematic mortality risk, *Insurance: Mathematics & Economics* **39**(2), 193–217.
- [11] Cairns, A.J.G., Blake, D. & Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration, *Journal of Risk and Insurance* **73**(4), 687–718.
- [12] Biffis, E. & Millossovich, P. (2006). A bidimensional approach to mortality risk, *Decisions in Economics and Finance* **29**(2), 71–94.
- [13] Norberg, R. (1991). Reserves in life and pension insurance, *Scandinavian Actuarial Journal* 3–24.
- [14] Norberg, R. (2001). On bonus and bonus prognoses in life insurance, *Scandinavian Actuarial Journal* **2001**(2), 126–147.
- [15] Duffie, D. (2001). *Dynamic Asset Pricing*, 3rd Edition, Princeton University Press.
- [16] Møller, T. (1998). Risk-minimizing hedging strategies for unit-linked life insurance contracts, *ASTIN Bulletin* **28**(1), 17–47.
- [17] Biffis, E. & Denuit, M. (2006). Lee-Carter goes risk-neutral, *Giornale dell'Istituto Italiano degli Attuari* **LXIX**, pp. 33–53.
- [18] Young, V.R. & Zariphopoulou, T. (2002). Pricing dynamic insurance risks using the principle of equivalent utility, *Scandinavian Actuarial Journal* **2002**(4), 246–279.
- [19] Milevsky, M.A., Promislow, S.D. & Young, V.R. (2006). Killing the law of large numbers: mortality risk premiums and the Sharpe Ratio, *Journal of Risk and Insurance* **73**(4), 673–686.

Related Articles

Credit Risk Models

Insurance Pricing/Nonlife

ENRICO BIFFIS

Privacy Protection in an Era of Data Mining and Record Keeping

Even though it is not guaranteed explicitly by the US constitution, the “right to privacy” is seen by many individuals as fundamental. There is widespread, and in some cases well-founded, fear of misuse of personal information (about one’s activities, associates, economic status, and health, for instance) by governments, organizations – for example, current or potential employers, and other individuals. Multiple factors currently exacerbate these fears: the legitimate need of governments for information, the near-complete reliance on electronic technologies to assemble, transmit, and store information, and the scope of the information collected: where one goes, what one eats, what one reads, with whom one associates and the footprint of one’s house all end up in someone’s database, not always accurately. Use of information for purposes other than that for which it was obtained has become rampant.

In this article, we provide a framework for thinking about privacy protection in an era of data mining and record keeping, and discuss approaches to the problem. While there are also important issues of organizational privacy – the protection of not simply proprietary information and trade secrets, but also economic information and information about employees, the focus is on personal privacy. It is also focused on the United States, which often lags behind the European Community in terms of protecting privacy.

Not uniquely, but possibly most visibly, privacy protection is where competing risks collide.

Background

By any standard, privacy (*see* **Syndromic Surveillance; Public Health Surveillance**) is an illusive concept. Physical privacy – protection of one’s person and home from “unreasonable searches and seizures” is articulated in Amendment IV to the US Constitution and well established legally. Breaches are for the most part readily detectable. Recourse is possible. None of these is true for information privacy. Few people know or have any meaningful

capability to detect or monitor who or which organizations hold what information about them, whether that information is correct, or for what purposes it is used, which may be entirely different from those for which it was collected. The threats are so diverse that new terms, such as *identity theft* have been coined to describe them.

The US Congress has reacted to some perceived and real threats to information privacy. Notable among congressional actions are the Fair Credit Reporting Act (1971)^a and the Health Insurance Portability and Accountability Act (HIPAA), enacted in 1996. Social security numbers cannot be used as identifiers, and have been removed from driver’s licenses and many databases.

Despite these actions, the public’s sense is that threats to information privacy are increasing rather than diminishing. Near-daily reports of intrusions into corporate databases containing credit card or other information add to the concerns.

What are the Risks and Impacts?

For an individual, there are seemingly evident risks associated with sensitive information being “in the wrong hands.” These risks are widely perceived as increasing. The most dramatic change over the past 5 years is not collection of the information, but rather its ready *accessibility* on the World Wide Web. More than 80% of the people in the United States are identified uniquely by their date of birth, gender, and five-digit ZIP code. Finding this information about someone on the web takes only a few minutes, using public sources such as voter registration and property ownership databases, and commercial services.

How real are the risks, and what are the consequences? Some risks, for instance, identity theft and being denied employment on the basis of medical conditions, are both real and carry severe, material consequences. Other risks are more nebulous. For instance, what is the risk associated with a third person knowing one’s salary, how much one’s house is worth, or what one has recently purchased online? And what are the consequences, other than the person’s knowing something one would prefer him or her not to know, if that information is acted on in a way that harms one?

One legal response is to remove consideration of harm. For some time, information collected by US “official statistics” agencies such as the Census

Bureau has been protected by law. Disclosure of identifiable information by employees is a felony. The disclosure is a crime in itself. The veracity of the information, which would protect disclosers under libel law, is also immaterial.

Although less widely recognized, there are also significant risks associated with incorrect information about individuals. Visible instances include people who have been unable to board airplanes because of the similarity of their names to those on “do not fly” lists and mistakes in medical treatment. Invisible instances are surely more numerous, and largely uncharacterized. Arguably, many people are at more risk from low-quality information (*see Managing Infrastructure Reliability, Safety, and Security; Human Reliability Assessment; Meta-Analysis in Nonclinical Risk Assessment*) about them than from correct information about them.

A final, especially invidious risk associated with information privacy is not knowing who has had access to what information. Moreover, information is not a consumable, and can be accessed an unlimited number of times.

From a social perspective, things are more complicated. The defining events for much of the current discussion about privacy are the terrorist attacks of September 11, 2001. The immense amount of information assembled *ex post facto* about the attackers and their internal and external interactions poses central questions: (a) Could this information have been used to prevent the attacks? (b) At what cost, not only financial, but also in terms of individual rights and liberties would this have been achieved? (c) What would be the recourse for those falsely suspected or accused? In the United States at least, the fact that the federal government has not always been (seen as) a responsible protector of information compounds the issues.

The impacts of the government’s knowing too little, from either a national security or policy perspective (What if the FDA had insufficient information to detect harmful drugs?) are obviously real. When the government has been at risk of having too little information, the Congress has often acted: examples include the USA Patriot Act [1] (which is controversial with respect to existence and implementation) and mandatory reporting of drug side effects. Less attention has been paid to information quality issues. One exception is the Fair Credit Reporting

Act, which provides individuals the ability to access and correct their credit records.

Individuals hold disparate views of the risks and impacts of the government’s knowing too much about its citizens. A significant complication, as discussed previously, is that “having access to information” and “causing harm from information” are quite different. Extreme liberals and conservatives argue, for very different reasons, that possession of information by the government is inherently bad. The prevailing view, however, seems to be more pragmatic – that governments can be trusted not to misuse information.

Governments are not the only organizations relevant to privacy. Organizations from e-tailors to insurance companies to universities assemble, hold and in some cases share or sell information about their customers or employees. Paradoxes abound. Students’ records have been protected since the Family Education Rights and Privacy Act (FERPA) was passed in 1974, while medical records received significant protection much more recently (HIPAA, 1996)^b. Customers of grocery and other stores readily surrender information about their buying habits in return for discounts. Web site “privacy policies” vary wildly. Much information about individuals is provided to the government by their employers. So, no one can answer confidently the question “Who knows what (true or false) about me?”

Nor are the mega problems the only ones. Researchers generate (for example, in medical and sociological studies) data containing sensitive information. At the same time, there are pressures to share data (e.g., for scientific verification), especially if public funds were used to generate them. Official statistics agencies wish to make their data available to researchers and policy makers: their data warehouses are not meant to be data cemeteries.

The Trade-offs

As noted in the section titled “Background”, privacy and information is a context in which risks inevitably compete, at least with respect to governments. There are conflicting, possibly irreconcilable, pressures. Beyond the use of census data for congressional apportionment – the census is prescribed by the US Constitution, the right of the government to collect information for use in setting policy and enforcing statutes is well established. The same is true of

information collected to preserve national security. In many settings, there are prohibitions against use of such information for other purposes.

There is, however, one significant difference between national security and official statistics: information provided to official statistics agencies is almost always provided knowingly, while information collected for national security may be obtained in a clandestine manner.

In part, the trade-offs are classical ones between individual risk and social benefit: more knowledge on the part of (a responsible) government engenders privacy risks to individuals, but leads to societal benefit. As might be expected, society's position on these trade-offs is ambivalent. To a degree that some consider remarkable, the US citizenry has been willing to surrender personal information in the name of national security. Nearly everyone responds to the census, and in the overwhelming majority of cases, responds accurately. At the same time, there is profound discomfort with the potential for abuse.

Trade-offs between individual privacy and societal benefit are not new. Those who pay taxes or serve in the armed forces relinquish their privacy. So do those who enroll in public schools, or receive economic assistance from the government. What distinguishes information privacy is that it can be given up involuntarily, unwittingly, and infinitely often, and that the information may be incorrect.

Proposed Solutions

There is no shortage of proposals for strengthening privacy protection in today's world. At a very high level, these can be grouped into *behavioral solutions*, *policy-driven solutions*, and *technological solutions*. Naturally, the lines between them are blurred and any effective long-term strategy must involve all three.

Behavioral solutions seek to make individuals aware of privacy issues, so that even in the absence of institutional safeguards, they can take as many steps as possible. Examples include verifying the legitimacy of those conducting government surveys, protecting credit cards from prying eyes, and not writing down computer passwords. But despite attention to privacy in the media and the ease of being (at least somewhat) safe, naïveté and carelessness remain rampant.

Policy-driven solutions are government-directed means by which individuals can protect their privacy.

Examples include being able to access and amend one's own records. A stronger version would be notification of who had accessed one's records, when, and for what purpose. Many organizations require computer users to use "strong" passwords and to change them frequently. Some policy-driven solutions go further, forbidding acts that an individual might not be aware of; FERPA and HIPPA are examples.

Technological solutions are meant to counter threats electronically. Encryption protects data from authorized access. Techniques for authenticating computer users are increasingly powerful. Water-marking may be able to trace improperly shared information. Two active and closely related fields – *privacy-preserving data mining* in [2] computer science and *statistical disclosure limitation* (SDL) (see [3, 4] **Near-Miss Management: A Participative Approach to Improving System Reliability**) in statistics – deal with allowing access to data for legitimate purposes without breaking confidentiality. Techniques are being created for sound statistical analysis of distributed databases. Prototype software has been developed that anonymizes Internet transactions.

Some Final Thoughts

As of 2007, the prospects for information privacy seem distinctly mixed. On the plus side, attention to the problem, from individuals, Congress, and researchers, is high. The US National Academies issued a number of reports. There seems to be wide understanding that compromises are necessary.

Even so, it is hard not to despair. Data volume is growing exponentially. Data quality is suspect, although to many, still insufficiently so. What is considered to be data is changing. Privacy issues are difficult enough for "classical" data: subject-indexed records containing text and numbers stored on machines, but what about video images on streets (London has more than 1 million street cameras) or at public events? [5–7] Voice prints? DNA profiles? Satellite images, such as those on Google Earth?

As always, humans are at once the weakest links in the chain and the only path to dealing with the problems. Truly draconian solutions, such as radio frequency IDs implanted in children or criminals who have paid their debt to society, are possible. But so

are sensible combinations of behavior, policy, and technology that offer genuine protection.

End Notes

^a. The Fair Credit Reporting Act is Public Law 91-508, codified at 15 USC §1681 et seq.

^b. HIPAA is Public Law No. 104-191, codified at 42 USC §201 et seq.

References

- [1] USA Patriot Act: Public Law 107-56.
- [2] Vaidya, J., Clifton, C. & Zhu M. (2006). *Privacy Preserving Data Mining*, Springer-Verlag, New York.
- [3] Doyle, P., Lane, J.I., Theeuwes, J.J.M. & Zayatz, L.V. (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies*, Elsevier, Amsterdam.

- [4] Willenborg, L.C.R.J. & de Waal, T. (2001). *Elements of Statistical Disclosure Control*, Springer-Verlag, New York.
- [5] Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure (1997). *For the Record: Protecting Electronic Health Information*, National Academies Press, Washington, D.C.
- [6] Duncan, G.T., Jabine, T.B. & de Wolf, V.A. (eds) (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, National Academies Press, Washington, D.C.
- [7] Waldo, J., Lin, H.S. & Millett, L.I. (eds) (2007). *Engaging Privacy and Information Technology in a Digital Age*, National Academies Press, Washington, D.C.

ALAN F. KARR

Probabilistic Design

The development of probabilistic methods in engineering is of real interest for design optimization or optimization of the inspection and maintenance strategy of structures subject to reliability and availability constraints, as well as for the requalification of a structure following an incident. One of the main stumbling blocks to the development of probabilistic methods is substantiation of probabilistic models used in the studies. In fact, it is frequently necessary to estimate an extreme value (*see Solvency; Extreme Value Theory in Finance; Mathematics of Risk and Reliability: A Select History*) based on a very small sample of existing data.

Whether a deterministic or probabilistic approach is implemented, sample or database treatment must be performed. A deterministic approach involves the identification of information like minimum, maximum values, and envelope curves, while a probabilistic vision concentrates on the dispersion or variability of the value through a variation interval or fractile-type data (without prejudice to a distribution as in some deterministic or parametric analyses), or a probability distribution. A fractile or quantile of the order α is a real number, X^* , satisfying $P(X \geq X^*) = \alpha$. Treatment is compatible with the intended application, for example, determining a good distribution representation around a central value or correctly modeling behavior in a distribution tail.

Tools to describe sample dispersions are taken from the statistics; however, their effectiveness is a function of the sample size. Methods are available that may be used to adjust a probability distribution on a sample, and then verify the adequacy of this adjusted distribution in the maximum failure probability region. It is obvious that if data is lacking or scarce, these tools are difficult to use. Under such circumstances, it is entirely reasonable to refer to expert opinion to model uncertainty associated with a value, and then transcribe this information in the form of a probability distribution. This article does not describe methods available in these circumstances, and the reader is referred, for example, to the maximum entropy principle (*see Volatility Smile; Bayesian Statistics in Quantitative Risk Assessment*) [1], and other references about expert opinion and application of Bayesian methods.

The practical approach of a probabilistic analysis may be summarized by three scenarios:

Scenario 1: If a lot of experience feedback data is available, the frequentist statistic is generally used. The objectivist or frequentist interpretation associates the probability with the observed frequency of an event. In this interpretation, the confidence interval of a parameter, p , has the property that the actual value of p is within the interval with a confidence level; this confidence interval is calculated on the basis of measurements.

Scenario 2: If data is not as abundant, expert opinion (*see Expert Elicitation for Risk Assessment; Expert Judgment; Sampling and Inspection for Monitoring Threats to Homeland Security*) may be used to obtain modeling hypotheses. The Bayesian analysis is used to correct *a priori* values established on the basis of expert opinion as a function of observed events. The subjectivist (or Bayesian) interpretation interprets probability as a degree of belief in a hypothesis. In this interpretation, the confidence interval is based on a probability distribution representing the analyst's degree of confidence in the possible values of the parameter and reflecting his/her knowledge of the parameter.

Scenario 3: If no data is available, probabilistic methods may be used that are designed to reason on the basis of a model that allows the value sought to be obtained from other values (referred to as the *input parameters*). The data to be gathered thus concerns the input parameters. The quality of the probabilistic analysis is a function of the credibility of statistics concerning these input parameters and that of the model. The following may be discerned:

- a structural reliability-type approach if the value sought is a probability,
- an uncertainty propagation-type approach if a statistic around the most probable value is considered.

Scenario 1, where a large enough sample is available, i.e., the sample allows “characterization of the relevant distribution with a known and adequate precision”, begs the following questions:

1. Is the distribution type selected relevant and justifiable? Of the various statistical models available, what would be the optimal choice of distribution?

2. Would altering the distribution (all other things being equal) entail a significant difference in the results of the application?
3. How can uncertainty associated with sample representativeness be taken into consideration (sample size, quality, etc.)?

Justification is difficult for scenarios 2 and 3. For example, if the parameters of a density are adjusted as of the first moments, it must be borne in mind that a precise estimation of the symmetry coefficient requires at least 50 values while kurtosis requires 100 data points, except for very specific circumstances. Furthermore, the critical values used by tests to reject or accept a hypothesis are frequently taken from results that are asymptomatic in the sense that the sample size tends toward infinity. Thus when the sample size is small, the results of conventional tests should be handled with caution.

With respect to question 2, a study examining sensitivity to the probability distribution used provides information. There are two methods available for the sensitivity study:

- It is assumed that the distribution changes while the first moments are preserved (mean, standard deviation especially) (moment identification-type method).
- The sample is redistributed to establish the reliability data distribution parameters (through a frequential or Bayesian approach).

Another method is to take the uncertainty associated with some distribution parameters into consideration by replacing the parameters' deterministic value with a random variable. Conventional criticism concerning the statistical modeling of a database concerns are

- difficulty in interpreting experience feedback for a specific application;
- database quality, especially if few points are available;
- substantiation of the probabilistic model built.

The probabilistic modeling procedure should attempt to answer these questions. If it is not possible to define a correct probabilistic model, it is obvious that, under these circumstances, the quantitative results in absolute value are senseless in the decision

process. However, the probabilistic approach always allows results to be used relatively, notably through

- a comparison of the efficiency of various solutions from the standpoint of reliability or availability, for example;
- classification of parameters that make the biggest contribution to the uncertainty associated with the response to direct R&D work to reduce said uncertainty.

This argument concerning the quality of uncertainty probabilistic models also has repercussions on the deterministic approach.

The deterministic approach involves validation of values used and also constitutes a sophisticated problem: it is not easy to prove that a value assumed to be conservative is realistic (especially if the sample is small), or to guarantee that the value is an absolute upper- or lowerbound. Conservative values used are frequently formally associated with small or large fractiles of the orders of the values studied. However, the concept of fractile is associated with the probability distribution adjusted on the sample, and even with one of the distribution tails.

The probabilistic approach seems to be even more suitable to deal with the problem. In fact, the probabilistic model reflects the level of knowledge of variables and models, and confidence in said knowledge. By means of sensitivity studies, this approach allows the impact of the probabilistic model choice on risk to be objectively assessed. Furthermore, in the event of new information impugning probabilistic modeling, and consequently the fractiles of a variable, the Bayesian theory, that combines objective and subjective (expertise) data, allows the probabilistic model and results of the probabilistic approach to be updated stringently.

The adjustment of a probability distribution and subsequent testing of the quality of the said adjustment around the central section (or maximum failure probability region) of the distribution constitute operations that are relatively simple to implement using the available statistical software packages (SAS, SPSS, Splus, Statistica, etc.), for conventional laws, in any case. However, the interpretation and verification of results still requires the expertise of a statistician. For example, the following points

- the results of an adjustment based on a histogram is sensitive to the intervals width;

- the maximum likelihood or moment methods are not suitable for modeling a sample obtained by overlaying phenomena beyond a given limit of an observation variable;
- moment methods assume estimations of kurtosis and symmetry coefficients that are only usually specified for large bases (with at least one hundred values for kurtosis);
- most statistical tests, specifically, the most frequently used Kolmogorov–Smirnov, Anderson–Darling, and Cramer–Von Mises tests, are asymptotic tests;
- in the Bayesian approach, the distribution selected *a priori* influences the result. Furthermore, the debate concerning whether the least informative law should be used has not been concluded.

Probabilistic versus Deterministic Approach of the Design

The basis of the deterministic approach is the so-called design values for the loads and the strength parameters. Loads, for instance, are the design water level and the design-significant wave height. Using design rules according to codes and standards, it is possible to determine the shape and the height of the cross section of the flood defense.

These design rules are based on limit states of the flood defense system's elements, such as overtopping, erosion, instability, piping, and settlement.

It is assumed that the structure is safe when the margin between the design value of the load and the characteristic value of the strength is large enough for all limiting states of all elements.

The safety level of the protected area is not explicitly known when the flood defense is designed according to the deterministic approach.

The most important shortcomings of the deterministic approach are as follows:

- It is a fact that the failure probability of the system is unknown.
- The complete system is not considered as an integrated entity. An example is the design of the flood defenses of the protected area of Figure 1. With the deterministic approach, the design of the sea dike is, in both cases, exactly the same. In reality, the left area is threatened by flood from two independent causes – the sea and the river.

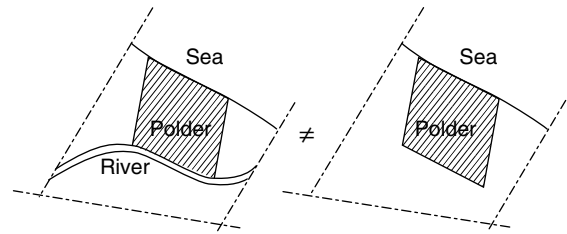


Figure 1 Different safety levels for the same design

Therefore, the safety level of the left area is less than the safety level of the right one.

- Another shortcoming of the deterministic approach is that the length of flood defense does not affect the design.

In the deterministic approach, the design rules are the same for all the sections, independent of the number of sections. It is, however, intuitively clear that the probability of flooding increases with the length of the flood defense (Figure 2).

- With the deterministic design methods, it is impossible to compare the strength of different types of cross sections such as dikes, dunes, and structures like sluices and pumping stations.
- Last but not least, the deterministic design approach is incompatible with other policy fields, for instance, the safety of industrial processes and the safety of transport of dangerous substances.

A fundamental difference with the deterministic approach is that the probabilistic design methods are based on an acceptable frequency or probability of flooding of the protected area.

The probabilistic approach results in a probability of failure of the whole flood defense system taking account of each individual cross section and each structure. So the probabilistic approach is an integral design method for the whole system.

Uncertainties

Uncertainties are everywhere. They surround us in everyday life. Among the numerous synonyms for “uncertainty” are unsureness, unpredictability, randomness, hazardness, indeterminacy, ambiguity, variability, irregularity, and so on. Recognition of the need to introduce the ideas of uncertainty in civil

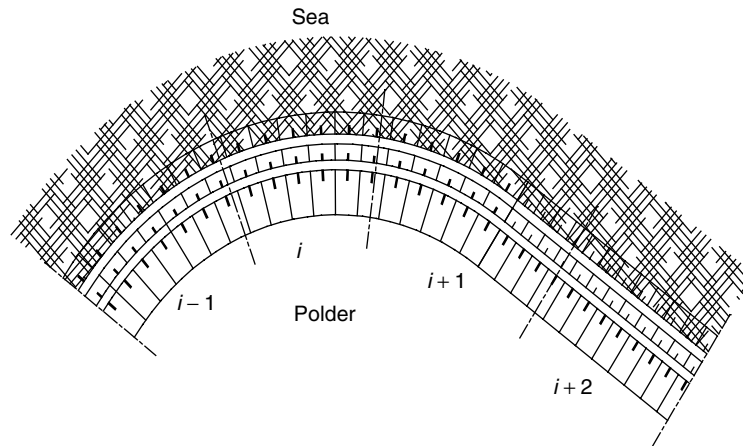


Figure 2 Sections of a dike

engineering today reflects, in part, some of the profound changes in civil engineering over the last few decades.

Recent advancements in statistical modeling have provided engineers with an increasing power for making decisions under uncertainty. The process and information involved in the engineering problem solving are, in many cases, approximate, imprecise, and subject to change. It is generally impossible to obtain sufficient statistical data for the problem at hand, and reliance must be placed on the ability of the engineer to synthesize existing information when required. Hence, to assist the engineer in making decisions, analytical tools should be developed to effectively use the existing uncertain information.

Uncertainties in decision and risk analysis can primarily be divided into two categories: uncertainties that stem from variability in known (or observable) populations and therefore represent randomness in samples (inherent uncertainty), and uncertainties that come from basic lack of knowledge of fundamental phenomena (epistemic uncertainty (*see Protection of Infrastructure*)).

Inherent uncertainties represent randomness or variations in nature. For example, even with a long history of data, one cannot predict the maximum water level that will occur in, for instance, the coming year, at the North Sea. It is not possible to reduce inherent uncertainties.

Epistemic uncertainties are caused by lack of knowledge of all the causes and effects in physical systems, or by lack of sufficient data. For example,

it might only be possible to obtain the type of a distribution, or an exact model of a physical system, when sufficient research can be and is done. Epistemic uncertainties may change as knowledge increases.

Generally, in probabilistic design, the following five types of uncertainties are discerned (see also [2]): inherent uncertainty in time and in space, parameter uncertainty and distribution type uncertainty (together also known as *statistical uncertainty*), and finally model uncertainty. Uncertainties such as construction costs uncertainties, damage costs uncertainties, and financial uncertainties are considered to be examples of model uncertainties.

Inherent Uncertainty in Time

When determining the probability distribution of a random variable that represents the variation in time of a process (like the occurrence of a water level), there essentially is a problem of information scarcity. Records are usually too short to ensure reliable estimates of low-exceedance probability quantiles in many practical problems. The uncertainty caused by this shortage of information is the statistical uncertainty of variations in time. This uncertainty can theoretically be reduced by keeping record of the process for the coming centuries.

Stochastic processes running in time (individual wave heights, significant wave heights, water levels, discharges, etc.) are examples of the class of inherent uncertainty in time. Unlimited data will not reduce

this uncertainty. The realizations of the process in the future remain uncertain. The probability density function (PDF) or the cumulative distribution function (CDF) and the autocorrelation function describe the process.

In case of a periodic stationary process like a wave field, the autocorrelation function will have a sinusoidal form and the spectrum, as the Fourier transform of the autocorrelation function, gives an adequate description of the process. Attention should be paid to the fact that the well-known wave energy spectra such as Pierson–Moskowitz and Jonswap are not always able to represent the wave field at a site. In quite a few practical cases, swell and wind wave form a wave field together. The presence of two energy sources may be clearly reflected in the double peaked form of the wave energy spectrum.

An attractive aspect of the spectral approach is that the inherent uncertainty can be easily transferred through linear systems by means of transfer functions. By means of the linear wave theory the incoming wave spectrum can be transformed into the spectrum of wave loads on a flood defense structure. The PDF of wave loads can be derived from this wave load spectrum. Of course, it is assumed here that no wave breaking takes place in the vicinity of the structure. In case of nonstationary processes that are governed by meteorological and atmospheric cycles (significant wave height, river discharges, etc.) the PDF and the autocorrelation function are needed. Here the autocorrelation function gives an impression of the persistence of the phenomenon. The persistence of rough and calm conditions is of utmost importance in workability and serviceability analyses.

If the interest is directed to the analysis of ultimate limit states (ULSs), e.g. sliding of the structure, the autocorrelation is eliminated by selecting only independent maxima for the statistical analysis. If this selection method does not guarantee a set of homogeneous and independent observations, physical or meteorological insights may be used to homogenize the dataset. For instance, if the fetch in the north westerly (NW) direction is clearly maximal, the dataset of maximum significant wave height could be limited to NW storms. If such insight fails, one could take only the observations exceeding a certain threshold peaks over threshold (POT) (see **Extreme Values in Reliability**) into account hoping that this will lead to the desired result. In case of a clear yearly seasonal

cycle, the statistical analysis can be limited to the yearly maxima.

Special attention should be given to the joint occurrence of significant wave height H_s and spectral peak period T_p . A general description of the joint PDF of H_s and T_p is not known. A practical solution for extreme conditions considers the significant wave height and the wave steepness s_p as independent stochastic variables to describe the dependence. This is a conservative approach as extreme wave heights are more easily realized than extreme peak periods. For the practical description of daily conditions (service limit state: SLS), the independence of s_p and T_p seems sometimes a better approximation. Also the dependence of water levels and significant wave height should be explored because the depth limitation to waves can be reduced by wind setup. Here the statistical analysis should be clearly supported by physical insight. Moreover, it should not be forgotten that shoals could be eroded or accreted due to changes in current or wave regime induced by the construction of the flood defense structure.

Inherent Uncertainty in Space

When determining the probability distribution of a random variable that represents the variation in space of a process (like the fluctuation in the height of a dike), there is essentially a problem of shortage of measurements. It is usually too expensive to measure the height or width of a dike in great detail. This statistical uncertainty of variations in space can be reduced by taking more measurements [2].

Soil properties can be described as stochastic processes in space. From a number of field tests, the PDF of the soil property and the (three-dimensional) autocorrelation function can be fixed for each homogeneous soil layer. Here, the theory is more developed than the practical knowledge that is available. Numerous mathematical expressions are proposed in the literature to describe the autocorrelation. No clear preference has, however, emerged yet as to which functions describe the fluctuation pattern of the soil properties best. Moreover, the correlation length (distance where correlation becomes approximately zero) seems to be of the order of 30–100 m, while the spacing of traditional soil mechanical investigations for flood defense structures is of the order of 500 m. So it seems that the intensity of the soil

mechanical investigations has to be increased considerably if reliable estimates have to be made of the autocorrelation function.

The acquisition of more data has a different effect in case of stochastic processes in space rather than in time. As structures are immobile, there is only one single realization of the field of soil properties. Therefore, the soil properties at the location could be exactly known if sufficient soil investigations are done. Consequently, the actual soil properties are fixed after construction, although not completely known to man. The uncertainty can be described by the distribution and the autocorrelation function, but it is, in fact, a case of lack of information.

Parameter Uncertainty

This uncertainty occurs when the parameters of a distribution are determined with a limited number of data. The fewer the number of data available, the higher is the parameter uncertainty. A parameter of a distribution function is estimated from the data and is thus a random variable. The parameter uncertainty can be described by the distribution function of the parameter. In van Gelder [3], an overview of the analytical and numerical derivation of parameter uncertainties for certain probability models (exponential, Gumbel, and lognormal) is given. The bootstrap method (*see Statistics for Environmental Toxicity; Credit Migration Matrices*) is a fairly easy tool used to calculate the parameter uncertainty numerically. Bootstrapping methods are described in, for example, Efron [4]. Given a dataset $x = (x_1, x_2, \dots, x_n)$, we can generate a bootstrap sample x^* , which is a random sample of size n drawn with replacement from the dataset x . The following bootstrap algorithm can be used for estimating the parameter uncertainty:

1. Select B independent bootstrap samples $x^{*1}, x^{*2}, \dots, x^{*B}$, each consisting of n data values drawn with replacement from x .
2. Evaluate the bootstrap corresponding to each bootstrap sample;

$$\tau^*(b) = f(x^{*b}) \text{ for } b = 1, 2, \dots, B \quad (1)$$
3. Determine the parameter uncertainty by the empirical distribution function of τ^* .

Other methods to model parameter uncertainties like Bayesian methods (*see Repair, Inspection, and*

Replacement Models; Near-Miss Management: A Participative Approach to Improving System Reliability) can be applied too. Bayesian inference lays its foundations upon the idea that states of nature can be and should be treated as random variables. Before making use of data collected at the site, the engineer can express his information concerning the set of uncertain parameters Λ for a particular model $f(x|\Lambda)$, which is a PDF for the random variable x . The information about Λ can be described by a prior distribution $B(\Lambda|I)$, i.e., prior to using the observed record of the random variable x . The basis upon which these prior distributions are obtained from the initial information I are described in, for instance, van Gelder [3]. Noninformative priors can be used if we do not have any prior information available. If $p(\lambda)$ is a noninformative prior, consistency demands that $p(H) dH = p(\lambda) d\lambda$ for $H = H(\lambda)$; thus a procedure for obtaining the ignorance prior should presumably be invariant under one-to-one reparametrization. A procedure that satisfies this invariance condition is given by the Fisher matrix of the probability model:

$$I(\lambda) = -E_{x|\lambda} [M^2/M^2 \log f(x|\lambda)] \quad (2)$$

giving the so-called noninformative Jeffrey's prior $p(\lambda) = I(\lambda)^{1/2}$.

The engineer now has a set observations x of the random variable X , which he assumes comes from the probability model $f_X(x|\lambda)$. Bayes' theorem provides a simple procedure by which the prior distribution of the parameter set Λ may be updated by the dataset X to provide the posterior distribution of Λ , namely,

$$f(\Lambda|X, I) = 1(X|\Lambda)B(\Lambda|I)/K \quad (3)$$

where

$f(\Lambda|X, I)$: posterior density function for Λ , conditional upon a set of data X and information I

$1(X|\Lambda)$: sample likelihood of the observations, given the parameters

$B(\Lambda|I)$: prior density function for Λ , conditional upon the initial information I

K : is the normalizing constant ($K = E_\lambda(X|\Lambda)B(\Lambda|I)$).

The posterior density function of Λ is a function weighted by the prior density function of Λ and the data-based likelihood function in such a manner as to combine the information content of both. If future observations X_F are available, Bayes' theorem can

be used to update the PDF on Λ . In this case, the former posterior density function for Λ now becomes the prior density function, since it is prior to the new observations or the utilization of new data. The new posterior density function would also have been obtained if the two samples X and X_F had been observed sequentially as one set of data. The way in which the engineer applies his information about Λ depends on the objectives in analyzing the data.

Distribution Type Uncertainty

This type represents the uncertainty of the distribution type of the variable. It is, for example, not clear whether the changes in the water level of the North Sea is exponentially or Gumbel distributed (*see Reliability of Large Systems; Extreme Values in Reliability*), or whether it has another distribution. A choice was made to divide statistical uncertainty into parameter- and distribution-type uncertainty although it is not always possible to draw the line; in case of unknown parameters (because of lack of observations), the distribution type will be uncertain as well.

Any approach that selects a single model and then makes inference conditionally on that model ignores the uncertainty involved in the model selection, which can play a big part in the overall uncertainty. This difficulty can be avoided, in principle, if one adopts a Bayesian approach and calculates the posterior probabilities of all the competing models following directly from the Bayes factors. A composite inference can then be made that takes account of model uncertainty in a simple way with the weighted average model:

$$f(h) = A_1 f_1(h) + A_2 f_2(h) + \cdots + A_n f_n(h) \quad (4)$$

where $\sum A_i = 1$.

The approach described above gives us some sort of Bayesian discrimination (*see Risk in Credit Granting and Lending Decisions: Credit Scoring*) procedure between competing models. This area has become very popular recently. Theoretical research comes from Kass and Raftery [5], and applications can be found mainly in the biometrical sciences [6] and econometrical sciences [7]. The very few applications of Bayesian discrimination procedures in civil engineering come from Wood and Rodriguez-Iturbe [8], Pericchi and Rodriguez-Iturbe [9, 10], and Perreault *et al.* [11].

Model Uncertainty

Many engineering models describing natural phenomena like wind and waves are imperfect. They can be imperfect because the physical phenomena are not known (for example, when regression models without the underlying physics are used), or they can be imperfect because some variables of lesser importance are omitted in the engineering model for reasons of efficiency.

Suppose that the true state of nature is X . Prediction of X may be modeled by X^* . As X^* is a model of the real world, imperfections may be expected; the resulting predictions will therefore contain errors and a correction N may be applied. Consequently, the true state of nature may be represented by Ang [12]:

$$X = NX^* \quad (5)$$

If the state of nature is random, the model X^* naturally is also a random variable, for which a normal distribution will be assumed. The inherent variability is described by the coefficient of variation (CV) of X^* , given by $\sigma(X^*)/\mu(X^*)$. The necessary correction N may also be considered a random variable, whose mean value $\mu(N)$ represents the mean correction for systematic error in the predicted mean value, whereas the CV of N , given by $\sigma(N)/\mu(N)$, represents the random error in the predicted mean value.

It is reasonable to assume that N and X^* are statistically independent. Therefore, we can write the mean value of X as

$$\mu(X) = \mu(N)\mu(X^*) \quad (6)$$

The total uncertainty in the prediction of X becomes

$$CV(X) = \sqrt{(CV^2(N) + CV^2(X^*) + CV^2(N)CV^2(X^*))} \quad (7)$$

In van Gelder [3], an example of model uncertainty is presented by fitting physical models to wave impact experiments.

We can ask ourselves if there is a relationship between model and parameter uncertainty. The answer is "No". Consider a model for predicting the weight of an individual as a function of his height. This might be a simple linear correlation of the form $W = aH + b$. The parameters a and b may be found from a least squares fit to some sample data. There

will be *parameter* uncertainty to a and b due to the sample being just that – a sample, not the whole population. Separately, there will be *model* uncertainty due to the scatter of individual weights either side of the correlation line.

Thus parameter uncertainty is a function of how well the parameters provide a fit to the population data, given that they would have been fitted using only a sample from that population, and that sample may or may not be wholly representative of the population. Model uncertainty is a measure of the scatter of individual points either side of the model once it has been fitted. Even if the fitting had been performed using the whole population, there would still be residual errors for each point since the model is unlikely to be exact.

Parameter uncertainty can be reduced by increasing the amount of data against which the model fit is performed. Model uncertainty can be reduced by adopting a more elaborate model (e.g., quadratic fit instead of linear). There is, however, no relationship between the two.

Uncertainties Related to the Construction

To optimize the design of a hydraulic structure, the total lifetime costs – an economic cost criterion – can be used. The input for the cost function consists of uncertain estimates of the construction cost and the uncertain cost in case of failure.

The construction costs consist of a part that is a function of the structure geometry (variable costs) and a part that can only be allocated to the project as a whole (fixed costs). For a vertical breakwater, for instance, the variable costs can be assumed to be proportional to the volumes of concrete and filling sand in the cross section of the breakwater.

The costs in case of ULS (see **Structural Reliability**) failure consist of replacement of (parts of) the structure and thus depend on the structure dimensions. The costs in case of SLS failure are determined by the costs of downtime and are thus independent of the structure geometry.

The total risk over the lifetime of the structure is given by the sum of all yearly risks, corrected for interest, inflation, and economical growth. This procedure is known as *capitalization*. The growth rate expresses that, in general, the value of all goods and equipment behind the hydraulic structure will increase during the lifetime of the structure.

Several cost components can be allocated to the building project as a whole. Examples of these cost components are

- cost of the feasibility study;
- cost of the design of the flood protection structure;
- site investigations, like penetration tests, borings and surveying;
- administration.

In principle, there are two ways in which a structure can fail. Either the structure collapses under survival conditions after which there will be more wave penetration in the protected area or the structure is too low and allows too much wave generation in the protected area due to overtopping waves. In both cases, possibly harbor operations have to be stopped, resulting in (uncertain amount of) damage (downtime costs).

If a structure fails to protect the area of interest against wave action, possibly the operations in this area will have to be stopped. The damage costs that are caused by this interruption of harbor operations are called *downtime costs*. The exact amount of downtime costs is very difficult to determine and therefore contain a lot of uncertainty. The downtime costs for one single ship can be found in literature, but the total damage in case of downtime does not depend solely on the downtime costs of ships. The size of the harbor and the type of cargo are also important variables in this type of damage. Furthermore, the availability of an alternative harbor is very important. If there is an alternative, ships will use this harbor. In that case, the damaged harbor will lose income because less ships make use of the harbor and possibly because of claims of shipping companies. On a macroeconomic scale, however, there is possibly minor damage since the goods are still coming in by way of the alternative harbor. This also shows that the availability of infrastructure in the area influences the damage in case of downtime. If an alternative harbor is not available, the economic damage may be felt beyond the port itself.

The location of the structure in relation to the harbor also influences the damage costs. If the structure protects the entrance channel, the harbor cannot be reached during severe storms, thus causing waiting times. These waiting times have the order of magnitude of hours to a few days. If the structure protects the harbor basin or a terminal, damage to the structure can cause considerable amounts of extra

downtime owing to the fact that the structure only partly fulfills its task over a longer period of time.

If the load on a structure component exceeds the admissible load, the component collapses. Several scenarios are now possible:

- The component is not essential to the functionality of the structure. Repair is not carried out and there is no damage in monetary terms. This is the case if, for instance, an armor block is displaced in the rubble foundation. It should be noted that this kind of damage can cause failure if a lot of armor blocks are displaced (preceding failure mode).
- The component is essential to the functionality of the structure. The stability of the structure is, however, not threatened. This is the case if, for instance, the crown wall of a caisson collapses. The result is a reduction of the crest height of the structure, which could threaten the functionality of the structure. Therefore, repair has to be carried out and there is some damage in monetary terms.
- The structure has become unstable during storm conditions. There is considerable damage to the structure, resulting in necessary replacement of (parts of) the structure. The damage in monetary terms is possibly even higher than the initial investment in the structure.

When optimizing a structural design, an estimate of the damage is needed. In the case of a structure component, this could be the cost of rebuilding. If large parts of the caissons have collapsed, the area would have to be cleared before rebuilding the structure. In that case, the damage would be higher than in the case of rebuilding alone. Furthermore, collapse, in general, leads to downtime cost which further increases the damage.

Reduction of Uncertainty

Inherent uncertainties represent randomness or variations in nature. Inherent uncertainties cannot be reduced. Epistemic uncertainties, on the other hand, are caused by lack of knowledge. Epistemic uncertainties may change as knowledge increases. In general, there are three ways to increase knowledge:

- gathering data
- research
- expert judgment.

Data can be gathered by taking measurements or by keeping record of a process in time. Research can, for instance, be undertaken with respect to the physical model of a phenomenon or into the better use of existing data. By using expert opinions, it is possible to acquire the probability distributions of variables that are too expensive or practically impossible to measure.

The goal of all this, obviously, is to reduce the uncertainty in the model. Nevertheless, it is also thinkable that uncertainty will increase. Research might show that an originally flawless model actually contains a lot of uncertainties. Or after taking some measurements, the variations of the dike height can be a lot larger. It is also thinkable that the average value of the variable will change because of the research that has been done.

The consequence is that the calculated probability of failure will be influenced by future research. To guarantee a stable and convincing flood defense policy after the transition, it is important to understand the extent of this effect.

Probabilistic Approach of the Design

The accepted probability of flooding is not the same for every polder or floodplain. It depends on the nature of the protected area, the expected loss in case of failure, and the safety standards of the country. For instance, for a protected area with a dense population or an important industrial development, a smaller probability of flooding is allowed than for an area of lesser importance.

For this reason accepted risk is a better measure than an accepted failure probability, because risk is a function of the probability and the consequences of flooding.

The most general definition of risk is the product of the probability and a power of consequences:

$$Risk = (probability) \cdot (consequence)^n \quad (8)$$

In many cases, such as economic analyses, the power n is equal to 1.

Figure 3 shows the elements of the probabilistic approach. First of all the flood defense system has to be described as a configuration of elements such as dike sections, sluices, and other structures. Then an inventory of all the possible hazards and failure modes must be made. This step is one of the most

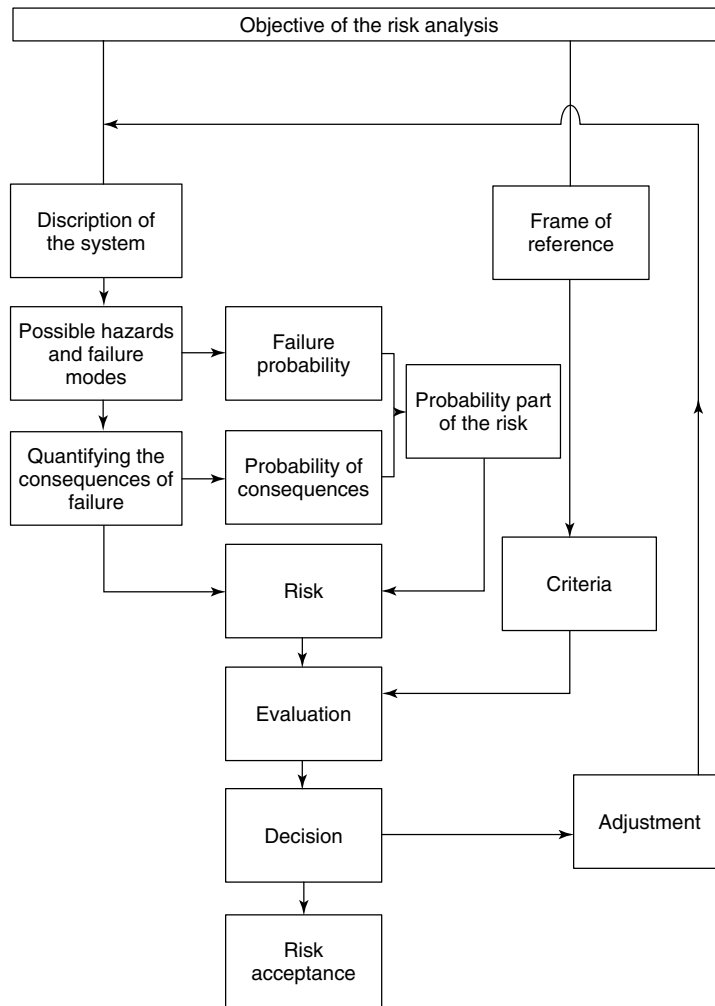


Figure 3 Probabilistic approach of the design

important of the analysis because missing a failure mode can seriously influence the safety of the design.

The next step can be the quantifying of the consequences of failure. Here, it is necessary to analyze the consequence of the failure for all possible ways of failure. Sometimes, the consequences of failure of an element of the system are different for each element.

The failure probability and the probability of the consequences form the probability part of the risk. When the risk is calculated, the design can be evaluated. For this, criteria must be available such as a maximum acceptable probability of a number of casualties or the demand of minimizing the total costs

including the risk. For determining the acceptable risk, we need a frame of reference. This frame of reference can be the national safety level aggregating all the activities in the country.

After the evaluation of the risk, one can decide to adjust the design or to accept it with the remaining risk.

System Analysis

Every risk analysis, which is the core of the probabilistic design, starts with a system analysis. There are several techniques to analyze a system, but in this

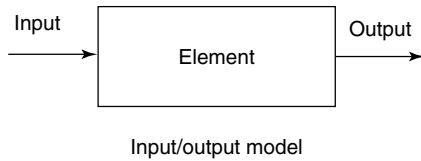


Figure 4 Input–output model

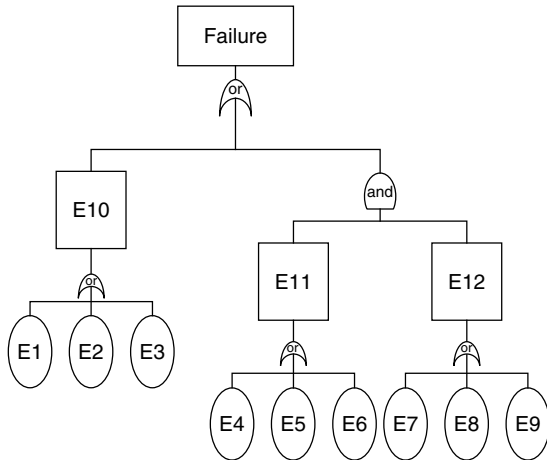


Figure 5 Fault tree

case we restrict ourselves to the input–output model (Figure 4) and the fault tree analysis.

With the input–output model the elements of the system are schematized as fuses in an electrical scheme. When an element fails, the connection is broken and there will be no current through the element. So in this case, there will be no output.

The fault tree arranges all the events in such a way that their occurrence leads to failure of the system. Figure 5 is an example of a fault tree. A fault tree (*see Canonical Modeling; Systems Reliability; Imprecise Reliability; Reliability Data*)

consists of basic events ($E1 \dots E9$), combined events ($E10 \dots E12$), a top event (failure), and gates (and, or). A gate is the relation of the events beneath the gate that lead to the event above it.

Simple Systems

The most simple systems are the parallel and the series systems (Figure 6). A parallel system that consists of two elements functions as long as one of the elements functions.

When a system fails if only one element fails, it is called a *series system*.

What can we say about the strength of these systems? Let us start with the series system. For instance, in the case of a chain which is loaded by a tensile force (Figure 7) the chain is as strong as the weakest link.

A parallel system that consists of two ductile steel columns in a frame is as strong as the sum of the strengths of the two columns (see Figure 8).

When the elements and their failure modes are analyzed, it is possible to make a fault tree. The fault tree gives the logical sequence of all the possible events that lead to failure of the system.

Take, for instance, the fault tree of a simple parallel system. The basic events are the failure of the single elements, and the failure of the system is called the *top event*. The system fails only when all single elements fail. So the gate between the basic events and the top event is a so-called and-gate (see Figure 9).

A series system of two elements fails if only one of the elements fails as depicted by the so-called or-gate between the basic events and the top event (see Figure 10).

When there are more failure modes possible for the failure of an element, then the failure modes are the basic events and the failure of an element is a so-called composite event.

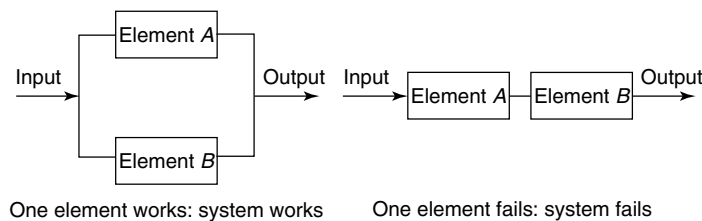


Figure 6 Parallel and series systems

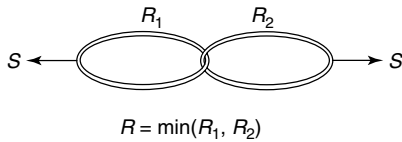


Figure 7 A chain as a series system

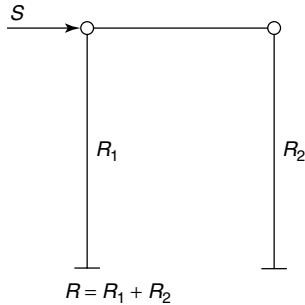


Figure 8 A frame as a parallel system

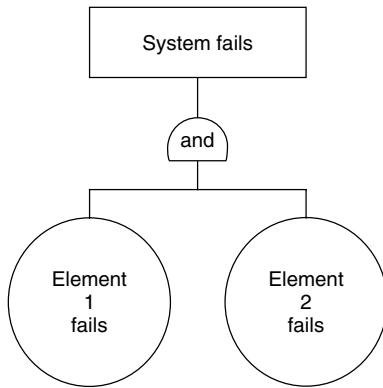


Figure 9 Fault tree of a parallel system

Figure 11 is an example (after [13]) of a parallel system of elements in which the system elements are in turn series systems of the possible failure modes.

Example: Risk Analysis of a polder

An overview of the flood defense system of a polder is given in Figure 12. Failure of the subsystems (dike, dune sluice, and levee) of the system leads to flooding of the polder.

The subsystems all consist of elements. The dikes can, for instance, be divided into sections. This is also shown in Figure 12. Failure of any of the elements

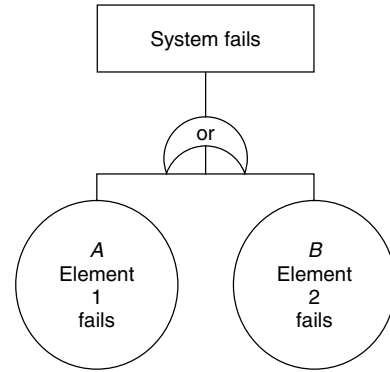


Figure 10 Fault tree of a series system

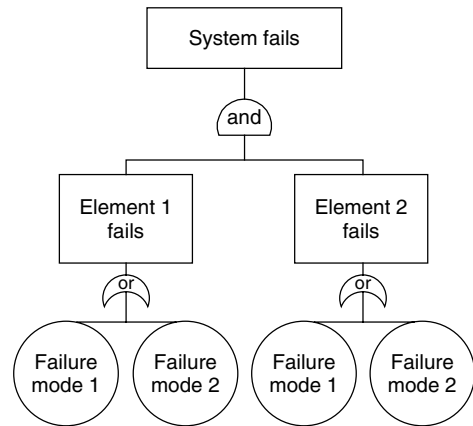


Figure 11 Elements of a parallel system as series systems of failure modes

of the subsystem “dike 1” leads to flooding of the polder.

For all the elements of the flood defense, all the possible failure modes can be the cause of failure. A failure mode is a mechanism that leads to failure. The most important failure modes for a dike section are given in Figure 13. The place of the failure modes in the system is demonstrated by a fault tree analysis in Figures 14 and 15.

An advantage of the probabilistic approach above the deterministic approach is illustrated in Figure 15, where human failure to close the sluice is included in the analysis.

The conclusion of this analysis is that any failure mechanism of any element of any subsystem of the flood defense system can lead to inundation of the polder. The system is, therefore, a series system.

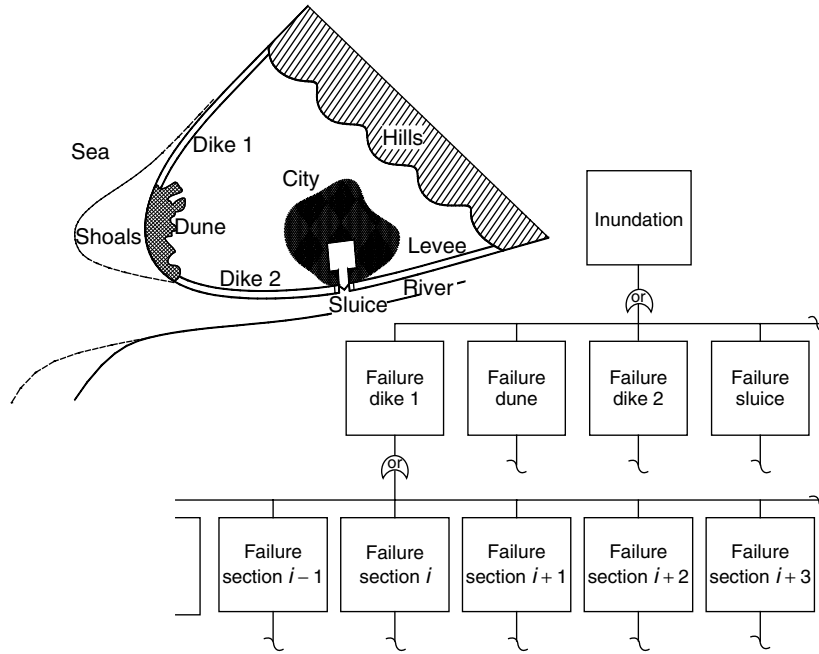


Figure 12 Flood defense system and its elements presented in a fault tree

Failure Probability of a System

This article gives an introduction of the determination of the failure probability of a system for which the failure probabilities of the elements are known.

Two fault trees given in Figure 16 – one for a parallel system and one for a series system – both consisting of two elements.

Event A is the event that element 1 fails and B is the event that element 2 fails.

The parallel system fails if both the elements fail. The failure probability is the probability of A and B . The series system fails if at least one of the elements fail. So the failure probability is the probability of A or B (see Figure 17).

The probability of A and B is equal to the product of the probability of A and the probability of B given A . The probability of A or B is equal to the sum of the probability of A and the probability of B minus the probability of A and B .

In practice, the evaluation of the probability of B given A is rather difficult because the relation between A and B is not always clear. If A and B are independent of each other, the probability of B given A is equal to the probability of B without A .

In this case, the probability of A and B is equal to the product of the probability of A and the probability of B :

$$P(B|A) = P(B) \implies P(A \cap B) = P(A)P(B) \quad (9)$$

If event A excludes B , then the probability B and A is zero:

$$P(B|A) = 0 \implies P(A \cap B) = 0 \quad (10)$$

If A includes B , then the probability of B given A is 1 and so the probability of A and B is equal to the probability of A

$$P(B|A) = 1 \implies P(A \cap B) = P(A) \quad (11)$$

In the same way, it is possible to determine the probability of A or B . If A and B are independent of each other the probability of A or B is

$$P(A \cup B) = P(A) + P(B) - P(A)P(B) \quad (12)$$

If event A excludes B , then the probability B and A is zero so the probability of A or B is

$$P(A \cup B) = P(A) + P(B) \quad (13)$$

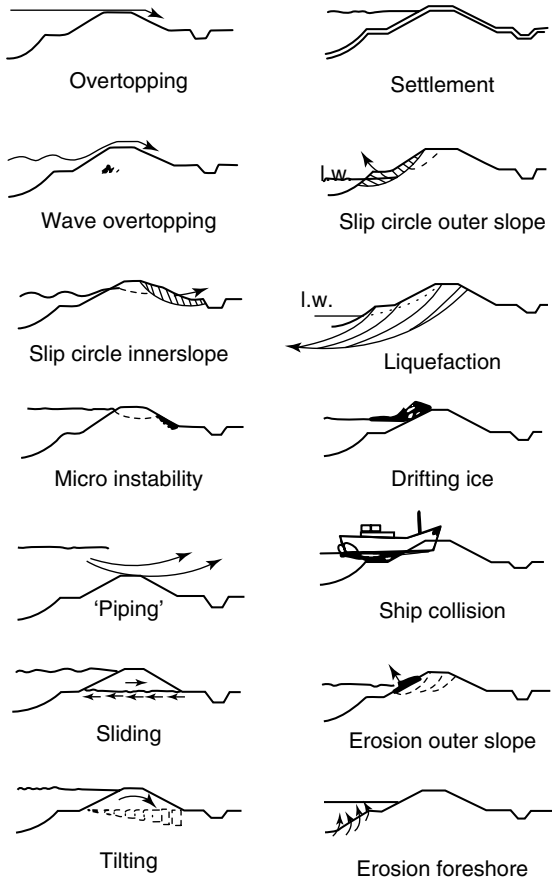


Figure 13 Failure modes of a dike [Reproduced from [13] with permission from Springer, © 1986.]

If A includes B , then the probability of B given A is 1 and so the probability of A and B is equal to the probability of A and the probability of A or B is

$$P(A \cup B) = P(A) + P(B) - P(A) \quad (14)$$

In many cases, the events A and B are each described by a stochastic variables Z_1 and Z_2 , respectively. Event A occurs when $Z_1 < 0$ and event B occurs when $Z_2 < 0$. This is further explained in the next section on probability estimation of an element. In many cases, as for instance, when there is a linear relation between Z_1 and Z_2 , the dependency of the events A and B can be described by the correlation coefficient. This is defined by

$$\rho = \frac{\text{Cov}(Z_1 Z_2)}{\sigma_{Z_1} \sigma_{Z_2}} \quad (15)$$

where

σ_{Z_1} : standard deviation of Z_1

σ_{Z_2} : standard deviation of Z_2

$\text{Cov}(Z_1 Z_2)$: covariance of Z_1 and Z_2

: $E((Z_1 - \mu_{Z_1})(Z_2 - \mu_{Z_2}))$

: expected value of $(Z_1 - \mu_{Z_1})(Z_2 - \mu_{Z_2})$

$f_{Z_1}(\xi_1)$: probability density function of Z_1

In the graph of Figure 18, the probability of A or B is plotted against the correlation coefficient. The probability of B is the lower limit of the failure probability and the sum of the probability of A and the probability of B is the upper limit of the failure probability.

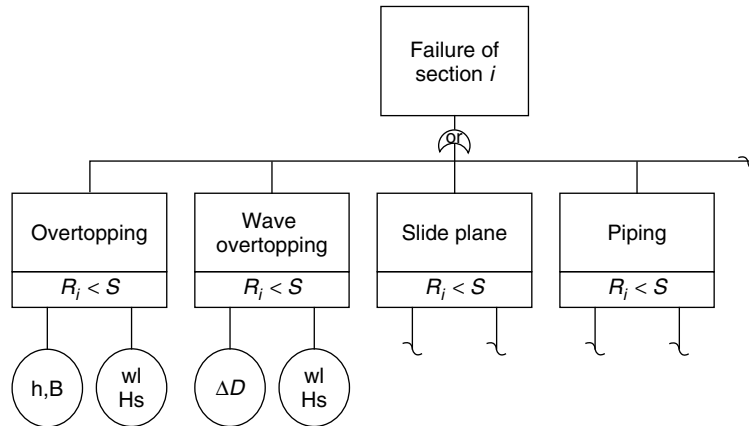


Figure 14 A dike section as a series system of failure modes

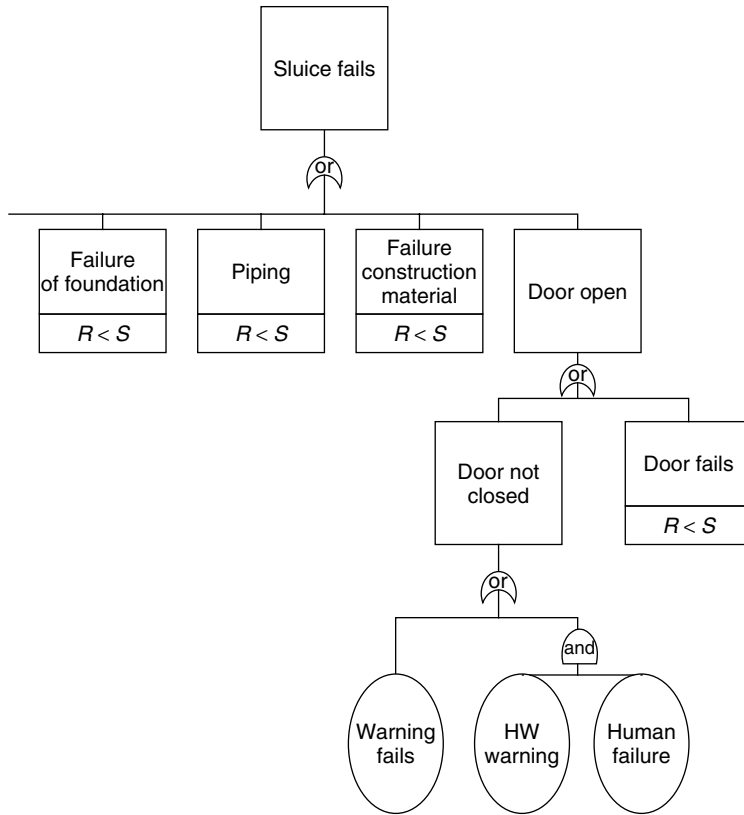


Figure 15 The sluice as a series system of failure modes

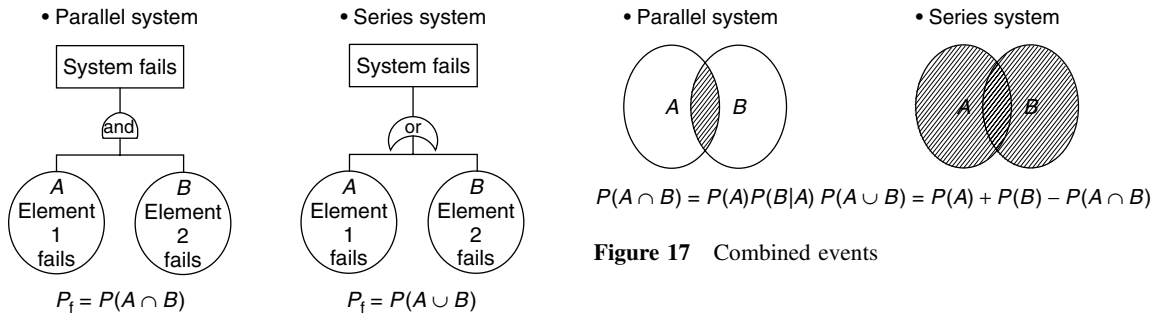


Figure 16 Fault trees for parallel and series systems

It can be seen that as long as the correlation coefficient is less than 90%, the failure probability is close to the upper limit.

In case of a series system with a large number of elements, the lower and upper bounds are as follows

- the maximum probability of the failure of a single element;
- and the sum of the failure probabilities of all the elements.

$$\max(P(i)) \leq P_f \leq \sum_{i=1}^n P(i) \quad (16)$$

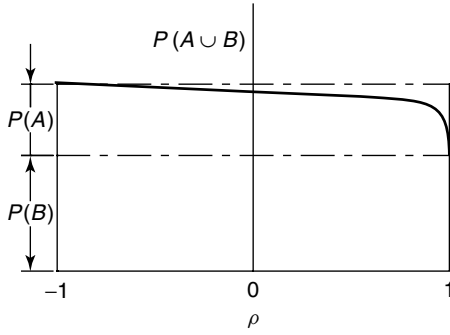


Figure 18 Probability of A or B, given ρ

Ditlevsen has narrowed these boundaries to get a better estimation of the failure probability.

$$P(1) + \sum_{i=2}^n \max \left(P(i) - \sum_{j=1}^{i-1} P(i \cap j) \right) \leq P_f$$

$$P_f \leq \sum_{i=1}^n P(i) - \sum_{i=2}^n \max_{j < i} (P(i \cap j)) \quad (17)$$

Let us now look at a series system that consists of two elements, each having two failure modes.

If the probabilities of the potential failure modes are known, it is possible to determine the upper limit of the failure probabilities of the element as the sum of the probabilities of the two different failure modes. The upper limit of the probability of failure of the system can be determined as the sum of the upper

limits of the failure probability of the two elements. So the upper limit of the failure probability is the sum of the probability of all the failure modes.

The scheme in Figure 19 for computation of the upper limit of the failure probability can also be used top down. For instance, if the allowable failure probability of the system is known, the scheme can be used to distribute this probability among the failure modes. In this way, we get a boundary condition for each failure mode that can be applied for the design.

Estimation of the Probability of a Failure Mode of an Element

After analyzing the failure probability of the system as a function of the probabilities of the failure modes, we need to know the probabilities of failure modes to estimate the failure probability. These probabilities can be determined by analyzing historical failure data or by probabilistic calculation of the limit states.

For most cases, there is not enough specific failure data available, so we have to determine the failure probabilities by computation.

The probabilistic computation uses the reliability function and the PDF of the variables as the base for the determination of the failure probability. A reliability function is a function of the strength and the load for a particular failure mode. In general, the formulation of the reliability function is given by $Z = R - S$ where R is the strength and S is the load. The failure mode will not occur as long as the

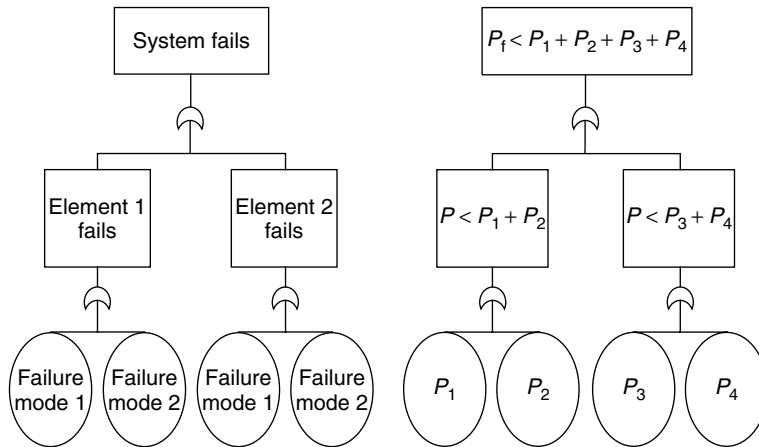


Figure 19 Probability of failure of a series system

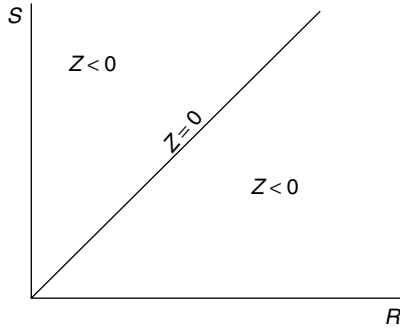


Figure 20 Reliability function

reliability function is positive. The graph of Figure 20 shows the reliability function. The line $Z = 0$ is a limit state. This line represents all the combinations of values of the strength and the loading for which the failure mode will just not occur. So it is a boundary between functioning and failure.

In the reliability function, the strength and load variables are assumed to be stochastic variables.

A stochastic variable is a variable that is defined by a probability distribution and a probability density function.

The probability distribution $F(x)$ returns the probability that the variable is less than x (as shown in Figure 21).

The PDF is the first derivative of the probability distribution.

If the distribution and the density of all the strength and load variables are known, it is possible to estimate the probability that the load has a value x and that the strength has a value less than x (see Figure 22).

$$\left. \begin{aligned} P(S = x) &= f_S(x) dx \\ P(R \leq x) &= F_R(x) \end{aligned} \right\} \implies P(S = x \cap R \leq x) = f_S(x) F_R(x) dx \quad (18)$$

The failure probability is the probability that $S = x$ and $R < x$ for every value of x . So we have to

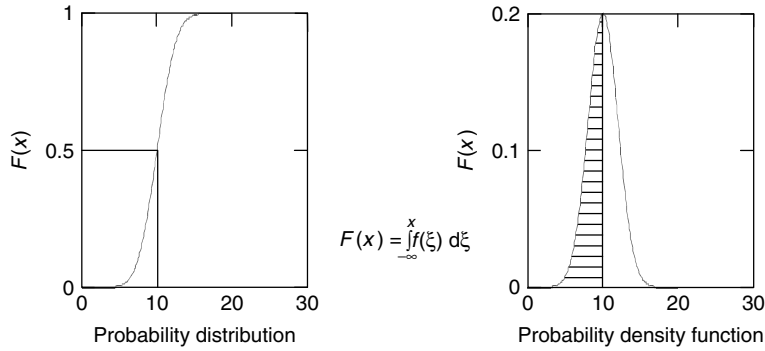


Figure 21 Probability distribution and probability density

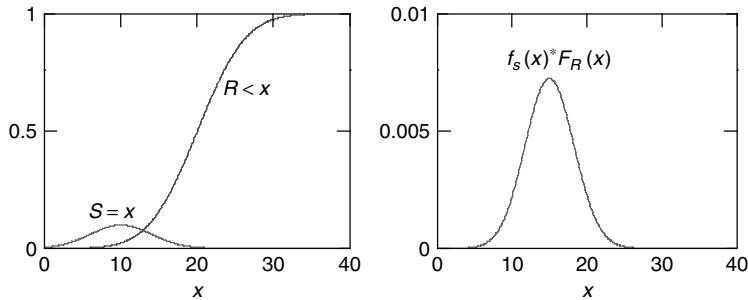


Figure 22 Components of the failure probability

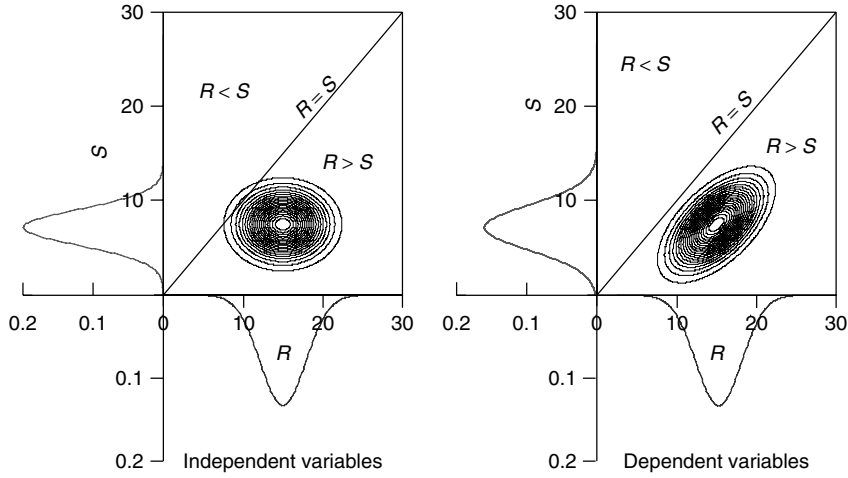


Figure 23 Joint probability density function

compute the sum of the probabilities for all possible values of x :

$$P_f = \int_{-\infty}^{\infty} f_S(x) F_R(x) dx \quad (19)$$

This method can be applied when the strength and the load are independent of each other.

Figure 23 gives the joint probability density of the strength and the load for a certain failure mode in which the strength and the load are not independent.

The strength is plotted on the horizontal axis and the load is plotted on the vertical axis.

The contours give the combinations of the strength and the load with the same probability density. In the area ($Z < 0$) the value of the reliability function is less than zero and the element will fail.

The failure probability can be determined by summation of the probability density of all the combinations of strength and load in this area.

$$P_f = \iint_{Z < 0} f_{RS}(r, s) dr ds \quad (20)$$

In a real case, the strength and the load in the reliability function are nearly always functions of multiple variables. For instance, the load can consist of the water level and the significant wave height. In this case, the failure probability is less simple to evaluate. Nevertheless, with numerical methods like numerical integration and Monte Carlo simulation (see **Risk Measures and Economic Capital for**

(Re)insurers; Structured Products and Hybrid Securities; Reliability Optimization; Uncertainty Analysis and Dependence Modeling), it is possible to solve the integral

$$P_f = \iint_{Z < 0} \dots \iint f_{r_1, r_2, \dots, r_n, s_1, s_2, \dots, s_m} \times (r_1, r_2, \dots, r_n, s_1, s_2, \dots, s_m) \times dr_1 dr_2 \dots r_n ds_1 ds_2 \dots ds_m \quad (21)$$

These methods that take into account the real distribution of the variables are called *level III probabilistic methods*. In the Monte Carlo simulation method, a large sample of values of the basic variables is generated and the number of failures is counted. The number of failures equals

$$N_f = \sum_{j=1}^N 1(g(\mathbf{x}_j)) \quad (22)$$

where N is the total number of simulations. The probability of failure can be estimated by

$$P_f \approx \frac{N_f}{N} \quad (23)$$

The CV of the failure probability can be estimated by

$$V_{P_f} \approx \frac{1}{\sqrt{P_f N}} \quad (24)$$

where P_f denotes the estimated failure probability.

The accuracy of the method depends on the number of simulations. The relative error made in the simulation can be written as

$$\varepsilon = \frac{\frac{N_f}{N} - P_f}{P_f} \quad (25)$$

The expected value of the error is zero. The standard deviation is given as

$$\sigma_\varepsilon = \sqrt{\frac{1 - P_f}{N P_f}} \quad (26)$$

For a large number of simulations, the error is normally distributed. Therefore, the probability that the relative error is smaller than a certain value E can be written as

$$P(\varepsilon < E) = \Phi\left(\frac{E}{\sigma_\varepsilon}\right) \quad (27)$$

$$N > \frac{k^2}{E^2} \left(\frac{1}{P_f} - 1 \right) \quad (28)$$

The probability of the relative error E being smaller than $k\sigma_\varepsilon$ now equals $\Phi(k)$. For desired values of k and E , the required number of simulations is determined as follows: For relative error of $E = 0.1$ lying within the 95% confidence interval ($k = 1.96$),

$$N > 400 \left(\frac{1}{P_f} - 1 \right) \quad (29)$$

The equation shows that the required number of simulations, and thus the calculation time, depend on the probability of failure to be calculated. Most structures in coastal and river engineering possess a relatively high probability of failure (i.e., a relatively low reliability) compared to structural elements/systems, resulting in reasonable calculation times for Monte Carlo simulation. The calculation time is independent of the number of basic variables and therefore Monte Carlo simulation should be favored over the Riemann method in case of a large number of basic variables (typically more than five). Furthermore, the Monte Carlo method is very robust, meaning that it is able to handle discontinuous failure spaces and reliability calculations in which more than one design point is involved (see below).

The problem of long calculation times can be partly overcome by applying importance sampling. This is not elaborated upon here. Reference is made to [14, 15].

If the reliability function (Z) is a sum of a number of normal distributed variables, then Z is also a normal distributed variable. The mean value and the standard deviation can easily be computed with these equations:

$$Z = \sum_{i=1}^n a_i x_i, \mu_Z = \sum_{i=1}^n a_i \mu_{x_i},$$

$$\sigma_Z = \sqrt{\sum_{i=1}^n (a_i \sigma_{x_i})^2} \quad (30)$$

This is the base of the level II probabilistic calculation. The level II methods approximate the distributions of the variables with normal distributions and they estimate the reliability function with a linear first order Taylor polynomial, so that the Z function is normal distributed.

If the distribution of the Z function is normal and the mean value and the standard deviation are known, it is rather easy to determine the failure probability (as shown in Figure 24).

By computing β as μ divided by σ it is possible to use the standard normal distribution to estimate the failure probability.

Tables of the standard normal distribution are available in the handbooks for statistics.

Nonlinear Z Function

A nonlinear Z function (see **Comonotonicity; Risk Classification/Life; Optimal Stopping and Dynamic Programming**) can be estimated with a

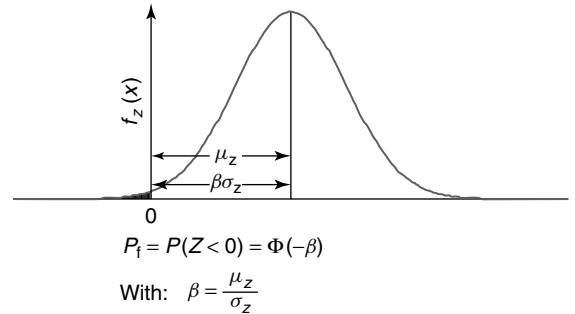


Figure 24 Probability density of the Z function

Taylor polynomial

$$Z(\vec{x}) \approx Z(\vec{x}^*) + \sum_{i=1}^n \frac{\partial Z}{\partial x_i} \cdot (x_i - x_i^*) \quad (31)$$

The function depends on the point where it is linearized. The mean value and the standard deviation of the linear Z function can be determined easily. If the reliability function is estimated by a linear Z function at the point where all the variables have their mean values ($x_i^* = \mu_{x_i}$), we call it a *mean value approach*.

The so-called design point approach estimates the reliability function by a linear function at a point on $Z = 0$ where the joint PDF has a maximum. Finding the design point is a maximization problem. For this problem, there are several numerical solutions, which are not discussed here.

Nonnormally Distributed Basic Variables

If the basic variables of the Z function are not normally distributed the Z function will be unknown and probably nonnormally distributed. To cope with this problem, the nonnormally distributed basic variables in the Z function can be replaced by a normally distributed variable. In the design point, the adapted normal distribution must have the same value as the real distribution. Because the normal distribution has two parameters (μ and σ), one condition is not enough to find the right normal distribution. Therefore, the value of the adapted normal PDF must also have the same value as the real PDF (see Figure 25).

The two conditions give a set of two equations with two unknowns, which can be solved:

$$\left. \begin{aligned} F_N(x^*) &= F_x(x^*) \\ f_N(x^*) &= f_x(x^*) \end{aligned} \right\} \implies \mu_N, \sigma_N \quad (32)$$

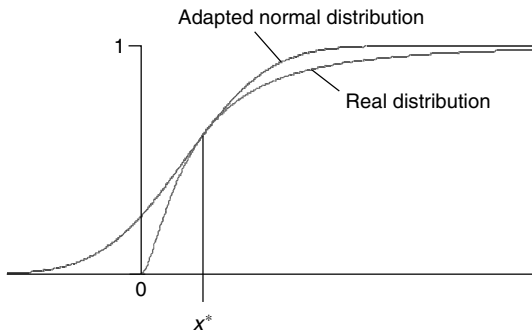


Figure 25 Adapted normal distribution

This method is known as the *approximate full distribution approach* (AFDA).

Acceptable Risk

A philosophy for acceptable risk (*see Actuary; Human Reliability Assessment; Multiattribute Modeling*) that takes into account the cost/benefit and the voluntariness aspect was developed by the Technical Advisory Committee for Water Retaining Structures [16]. The safety standard consists of a flexible evaluation of the individual and the societal acceptable risk but adds to these an economic approach taking the material damage into account. The latter provides the link with the safety philosophy of the Dutch dikes that was developed after the 1953 flood in the southwest of the Netherlands.

Personally Acceptable Level of Risk

The smallest component of the socially accepted level of risk is the personal assessment of risks by the individual. As an attempt to model this appraisal procedure quantitatively is not feasible, it is proposed to look with the insight gained to the preferences revealed in the accident statistics.

The actual personal risk levels inherent to various activities show statistical stability over the years and are approximately equal for the western countries, indicating a consistent pattern of preferences. The probability of losing one's life in normal daily activities such as driving a car or working in a factory appears to be 1 or 2 orders of magnitude lower than the overall probability of dying. Only a purely voluntary activity such as mountaineering entails a higher risk (Figure 26). This observation of public tolerance of 1000 times greater risks from voluntary than from involuntary activities with the same benefit was already made in the early 1970s.

In view of the consistency and the stability of the death risks presented, apart from a slightly downward trend due to technical progress, it would appear permissible to deduce there from a guideline for decisions with regard to the personally acceptable probability of failure

$$P_{fi} = \frac{\beta_i 10^{-4}}{P_{d|fi}} \quad (33)$$

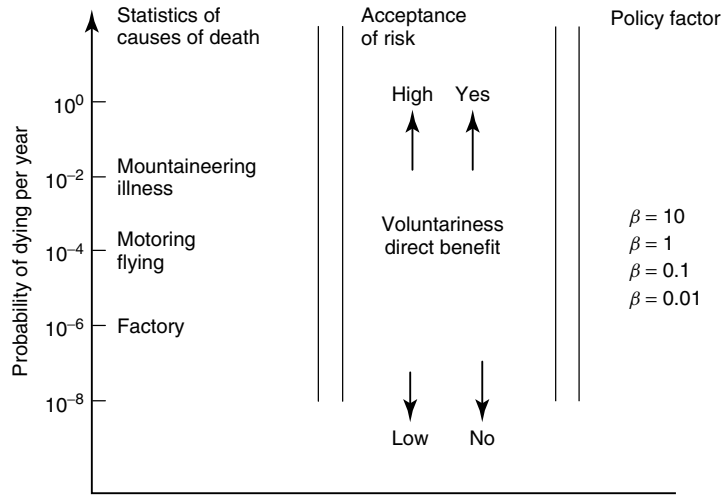


Figure 26 Personal risks in western countries, deduced from the statistics of causes of death and the number of participants per activity

where $P_{d|fi}$ denotes the probability of being killed in the event of an accident.

In this expression, the policy factor, β_i varies with the degree of voluntariness with which an activity is undertaken and with the benefit perceived. It ranges from 100 in the case of complete freedom of choice like mountaineering, to 0.01 in case of an imposed risk without any perceived direct benefit. This last case includes the individual risk criterion proposed by VROM [17], for the siting of a hazardous installation near a housing area without any direct benefit to the inhabitants.

Socially Acceptable Level of Risk

The basis of the framework with respect to societal risk is an evaluation of risks due to a certain activity on a national level. The risk evaluation on a national level has to be translated to local installations or activities to support a systematic appraisal by the local authorities.

If a risk criterion is defined on a local level (such as in [18]), the height of the national risk criterion is determined by the number of locations, where the activity takes place, and the type of PDF of the consequences of an accident. The resulting national norm has to be evaluated, as it was not intentionally formulated. It seems preferable to start with a risk criterion on a national level and to evaluate the acceptable local risk level in view of

the actual number of installations, the cost/benefit aspects of the activity and the general progress in safety in an iterative process with, say, a 10-year cycle (Figure 27).

Nationally Acceptable Level of Risk

The determination of the socially acceptable level of risk starts from the proposition that the result of a social process of risk appraisal is reflected in the accident statistics. It seeks to derive a standard from these revealed preferences. Using the circle of acquaintances as an instrument of observation, the very low probabilities of a fatal accident, which appear socially acceptable, are perceptible. The recurrence time is within the order of magnitude of a human life span.

In seeking to establish a norm for the acceptable level of risk for engineering structures it is more realistic to base oneself on the probability of a death due to a nonvoluntary activity in the factory, on board a ship, at sea, etc. which is approximately equal to: $1.4 \times 10^{-3} \text{ year}^{-1}$.

If this observation-based frequency is adopted as the norm for assessing the safety of activity i , then after rearranging the expression, and adopting a rather arbitrary distribution over some 20 categories of activities, each claiming an equal number of lives per year, the following norm is obtained for an activity i

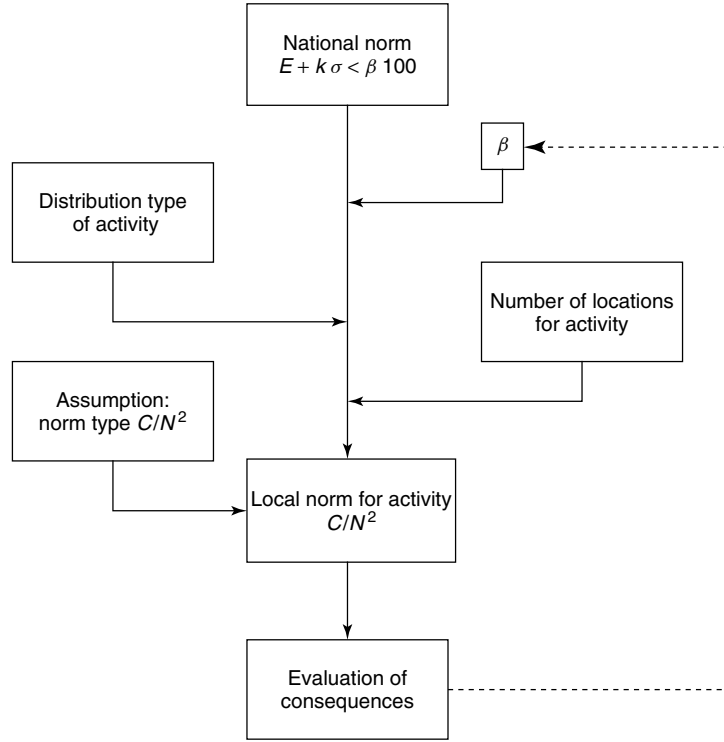


Figure 27 Framework for risk management

with N_{pi} participants in the Netherlands:

$$P_{fi} N_{pi} P_{d|fi} < \beta_i \times 100 \quad (34)$$

This norm should be interpreted in the sense that an activity is permissible as long as it is expected to claim fewer than $\beta_i \times 100$ deaths per year (in general: $\beta \times 7.10^{-6} \times$ national population size). The formula does not account for the standard deviation, which will certainly influence acceptance by a risk-averse community.

Risk aversion can be represented mathematically by increasing the mathematical expectation of the total number of deaths, $E(N_{di})$ by the desired multiple k of the standard deviation before the situation is tested against the norm:

$$E(N_{di}) + k \sigma(N_{di}) < \beta_i \times 100 \quad (35)$$

where

k : risk aversion index.

To determine the mathematical expectation and the standard deviation of the total number of deaths occurring annually in the context of activity i , it

is necessary to take into account the number of independent places N_A where the activity under consideration is carried out.

The model is tested for several activities in Vrijling *et al.* [19]. The agreement between the norm derived in this study for reasonable values of N_A and $0.01 < \beta_i < 100$ and the risk accepted in practice in the Netherlands seems to support the model.

Locally Acceptable Level of Risk

The translation of the nationally acceptable level of risk to a risk criterion for one single installation or location where an activity takes place depends on the distribution type of the number of casualties for accidents of the activity under consideration as shown above. To relate the new framework to the present one, it is assumed that on a local level the societal risk criterion is of the type proposed by VROM [17]:

$$1 - F_{N_{dij}}(x) < \frac{C_i}{x^2} \quad \text{for all } x \geq 10 \quad (36)$$

Assuming a Bernoulli distribution of the number of casualties, in the national criterion equation, and

taking account of N_{Ai} independent locations, gives the value of C_i

$$C_i = \left[\frac{-k\sqrt{N_{Ai}} + \sqrt{k^2 N_{Ai} + 4 \frac{N_{Ai}}{N} \beta_i \times 100}}{2 \frac{N_{Ai}}{N}} \right]^2 \quad (37)$$

If the expected value of the number of deaths is much smaller than its standard deviation, which is often true for calamities, the previous result reduces to

$$C_i = \left[\frac{\beta_i \times 100}{k\sqrt{N_{Ai}}} \right]^2 \quad (38)$$

Similar results are obtained for the exponential distribution.

The national societal acceptable risk criterion leads to a local acceptable risk criterion of the VROM type [17], which is inversely proportional to the number of independent places N_A and the square of the policy factor β_i :

$$1 - F_{N_{dij}}(x) \leq \frac{C_i}{x^2} \quad \text{for all } x \geq 100$$

$$\text{where } C_i = \left[\frac{\beta_i \times 100}{k\sqrt{N_{Ai}}} \right]^2 \quad (39)$$

The VROM rule is a special case of this general framework for acceptable risk: with $C_i = 10^{-3}$, $N_A = 1000$ (the approximate number of chemical installations), and $k = 3$, it follows that $\beta = 0.03$ which, according to Figure 26 is not unreasonable for an involuntarily imposed risk.

Economically Optimal Level of Risk

The problem of the acceptable level of risk can also be formulated as an economic decision problem. The expenditure I for a safer system is equated with the gain made by the decreasing present value of the risk (Figure 28).

The optimal level of safety indicated by P_{topt} corresponds to the point of minimal cost.

$$\min(Q) = \min(I(P_f) + PV(P_f S)) \quad (40)$$

where

Q : total cost

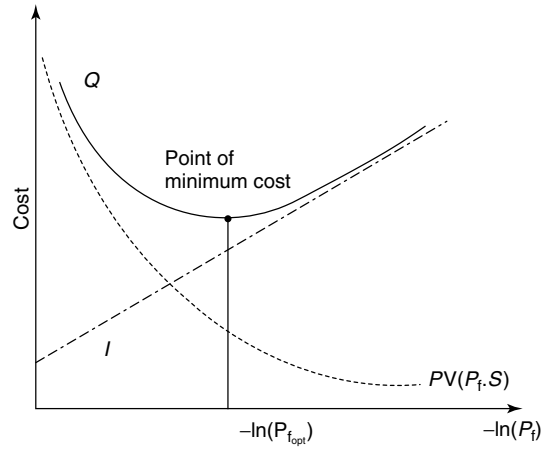


Figure 28 The economically optimal probability of failure of a structure

$I(P_f)$: investment to obtain a safety level corresponding P_f

PV : present value operator

S : total damage in case of failure.

If despite ethical objections, the value of a human life is rated at s , the amount of damage is increased to

$$P_{d|fi} N_{pi} s + S \quad (41)$$

where

N_{pi} : number of participants in activity I .

This extension makes the optimal failure probability a decreasing function of the expected number of deaths.

The valuation of human life is chosen as the present value of the net national product per inhabitant. The advantage of taking the possible loss of lives into account in economic terms is that the safety measures are affordable in the context of the national income. Risk aversion may also be included in the economic approach as shown by van Gelder and Vrijling [20].

References

- [1] Shannon, A. (1948). Mathematical theory of communication, *Bell Systems Technical Journal* **27**, 379–423, 623–656.
- [2] Vrijling, J.K. & van Gelder, P.H.A.J.M. (1998). The effect of inherent uncertainty in time and space on the reliability of flood protection, *ESREL'98: European Safety and Reliability Conference 1998*, 16–19 June 1998, Trondheim, pp. 451–456.

- [3] van Gelder, P.H.A.J.M. (2000). Statistical methods for the risk-based design of civil structures, *Communications on Hydraulic and Geotechnical Engineering*, ISSN:0169-6548 00-1, Tu Delft p. 248.
- [4] Efron, B. (1982). The jackknife, the bootstrap and other resampling plans, *CBMS-NSF Regional Conference Series in Applied Mathematics*, Monograph 38, SIAM, Philadelphia.
- [5] Kass, R.E. & Raftery, A.E. (1994). Bayes factors, Technical Report No. 254, Department of Statistics, University of Washington.
- [6] Volinsky, C.T., Madigan, D., Raftery, A.E. & Kronmal, R.A. (1996). Bayesian model averaging in proportional hazard models: assessing stroke risk, Technical Report No. 302, Department of Statistics, University of Washington.
- [7] De Vos, A.F. (1996). Fair and predictive Bayes factors for comparison of regression models, Technical Report, Vrije Universiteit, Amsterdam.
- [8] Wood, E.F. & Rodriguez-Iturbe, I. (1975). A Bayesian approach to analyzing uncertainty among flood frequency models, *Water Resources Research* **11**(6), 839–843.
- [9] Pericchi, L.R. & Rodriguez-Iturbe, I. (1983). On some problems in Bayesian model choice in hydrology, *The Statistician* **32**, 273–278.
- [10] Pericchi, L.R. & Rodriguez-Iturbe, I. (1985). On the statistical analysis of floods, *A Celebration of Statistics; The ISI Centenary Volume* A.C. Atkinson & S.E. Fienberg, eds, Springer, New York, pp. 511–541.
- [11] Perreault, L., Haché, M. Slivitzky, M. & Bobée, B. (1999). Detection of changes in precipitation and runoff over eastern Canada and U.S. using a Bayesian approach, *Stochastic Environmental Research and Risk Assessment* **13**(3), 201–216.
- [12] Ang, A.H.S. (1973). Structural risk analysis and reliability-based design, *Journal of Structural Division, ASCE* **99**(ST9), 1891–1910.
- [13] Vrijling, J.K. (1986). Engineering reliability and risk in water resources, *Probabilistic Design of Water Retaining Structures*, Martinus Nijhoff, Dordrecht.
- [14] Bucher, G.C. (1987). *Adaptive Sampling, An Iterative Fast Monte Carlo Procedure*, Internal Report, University of Innsbruck, Innsbruck.
- [15] Ditlevsen, O. & Madsen, H.O. (1996). *Structural Reliability Methods*, John Wiley & Sons, Chichester.
- [16] Technical Advisory Committee on Water Retaining (TAW) (1984). *Some Considerations on Acceptable Risk in the Netherlands*, Dienst Weg-en Waterbouwkunde, Delft.
- [17] Ministry of Housing, Spatial Planning and Environment, Land Use Planning and Environment (VROM) (1989). *Premises for Risk Management Risk Limits in the Context of Environmental Policy*, Annex to the Dutch National Environmental Policy Plan, Directorate General for Environmental Protection, The Hague.
- [18] Health and Safety Executive (HSE) (1989). *Risk Criteria for Land-Use Planning in the Vicinity of Major Industrial Hazards*, HM Stationery Office.
- [19] Vrijling, J.K., van Hengel, W. & Houben, R.J. (1995). A framework for risk evaluation, *Journal of Hazardous Materials* **43**(3), 245–261.
- [20] van Gelder, P.H.A.J.M. & Vrijling, J.K. (1997). Risk-averse reliability-based optimization of sea-defences, *Risk Based Decision Making in Water Resources*, ASCE Vol. 8, pp. 61–76.

Further Reading

Thoft-Christensen, P. & Baker, M.J. (1982). *Structural Reliability Theory and Its Applications*, Springer, Berlin/Heidelberg/New York, pp. 267.

PETRUS H.A.J.M. VAN GELDER
AND JOHANNES K. VRIJLING

Probabilistic Risk Assessment

Definition

Probabilistic risk assessment (PRA) (*see* **Environmental Health Risk Assessment**) is a systematic procedure for investigating how complex systems are built and operated. The PRAs model how human, software, and hardware elements of the system interact with each other. Also, they assess the most significant contributors to the risks of the system and determine the value of the risk. A formal definition proposed by Kaplan and Garrick [1] provides a simple and useful description of the elements of risk assessment that involves addressing three basic questions:

1. What can go wrong that could lead to exposure of hazards?
2. How likely is this to happen?
3. If it happens, what consequences are expected?

The PRA procedure involves quantitative application of the above triplets, in which probabilities (or frequencies) of scenarios of events leading to exposure of hazards are estimated and the corresponding magnitude of health, safety, environmental, and economic consequences for each scenario are predicted. The risk value (i.e., expected loss) of each scenario is often measured as the product of the scenario frequency and its consequences. The main result of the PRA is not the actual value of the risk computed (the so-called bottom-line number); rather it is the determination of the system elements that substantially contribute to the risks of that system, uncertainties associated with such estimates, and effectiveness of various risk reduction strategies available. In the remainder of this article, major steps in conducting a PRA are discussed.

Steps in Conducting a Probabilistic Risk Assessment

The following subsections provide a discussion of the essential components of PRA as well as the steps that must be performed in a PRA analysis.

The NASA PRA Guide [2] describes the components of the PRA as shown in Figure 1. In the following sections, each component of PRA is discussed in more details.

Objectives and Methodology

Preparing for a PRA begins with a review of the objectives of the analysis. Among many objectives that are possible, the most common ones include design improvement, risk acceptability, decision support, regulatory and oversight support, and operations and life management. Once the objectives are clarified, an inventory of possible techniques for the desired analyses should be developed. This step, in essence, provides a road map for the analysis. See Modarres [3] and Kumamoto and Henley [4] for the inventory of methodological approaches to PRA.

Familiarization and Information Assembly

A general knowledge of the physical layout of the overall system (e.g., facility, design, process, aircraft, or spacecraft), administrative controls, maintenance and test procedures, as well as barriers and subsystems, whose job is to protect, prevent, or mitigate hazard exposure conditions, is necessary to begin the PRA. A detailed inspection of the overall system must be performed in the areas expected to be of interest and importance to the analysis. The following items should be performed in this step:

1. Major critical barriers, structures, emergency safety systems, and human interventions should be identified.
2. Physical interactions among all major subsystems (or parts of the system) should be identified and explicitly described. The result should be summarized in a dependency matrix (*see* **Default Correlation**).
3. Past major failures and abnormal events that have been observed in the facility should be noted and studied. Such information would help ensure inclusion of important applicable scenarios.
4. Consistent documentation is critical to ensure the quality of the PRA. Therefore, a good filing system must be created at the outset and maintained throughout the study.

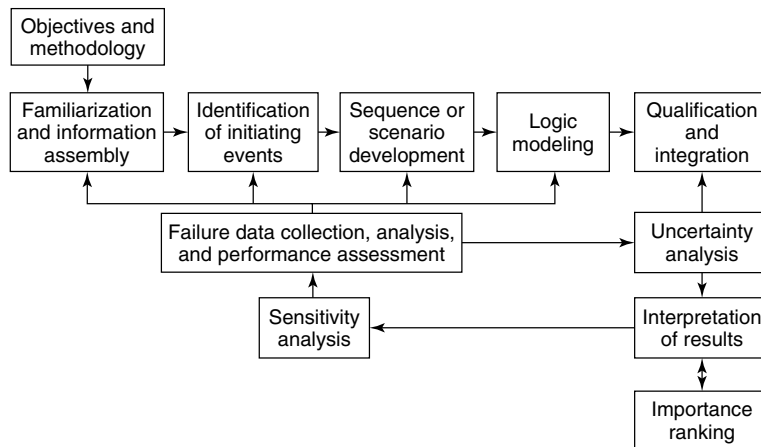


Figure 1 Components of the overall PRA process [Reproduced from [2]. National Aeronautics and Space Administration, 2002.]

Identification of Initiating Events

This task involves identifying those events (abnormal events or conditions) that could, if not correctly and timely responded to, result in hazard exposure (*see History and Examples of Environmental Justice; Combining Information; Bonus–Malus Systems; Longevity Risk and Life Annuities*). The first step involves identifying sources of hazard and barriers around these hazards. The next step involves identifying events that can lead to a direct threat to the integrity of the barriers.

During the normal operation mode of a system, loss of certain functions or subsystems will cause the process to enter an off-normal (transient) state transition. Once in this transition, there are two possibilities. First, the state of the system could be such that no other function is required to maintain the process in a safe condition (safe refers to a mode where the chance of exposing hazards beyond the system boundaries is negligible). The second possibility is a state wherein other functions (and thus systems) are required to prevent exposing hazards beyond the system boundaries. For this second possibility, the loss of the function or the system is considered an initiating event. Since such an event is related to the normally operating equipment, it is called an *operational initiating event*.

One method for determining the operational initiating events begins with first drawing a functional block diagram (*see Failure Modes and Effects*

Analysis) of the system. From the functional block diagram, a hierarchical relationship is produced, with the process objective being the successful completion of the desired system. Each function can then be decomposed into its subsystems and components and combined in a logical manner to represent operations needed for the success of that function.

An alternative to the use of functional hierarchy for identifying initiating events is the use of failure mode and event analysis (FMEA) (*see Human Reliability Assessment; Reliability Integrated Engineering Using Physics of Failure*) (*see Stamatis [5]*). The difference between these two methods is noticeable, namely, the functional hierarchies are deductively and systematically constructed, whereas FMEA is an inductive and experiential technique. In both of the above methods, one can always supplement the set of initiating events with generic initiating events (if known). For example, *see Sattison et al. [6]* for these initiating events for nuclear reactors, and NASA Guide [2] for space vehicles. The following procedures should be followed in this step of the PRA:

1. Select a method for identifying specific operational and nonoperational initiating events. Two of the representative methods are functional hierarchy and FMEA. If a generic list of initiating events is available, it can be used as a supplement.
2. Using the method selected, identify a set of initiating events.

3. Group the initiating events having the same effect on the system; for example, those requiring the same mitigating functions to prevent hazard exposure are grouped together.
4. Develop a systemic event tree for each initiating event, delineating the success conditions, initiating event progression phenomena, and end effect of each scenario.

For specific examples of scenario development, see [2–4].

Sequence or Scenario Development

The goal of scenario development (*see Managing Infrastructure Reliability, Safety, and Security; Environmental Exposure Monitoring; Dynamic Financial Analysis*) is to derive a complete set of scenarios that encompasses all of the potential exposure propagation paths that can lead to loss of containment or confinement of the hazards, following the occurrence of an initiating event. The scenarios that describe the functional response of the process to the initiating events are frequently displayed by event trees. The event-tree (*see Systems Reliability; Expert Judgment; Digital Governance, Hotspot Detection, and Homeland Security*) development techniques are discussed in [2–4].

In PRA, two types of event trees can be developed: functional and systemic. The functional event tree uses mitigating functions as its heading. The main purpose of the functional tree is to better understand the scenario of events at an abstract level, following the occurrence of an initiating event. The functional tree also guides the PRA analyst in the development of a more detailed systemic event tree. The systemic event tree reflects the scenarios of specific events (specific human actions, protective or mitigative subsystem operations, or failures) that lead to a hazard exposure. Therefore, a systemic event tree fully delineates the overall system response to an initiating event and serves as the main tool for further analyses in the PRA. For detailed discussion on specific tools and techniques used for this purpose, see Modarres [7]. The following procedures should be followed in this step of the PRA:

1. Identify the mitigating functions for each initiating event (or group of events).
2. Identify the corresponding human actions, systems, or hardware operations associated with each function, along with their necessary conditions for success.
3. Develop a functional event tree for each initiating event (or group of events).

Logic Modeling

Event trees commonly involve branch points at which a given subsystem (or event) either works (or happens) or does not work (or does not happen). Sometimes, failure of these subsystems (or events) is rare and there may not be an adequate record of observed failure events to provide a historical basis for estimating frequency of their failure. In such cases, other logic-based analysis methods such as fault trees or master logic diagrams may be used, depending on the accuracy desired. The fault-tree analysis involves developing a logic model (*see Early Warning Systems (EWSs) for Predicting Financial Crisis; Expert Elicitation for Risk Assessment*) in which the subsystem is broken down into its basic components or segments for which adequate data exist. For more details about fault trees associated to the event headings of an event tree, see Modarres *et al.* [8]. The following procedures should be followed as a part of developing the fault tree:

1. Develop a fault tree for each event in the event-tree heading for which actual historical failure data does not exist.
2. Explicitly model dependencies of a subsystem on other subsystems and intercomponent dependencies (e.g., common-cause failures). For common-cause failures, see Mosleh *et al.* [9].
3. Include all potential reasonable and probabilistically quantifiable causes of failure, such as hardware, software, test and maintenance, and human errors, in the fault tree.

The following steps should be followed in the dependent failure analysis:

1. Identify the hardware, software, and human elements that are similar and could cause dependent or common-cause failures.
2. Items that are potentially susceptible to common-cause failure should be explicitly incorporated

into the corresponding fault trees and event trees of the PRA where applicable.

3. Functional dependencies should be identified and explicitly modeled in the fault trees and event trees.

Failure Data Collection, Analysis, and Performance Assessment

A critical building block in assessing the reliability and availability of complex systems is the data on the performance of its barriers to contain hazards. In particular, the best resources for predicting future availability are past field experiences and tests. Hardware, software, and human reliability data are inputs to assess performance of hazard barriers, and the validity of the results depends highly on the quality of the input information. Collection of the various failure data consists fundamentally of the following steps: collecting generic data, assessing generic data, statistically evaluating facility-specific or overall system-specific data, and developing failure probability distributions using test and/or facility-specific and system-specific data. Three types of events identified during the risk scenario definition and system modeling must be quantified for the event trees and fault trees to estimate the frequency of occurrence of sequences: initiating events, component failures, and human error.

Typically, the necessary data include time of failures, repair times, test frequencies, test downtimes, and common-cause failure events (*see Systems Reliability; Reliability Data; Common Cause Failure Modeling*). Further uncertainties associated with such data must also be characterized. Kapur and Lamberson [10], Modarres *et al.* [8], and Nelson [11] discuss the methods available for analyzing data to obtain the probability of failure or the probability of occurrence of equipment failure. Also, Crow [12] and Ascher and Feingold [13] discuss analysis of data relevant to repairable systems. Finally, Mosleh *et al.* [9] discuss the analysis of data for dependent failures, Poucet [14] reviews human reliability issues (*see Human Reliability Assessment; Influence Diagrams*), and Smidts [15] examines software reliability models. Establishment of the database to be used will generally involve collection of some facility-specific or system-specific data combined with the use of generic performance data when specific data are absent or sparse. For example, [16–18] describe generic data

for electrical, electronic, and mechanical equipment. The following procedures should be followed as part of the data analysis task:

1. Determine generic values of material strength or endurance, load or damage agents, failure times, failure occurrence rate, and failures on demand for each item (hardware, human action, or software) identified in the PRA models. This can be obtained either from facility-specific or system-specific experiences, from generic sources of data, or both.
2. Gather data on hazard barrier tests, repair, and maintenance data primarily from experience, if available. Otherwise use generic performance data.
3. Assess the frequency of initiating events and other probability of failure events from experience, expert judgment, or generic sources.
4. Determine the dependent or common-cause failure probability for similar items, primarily from generic values. However, when significant specific data are available, they should be primarily used.

Quantification and Integration

Fault trees and event trees are integrated and their events are quantified to determine the frequencies of scenarios and associated uncertainties in the calculation of the final risk values. This integration depends somewhat on the manner in which system dependencies have been handled. We describe the more complex situation, in which the fault trees are dependent, i.e., there are physical dependencies (e.g., through support units of the main hazard barriers such as those providing motive, proper working environment, and control functions).

Normally, the quantification will use a Boolean reduction process to arrive at a Boolean representation for each scenario. Starting with fault-tree models for the various systems or event headings in the event trees, and using probabilistic estimates for each of the events modeled in the event trees and fault trees, the probability of each event-tree heading (often representing failure of a hazard barrier) is calculated (if the heading is independent of other headings). The fault trees for the main subsystems and support units (e.g., lubricating and cooling units and power units) are merged where needed, and the equivalent Boolean

expression representing each event in the event-tree model is calculated. The Boolean expressions are reduced to arrive at the smallest combination of basic failure events (the so-called minimal cut sets) that lead to exposure of the hazards. The minimal cut sets for the event-tree event headings are then appropriately combined to determine the cut sets for the event-tree scenarios. If possible, all minimal cut sets must be generated and retained during this process; unfortunately, in complex systems and facilities this leads to an unmanageably large collection of terms and a combinatorial outburst. Therefore, the collection of cut sets are often truncated (i.e., probabilistically small and insignificant cut sets are discarded on the basis of the number of terms in a cut set or on the probability of the cut set).

Even though the individual cut sets discarded may be several orders of magnitude less probable than the average of those retained, the large number of them discarded may sum to a significant part of the risk. The actual risk might thus be larger than what the PRA results indicate. This can be discussed as part of the modeling uncertainty characterization. Detailed examination of a few PRA studies of very complex systems, for example, nuclear power plants, shows that cut set truncation will not introduce any significant error in the total risk assessment results (see Dezfuli and Modarres [19]).

Other methods for evaluating scenarios also exist that directly estimate the frequency of scenario without specifying cut sets. This is often done in highly dynamic systems whose configuration changes as a function of time leading to dynamic event tree and fault trees (see **Canonical Modeling**). For more discussion on these systems, see Chang *et al.* [20], NASA Procedures PRA Guide [2], and Dugan *et al.* [21]. Employing advanced computer programming concepts, one may also directly simulate the operation of parts to mimic the real system for reliability and risk analysis (see Azarkhail and Modarres [22]). An example of PRA for compressed natural gas safety is presented in Chamberlain and Modarres [23]. Interested reader is also referred to Modarres [7] for a handful of numerical examples on PRA in a wide range of applications. The following procedures should be followed as part of the quantification and integration step in the PRA:

1. Merge corresponding fault trees associated with each failure or success event modeled in the

event-tree scenarios (i.e., combine them in a Boolean form). Develop a reduced Boolean function for each scenario (i.e., truncated minimal cut sets).

2. Calculate the total frequency of each sequence, using the frequency of initiating events, the probability of barrier failure including contributions from test and maintenance frequency (outage), common-cause failure probability, and human error probability.
3. Use the minimal cut sets of each sequence for the quantification process. If needed, simplify the process by truncating on the basis of the cut sets or probability.
4. Calculate the total frequency of each scenario.
5. Calculate the total frequency of all scenarios of all event trees.

Uncertainty Analysis

Uncertainties are part of any assessment, modeling, and estimation. In engineering calculations, we routinely ignore the estimation of uncertainties associated with failure models and parameters, because the uncertainties are very small and more often analyses are done conservatively (e.g., by using high safety factor, design margin). Since PRAs are primarily used for decision making and management of risk, it is critical to incorporate uncertainties in all facets of the PRA.

In PRAs, uncertainties are primarily shown in the form of probability distributions. For example, the probability of failure of a subsystem (e.g., a hazard barrier) may be represented by a probability distribution showing the range and likelihood of risk values.

The process involves characterization of the uncertainties associated with frequency of initiating events, probability of failure of subsystems (or barriers), probability of all event-tree headings, strength or endurance of barriers, applied load or incurred damage by the barriers, amount of hazard exposures, consequences of exposures to hazards, and sustained total amount of losses. Other sources of uncertainties are in the models used – for example, the fault-tree and event-tree models, stress-strength and damage-endurance models used to estimate failure or capability of some barriers, probabilistic failure models of hardware, software, and human, correlation between amount of hazard exposure and the consequence,

exposure models and pathways, and models to treat inter- and intrabarrier failure dependencies. Another important source of uncertainty is incompleteness of the risk models and other failure models used in the PRAs.

Once uncertainties associated with hazard barriers have been estimated and assigned to models and parameters, they must be “propagated” through the PRA model to find the uncertainties associated with the results of the PRA, primarily with the bottom-line risk calculations, and with the list of risk significant elements of the system. Propagation is done using one of the several techniques, but the most popular method used is Monte Carlo simulation (*see Risk Measures and Economic Capital for (Re)insurers; Simulation in Risk Management; Expert Elicitation for Risk Assessment; Reliability Optimization*). The results are then shown and plotted in the form of probability distributions. Steps in uncertainty analysis include the following:

1. Identify models and parameters that are uncertain and the method of uncertainty estimation to be used for each.
2. Describe the scope of the PRA and significance and contribution of elements that are not modeled or considered.
3. Estimate and assign probability distributions depicting model and parameter uncertainties in the PRA.
4. Propagate uncertainties associated with the hazard barrier models and parameters to find the uncertainty associated with the risk value.
5. Present the uncertainties associated with risks and contributors to risk in an easy way to understand and visually straightforward to grasp.

Sensitivity Analysis

Sensitivity analysis is the method of determining the significance of choice of a model or its parameters, assumptions for including or not including a barrier, phenomena or hazard, performance of specific barriers, intensity of hazards, and significance of any highly uncertain input parameter or variable to the final risk value calculated. The process of sensitivity analysis is straightforward. The effects of the input variables and assumptions in the PRA are measured by modifying them by several folds, factors, or even one or more order of magnitudes, one at a time,

and measure relative changes observed in the PRA’s risk results. Those models, variables, and assumptions whose change leads to the highest change in the final risk values are determined as “sensitive”. A good sensitivity analysis strengthens the quality and validity of the PRA results. Interested readers are referred to Modarres [7] for more detailed discussions on sensitivity analysis. The steps involved in the sensitivity analysis are as follows:

1. Identify the elements of the PRA (including assumptions, failure probabilities, models, and parameters) that analysts believe might be sensitive to the final risk results.
2. Change the contribution or value of each sensitive item in either direction by several factors in the range of 2–100. Note that certain changes in the assumptions may require multiple changes of the input variables. For example, a change in failure rate of similar equipment requires changing of the failure rates of all these equipment in the PRA model.
3. Calculate the impact of the changes in step 2, one at a time, and list the elements that are most sensitive.
4. On the basis of the results in step 3, propose additional data, any changes in the assumptions, use of alternative models, and modification of the scope of the PRA analysis.

Risk Ranking and Importance Analysis

Ranking the elements of the system with respect to their risk or safety significance is one of the most important outcomes of a PRA. Ranking is simply arranging the elements of the system on the basis of their increasing or decreasing contribution to the final risk values. The ranking process should be done with much care. In particular, during the interpretation of the results, since formal importance measures are context dependent and their meaning varies depending on the intended application of the risk results, the choice of the ranking method is important.

There are several unique importance measures in PRAs. For example, Fussell–Vesely [24], Birnbaum [25], risk reduction worth (RRW), and risk achievement worth (RAW) [8] are identified as appropriate measures for use in PRAs, and all are representative of the level of contribution of various elements

of the system as modeled in the PRA and enter in the calculation of the total risk of the system. Another important set of importance measures focus on ranking the elements of the system with respect to their contribution to the total uncertainty of the risk results obtained from PRAs. This process is called *uncertainty ranking* and is different from component, subsystem, and barrier ranking. In this importance ranking, the analyst is only interested to know which of the system elements drive the final risk uncertainties, so that resources can be focused on reducing important uncertainties.

For additional discussions on the risk ranking methods and their implications in failure and success domains, see Azarkhail and Modarres [26]. The following are the major steps of importance ranking:

1. Determine the purpose of the ranking and select appropriate ranking importance measure that has consistent interpretation for the use of the ranked results.
2. Perform risk ranking and uncertainty ranking, as needed.
3. Identify the most critical and important elements of the system with respect to the total risk values and total uncertainty associated with the calculated risk values.

Interpretation of Results

When the risk values are calculated, they must be interpreted to determine whether any revisions are necessary to refine the results and the conclusions. There are two main elements involved in the interpretation process. The first is to understand whether the final values and details of the scenarios are logically and quantitatively meaningful. This step verifies the adequacy of the PRA model and the scope of analysis. The second is to characterize the role of each element of the system in the final results. This step highlights additional analyses data and information gathering that would be considered necessary.

The interpretation step is a continuous process with receiving information from the quantification, sensitivity, uncertainty, and importance analysis activities of the PRA. The process continues until the final results can be best interpreted and used in the subsequent risk-management steps. The basic steps of the PRA results interpretation are as follows:

1. Determine accuracy of the logic models and scenario structures, assumptions, and scope of the PRA.
2. Identify system elements for which better information would be needed to reduce uncertainties in failure probabilities and models used to calculate performance.
3. Revise the PRA and reinterpret the results until attaining stable and accurate results.

References

- [1] Kaplan, S. & Garrick, J. (1981). On the quantitative definition of risk, *Risk Analysis* 1(1), 11–28.
- [2] Stamatelatos, M., Apostolakis, G., Dezfuli, H., Everline, C., Guarro, S., Moieni, P., Mosleh, A., Paulos, T. & Youngblood, R. (2002). *Probabilistic Risk Assessment Procedures Guide for NASA Managers and Practitioners*, Version 1.1, National Aeronautics and Space Administration, Washington, DC.
- [3] Modarres, M. (1993). *What Every Engineer Should Know about Reliability and Risk Analysis*, Marcel Dekker, New York.
- [4] Kumamoto, H. & Henley, E.J. (1996). *Probabilistic Risk Assessment for Engineers and Scientists*, IEEE Press, New York.
- [5] Stamatis, D.H. (2003). *Failure Mode and Effect Analysis: FMEA from Theory to Execution*, 2nd Edition, ASQ Quality Press, Wisconsin.
- [6] Sattison, M.B. & Hall, K.W. (1990). *Analysis of Core Damage Frequency: Zion, Unit 1 Internal Events*, NUREG/CR-4550, Vol. 7, Rev. 1.
- [7] Modarres, M. (2006). *Risk Analysis in Engineering, Techniques, Tools and Trends*, CRC Press, Boca Raton.
- [8] Modarres, M., Kaminskiy, M. & Krivtsov, V. (1999). *Reliability Engineering and Risk Analysis: A Practical Guide*, Marcel Dekker, Basel.
- [9] Mosleh, A., Fleming, K.N., Parry, G.W., Paula, H.M., Worledge, D.H. & Rasmuson, D.M. (1988). *Procedure for Treating Common Cause Failures in Safety and Reliability Studies*, NUREG/CR-4780, U.S. Nuclear Regulatory Commission, Washington, DC, Vols I and II.
- [10] Kapur, K.C. & Lamberson, L.R. (1977). *Reliability in Engineering Design*, John Wiley & Sons, New York.
- [11] Nelson, W. (1990). *Accelerated Testing: Statistical Models, Test Plans and Data Analyses*, John Wiley & Sons, New York.
- [12] Crow, L.H. (1990). Evaluating the reliability of repairable systems, in *Proceedings of Annual Reliability Maintainability Symposium*, IEEE.
- [13] Ascher, H. & Feingold, H. (1984). *Repairable Systems Reliability: Modeling and Inference, Misconception and Their Causes*, Marcel Dekker, New York.

- [14] Poucet, A. (1988). Survey of methods used to assess human reliability in the human factors reliability benchmark exercise, *Reliability Engineering and System Safety* **22**, 257–268.
- [15] Smidts, C. (1996). Software reliability, in *The Electronics Handbook*, J.C. Whitaker, ed, CRC Press & IEEE Press, Boca Raton.
- [16] Center for Chemical Process Safety (1989). *Guidelines for Process Equipment Reliability Data, with Data Tables*, American Institute of Chemical Engineers (AIChE), New York.
- [17] Department of Defense (1995). *Military Handbook, Reliability Prediction of Electronic Equipment (MIL-HDBK-217F)*.
- [18] IEEE Standards (1984). *Guide to the Collection and Presentation of Electrical, Electronic, Sensing Component and Mechanical Equipment Reliability Data for Nuclear Power Generating Stations*, IEEE Std. 500, New York.
- [19] Dezfuli, H. & Modarres, M. (1984). A truncation methodology for evaluation of large fault trees, *IEEE Transactions on Reliability* **R-33**, 325–328.
- [20] Chang, Y.H., Mosleh, A. & Dang, V. (2003). *Dynamic Probabilistic Risk Assessment: Framework, Tool, and Application*, Society for Risk Analysis Annual Meeting, Baltimore.
- [21] Dugan, J., Bavuso, S. & Boyd, M. (1993). Dynamic fault tree models for fault tolerant computer systems, *IEEE Transactions on Reliability* **40**(3), 363.
- [22] Azarkhail, M. & Modarres, M. (2006). An intelligent-agent-oriented approach to risk analysis of complex dynamic systems with applications in planetary missions, in *Proceedings of the Eighth International Conference on Probabilistic Safety Assessment and Management*, ASME, New Orleans.
- [23] Chamberlain, S. & Modarres, M. (2005). Compressed natural gas safety: a quantitative analysis, *Risk Analysis Journal* **25**, 377–389.
- [24] Fussell, J. (1975). How to hand calculate system reliability and safety characteristics, *IEEE Transaction on Reliability* **R-24**(3), 169–174.
- [25] Birnbaum, Z.W. (1969). On the importance of different components in a multicomponent system, in *Multivariate Analysis II*, P.R. Krishnaiah, ed, Academic Press, New York.
- [26] Azarkhail, M. & Modarres, M. (2004). A study of implications of using importance measures in risk-informed decisions, *PSAM-7, ESREL 04 Joint Conference*, Berlin.

MOHAMMAD MODARRES

Product Risk Management: Testing and Warranties

The main purpose of product testing is to identify and rectify defects in order to reduce unexpected failures in the field. Manufacturers employ various testing strategies during a product's life cycle to validate their product's design and gain insight into the product performance in a customer's environment. Product testing should be seen as a key ingredient for competitive advantage, as a thoroughly tested product will lead to high consumer satisfaction and market superiority. This article presents tests performed during the prototype and production phases. These testing methodologies are applicable to the manufacture of biomedical devices, computer hardware and peripheral equipment, semiconductor equipment, and consumer electronics, among other items.

Prototype Phase Testing

During the prototyping stage, engineers build working samples of the product they have designed. Design verification testing (DVT) is used on the prototypes to verify functionality against predetermined functional specifications and design goals. This information is used to validate the design or identify areas of the design that need to be revised. DVT is also used to ensure that any new functions introduced late in the product cycle do not adversely affect any previously available functions ("regression testing"). DVT marks the beginning of the design refinement phase in the design cycle. During this cycle, the design is revised and improved to meet performance requirements and design specifications. Highly accelerated life testing (HALT) provides a quick way to precipitate failures and reveal design weaknesses. HALT is often used only on high risk modules or key components. During the HALT process, the products are stressed with stimuli beyond the expected field environments to determine their operating and destruction limits. Initially, each stress is stepped up individually to the point of failure and, at the end of HALT, different stresses are combined to maximize their effects. Typical stresses used in HALT are thermal and vibration stress, rapid thermal transients, humidity, voltage

margin, and power cycles. The increased stress levels also provide a simulated aging effect on the product, so that failures that normally show up in the field only after a long time at the usual stress levels can be quickly discovered during HALT.

One important element of the HALT process is the root cause analysis to determine which failures are relevant and to evaluate their impact of the products in the field. If corrective actions can be implemented during testing, then the follow-on test would provide feedback on how effective the corrective actions are.

Time-to-failure data from HALT may be analyzed to draw inferences about product reliability. Product life is related to the stress level by an acceleration equation. This allows the manufacturer to extrapolate a reliability estimate for the working environment from units tested at intensified stress levels. The most frequently used models are the Arrhenius and Eyring models for temperature acceleration, and the power law model (Nelson [1], Meeker and Escobar [2], and NIST [3]) for voltage and vibration stresses.

Let X_s and X_u be two random variables representing the time to failure at stress and at use, respectively. Then, under a linear acceleration relationship between stress and use, we have $X_s = AF \times X_u$, where AF is a constant representing the acceleration factor.

In the Arrhenius model, which is based on the reaction rate of a failure causing chemical or physical process due to elevated temperature, the acceleration factor between a higher temperature T_2 and a lower temperature T_1 is given by

$$AF = \exp \left\{ \frac{\Delta H}{k} \left(\frac{1}{T_1} - \frac{1}{T_2} \right) \right\} \quad (1)$$

with temperatures T_1 and T_2 measured in kelvin and k denoting the Boltzmann's constant (8.617×10^{-5} in electronvolt per kelvin). The constant ΔH is the activation energy, the minimum energy necessary for a specific chemical reaction to occur. The value of ΔH depends on the failure mechanism and the material involved, and typically ranges from 0.3 or 0.4 up to 1.5 or higher. Applications of the Arrhenius model include battery cells and dielectrics.

The Eyring model, derived from quantum mechanics, includes stress due to elevated temperature and can be expanded to include other type of stresses. The temperature term is very similar to the Arrhenius

model, with the acceleration factor

$$AF = \left(\frac{T_1}{T_2}\right)^\alpha \exp \left\{ \frac{\Delta H}{k} \left(\frac{1}{T_1} - \frac{1}{T_2} \right) \right\} \quad (2)$$

The generalized Eyring model is used to model the situation where the device is subjected to two types of stresses, a thermal stress and a nonthermal stress, and is applicable in the accelerated testing of thin films. Let S_1 and S_2 denote two different levels of a nonthermal stress such as voltage or current. Then the acceleration factor is as follows:

$$AF = \left(\frac{T_1}{T_2}\right)^\alpha \exp \left\{ \frac{\Delta H}{k} \left(\frac{1}{T_1} - \frac{1}{T_2} \right) \right. \\ \left. + B(S_1 - S_2) + C \left(\frac{S_1}{T_1} - \frac{S_2}{T_2} \right) \right\} \quad (3)$$

The model parameters are α , ΔH , B , and C . As in the Arrhenius Model, k is Boltzmann's constant, ΔH is the activation energy, and temperatures T_1 and T_2 are measured in Kelvin. The parameter C is used to model the interaction between the thermal and nonthermal stresses. The model can also be expanded to include additional nonthermal stresses. Two more parameters are added to the model with each additional nonthermal stress.

The power law model uses voltage as the only stress and is a simplified version of the generalized Eyring model where $\alpha = \Delta H = C = 0$, $B = -\beta$ and the voltage stress is measured by $S = \ln(V)$. The acceleration factor is as follows:

$$AF = \left(\frac{V_1}{V_2}\right)^{-\beta} \quad (4)$$

The power law model is applicable for testing capacitors.

Throughout the prototype phase testing, problems are identified and corrective actions are taken, which may result in design revision. *Reliability growth* refers to the improvement of the product reliability over this period of time due to changes in the product's design. For example, Crow [4] modeled reliability growth processes for a complex repairable system and provided procedures for estimating the parameters of the underlying stochastic process. Let $N(t)$ denote the cumulative number of failures at time t . The intensity function $\rho(t) = d[EN(t)]/dt$ is constant for a time-homogeneous Poisson process. In the Crow model, the intensity function is $\rho(t) = \lambda\beta t^{\beta-1}$, i.e., a Weibull process (Finkelstein

[5]), owing to the fact that the time to first failure under this growth process has a Weibull distribution. Crow first developed the model at the US Army Material Systems Analysis Activity (AMSAA) and it is also known as the *Crow-AMSAA model*. The model has been widely adopted to model the reliability of semiconductor test equipment (Jin *et al.* [6]) and equipment reliability in commercial nuclear power plant (Sun *et al.* [7]). It is the standard model for repairable systems in the International Electrotechnical Commission (IEC) standards.

HALT and reliability growth analysis supply manufacturers with an initial reliability assessment derived from product testing, which can be combined with estimates from reliability prediction to forecast warranty costs while the product is still in the prototype phase. Manufacturers use reliability prediction during the product design and development period to evaluate the life expectancy of their products. Reliability prediction is relatively inexpensive and gives designers the capability to compare different concepts and designs in terms of reliability. Before the prototypes are built, this is their only means of setting product reliability goals and making competitive evaluations.

The most commonly used predictive models are based on either MIL-HDBK-217 or Telcordia SR-332, using only the parts count and parts type information from the product bill of material (BOM). MIL-HDBK-217 is a military handbook for reliability prediction of electronic equipment. It contains the failure rate for most part types such as ICs, transistors, diodes, capacitors, and connectors and is based on estimates from field data. The Telcordia Standard is a modification of MIL-HDBK-217 that allows the parts count data to be enhanced by laboratory test data and field performance data. It also provides a method for predicting failure rates during the first year of operation accounting for burn-in times and temperatures.

Production Phase Testing

At the start of volume production, product verification testing (PVT) is performed on the first few production units to demonstrate that the design has been correctly implemented in production. In contrast to the prototype phase testing, where subsystems may be tested separately, subsystems are integrated into a complete

system and tested in PVT. This may reveal interaction failure modes. Test results from PVT are analyzed to set the initial burn-in duration for volume production and to support key product checkpoints such as the early ship and general availability programs.

A reliability qualification test (RQT) is used to demonstrate the product reliability that was previously derived using reliability prediction during the design cycle. RQT is usually performed at the system level concurrent with or immediately after PVT. It is typically set up as an acceptance test. MIL-STD-781D provides the acceptance criteria and decision risks for various fixed-length and sequential test plans for products characterized by an exponential time-to-failure distribution. To shorten the test duration, RQTs are occasionally run as an accelerated test. This may be accomplished by simply increasing the duty cycle from the end-use environment. The basic assumption made in this type of accelerated test is that the product does not deteriorate while in standby mode. In more complex situations, test engineers will rely on the HALT results to determine the types of stresses, stress levels, and acceleration factor to apply in RQT.

The highly accelerated stress screen (HASS) is used to precipitate early failures in production using a stress screen. Initially, the screen profile is developed on the basis of the failure modes detected in HALT. The test engineer then runs a sample of products through multiple passes of the proposed screen. If a failure occurs, stress levels are reduced to calibrate the safety of the screen. Once the product is released, field performance is monitored to ensure that all observed failure mechanisms are accounted for in HASS. Early life failures in the field may indicate a need to refine the stress screen in HASS. Field returns that are diagnosed as “no problem found” (NPF) are often tested under the HASS screen to confirm that it can detect intermittent design problems. Once the manufacturing process is stabilized and the defect rate in HASS is under control, HASS is replaced by a highly accelerated stress audit (HASA) program. HASA implements the same screens as HALT, but only on a sample basis. HASA results are regularly analyzed to see if there is a need to change the sample size of the audit.

An ongoing reliability test (ORT) is used to monitor a product on a continual basis and obtain a reliability estimate by testing samples of the production units. ORT is similar to the RQT except that the RQT is performed only once prior to release of the product,

whereas the ORT is an ongoing test, performed by rotating in samples from the manufacturing line. As an ongoing test, ORT can detect changes in failure modes and product reliability caused by design or manufacturing process changes.

Warranty Analysis

Warranty costs have an enormous impact on business profitability. Warranties are driven by service support, material, failure analysis, inventory, and logistics costs. They incur expenses across multiple organizations within a company. Warranty analysis (see **Warranty Analysis**) is used to identify the most likely warranty events and develop strategies for cost reduction.

During product development, warranty analysis is performed by using the reliability prediction results to set up a cost model that examines the trade-offs between design, service, and warranty policies. Reliability prediction identifies the most likely warranty events, and the projected warranty costs are calculated for each identified event. As an alternative to reliability prediction, some manufacturers may choose to use a current product in the same market segment with similar design and function as an internal benchmark. The cost analysis can identify the key contributors to the warranty budget, such as product failure rates and support processes. The results are used to evaluate alternative designs and support strategies. They also establish a preliminary requirement for logistics support on issues such as spare parts inventory and field personnel.

Later in the design phase, prototype test results are utilized to update the reliability estimates and refine the cost model. Cost analysis can confirm whether sufficient testing has been carried out. HALT results are used at this stage to develop diagnostic tools so that the support organization may service the warranty events more efficiently in order to reduce the overall warranty cost.

During the product design and development stage, warranty analysis is built around activities for forecasting liabilities associated with future warranty claims. The actual warranty costs are found only after product launch, by tracking shipments and warranty return data. Most manufacturers have a field failure tracking system that documents the returned material by returned date, part number, manufacture date and

location, and failure diagnostics for a given product. This data is used to evaluate the product reliability and warranty cost exposure over the life of the warranty. After the product is released to the field, a sustaining engineering organization is responsible for the update of the product design, introduction of new materials to be used on the product, and the revision of product, process, and test specifications. Sustaining engineering analyzes the field failure data regularly to discover changes in failure modes over time and revise the HASS screen accordingly. In a product recall situation, the field failure data are examined to see if the problem can be isolated to a specific manufacture location and a narrow range of manufacture date. As a result, the company may be able to reduce a costly large-scale recall to a smaller isolated replacement program.

References

- [1] Nelson, B. (1990). *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis*, John Wiley & Sons.
- [2] Meeker, W. & Escobar, L. (1993). A review of recent research and current issues in accelerated testing, *International Statistical Review* **61**, 147–168.
- [3] NIST *NIST/SEMATECH e-Handbook of Statistical Methods*, <http://www.itl.nist.gov/div898/handbook/apr/section1/apr15.htm> (accessed 2007).
- [4] Crow, L. (1974). Reliability analysis for complex, repairable systems, in *Reliability and Biometry*, F. Proschan & R.J. Serfling, eds, SIAM, Philadelphia, pp. 379–410.
- [5] Finkelstein, J. (1976). Confidence bounds on the parameters of the Weibull process, *Technometrics* **18**, 115–117.
- [6] Jin, T., Liao, H., Xiong, Z. & Sung, C. (2006). Computerized repairable inventory management with reliability growth and system installations increase, *Conference of Automation Science Engineering (CASE 2006)*, Oct. 8–9, 2006, Shanghai.
- [7] Sun, A., Kee, E., Yu, W., Popova, E., Grantom, R. & Richards, D. (2005). Application of Crow-AMSAA analysis to nuclear power plant equipment performance, *13th International Conference on Nuclear Engineering*, May 16–20, 2005, Beijing.

Related Articles

Accelerated Life Testing

Design of Reliability Tests

Reliability Data

Stress Screening

WAI CHAN

Professional Organizations in Risk Management

Defining Financial Risk Management

Financial risk management (*see Solvency; Extreme Value Theory in Finance*) can be as broadly or as narrowly defined as the professional organization wants it to be. There is no globally accepted definition. It changes depending on who you are speaking with, where in the world the conversation takes place, what day it is during the calendar year, and the interests of those engaged in discussion.

The Global Association of Risk Professionals (GARP) defines financial risk management to include credit, market, and operational risk theories, concepts, and practices in addition to the quantitative aspects associated with each area. Each of these areas of financial risk management can cover specific subjects that are so broad in scope or so complex in nature that many financial risk-management practitioners specialize in them alone, or even a subsection within that particular subject area.

Included in GARP's financial risk definition would be activities engaged in by a wide variety of organizations such as banks, securities firms, asset management firms, pension funds, insurance companies, and in many instances supervisors or regulators. Each of these organizations individually will have unique risk-management issues to address and their internal definition of financial risk management will take on broadly different overtones by country, region, industry, company, department, or product line, or a combination of any or all of these geographical or corporate characteristics. These varied approaches to financial risk management and the sheer complexity of attempting to respond to the needs of a widely dispersed clientele present material challenges to any financial services professional organization.

The Role of a Professional Organization

As with the definition of financial risk management, the role of a professional organization within the field of financial risk management will vary

widely. Professional organizations obviously represent different membership constituencies, an important and major distinguishing factor among professional organizations.

Certain organizations include only corporations as members, while others will have only individuals who have joined personally, with or without the support of their employers. Some will combine the two membership types but understandably end up bifurcating services and service levels depending on who is paying the bills.

There are pros and cons associated with each type of membership organization. For example, a corporate-only membership may consist of a sterling list of members with each paying a yearly fee based on size. In those types of professional organizations, the corporate members will generally view their membership fee as an initial payment for programs offered by the association, expecting to be charged a below market rate, or no charge, for programs offered by the organization. Many corporate members will also expect to be represented on the organization's board and as members of any of the association's working groups. All of these expectations can be very positive for the association and its members. The primary issue to be concerned about with this structure is whether the association can remain nimble, being able to respond to issues and implement new initiatives in a timely manner without having to deal with a large committee and approval process.

Organizations such as GARP have individuals as their core membership. These individuals join on their own or with their employer's sponsorship. The advantage of this type of organizational structure is that the association can be quickly responsive to changes in the industry as the decision-making process is much quicker. In addition, because the individuals who have joined have done so on their own, their level of interest is generally exceptional, allowing for a large pool of individuals that would be available to assist in moving the association's programs forward.

Professional Organization Services

Educational services are generally the primary focus of most professional organizations. In the financial services industry this role is especially important given the number of new initiatives regularly being

put forth by regulators and the dynamic nature of the industry and its seemingly insatiable desire and ability to invent new products and services.

Educational services can take the form of offering professional certification programs, courses and workshops, conferences and events, and take on the role of educating regulators and assisting in the development of legislative or regulatory initiatives.

Professional certification programs are growing dramatically. This is the result of the increasing complexity of the financial services industry and the need to ensure that those who are working in certain subject areas can do so competently and with a knowledge base that is up-to-date and represents globally accepted approaches.

The rapid globalization of the financial services industry has enhanced the need to ensure that those in the financial services industry are speaking in terms that are common across borders and which represent accepted industry practices. A failure to do so will quickly result in misunderstandings or errors that will rapidly lead to financial and reputation risk.

Lobbying is another form of educational service. Not all organizations engage in lobbying activities, and for those it is a specific part of their mission.

The concept of lobbying is viewed differently in different parts of the world. In the United States it commonly incorporates the hiring of individuals whose primary purpose is to directly influence the votes of government legislators and other regulators as they look to decide specific issues or issues. In other regions it may include the provision of advice to legislators or regulators but not with the objective of forming the wording for a specific initiative that would benefit a specific group or organization.

GARP includes with its educational activities the role of being an industry voice for reviewing and providing insight into the effects of legislative or regulatory initiatives on financial institutions. This role is one of education and dialog encouragement among and between regulatory bodies and industry practitioners rather than the wielding of direct influence in specifically determining the wording for or effect of a rule or regulation. This is done through the formation of industry working groups whose purpose is to review issues of importance to the financial services industry as a whole and through written and oral discussion recommend approaches that would be in the best interests of the financial services industry.

Other organizations may define their role as it relates to lobbying differently. For example, the United States based Securities Industry and Financial Markets Association is a very effective lobbying organization that works on behalf of its institutional membership to influence regulatory initiatives. This is an important and material part of its mission, along with certain other educational activities.

Lobbying on behalf of an organization's membership is a form of financial risk management. For example, an organization seeking to establish standards relating to structured financial products or the adoption of approaches to influence capital allocation inherently includes in its analysis a number of financial risk-management concepts. Lobbying is intended to influence the development of rules and regulations by which every organization must operate, and as such it has direct risk-management implications.

Networking is another important role played by a professional risk-management organization. This role is simply one of providing a forum for risk professionals either by city, region, or country to share ideas and best practices. These networking professional organizations play an essential role for the sharing of career-related information and for offering professional development opportunities among their members.

Still, other financial associations have adopted a hybrid model, taking on the role of an educational institution and networking forum, a lobbying and educational group, or an organization that seeks to establish best practices through the issuance of "white papers" while not directly lobbying on behalf of its members. These latter organizations attempt to influence the thinking of rule-making bodies through the submission of formal letters expressing opinions to government regulatory bodies or publishing papers approved by the majority of its members.

Each membership model will have a direct effect on the organization's business approach and the organization's governance. In turn that will determine the latitude it may have in addressing new markets, issues, and responding to member requests.

There does not appear to be any one ideal model in terms of networking, lobbying, education, or a combination thereof, nor is there one prevailing membership model. Most organizational business models have evolved over time with regard to their scope of activities and membership structure. However, the rapid globalization of the marketplace is starting to

take a toll on some of these historical models. It is forcing many professional organizations to redefine their roles and services. But, unfortunately, it is leaving many with limited options for a variety of reasons.

Many organizations were slow to respond to the changing marketplace, allowing others to expand their services and/or leapfrog over them and take the spot as the global contact for that particular profession. There is, in most cases, little need to have two or more organizations providing the same or similar services to the same group of professionals.

In other cases corporate charters, which were viable when first put into place, now restrict their expansion outside certain geographic boundaries. For others, the need on the part of their membership for services that reflect global references and understanding cannot be fulfilled by a purely regional or local organization.

Professional Organizations as a Business

As noted above, the individual roles of professional organizations can be varied. They can be broad and cater to a global constituency or they can be narrow by responding to only a local market niche.

In a world where new financial products are being developed daily, where financial issues are becoming more complex by the day, and where it is virtually impossible to move away from having to consider issues based in a broader global context, each organization must be run as a business. This is a point missed by many professional organizations and their membership. Professional organizations are no different from any other business entity other than the fact that they, in most cases, do not have formal equity holders – their shareholders are their members.

Professional organizations must provide relevant services just as any other business entity. Whether the organization is supported by a dues paying individual or corporate membership, or its revenues are derived from the providing of a value-added service where the failure of its membership and/or the public to purchase the service will directly affect the organization's revenues, any professional organization must carve out its market niche and provide exceptional service. A failure to do so, as with any other business organization, will result in a diminished reputation and a diminution of the

financial resources necessary to run the organization and provide new and/or improved services.

Providing relevant services is now becoming a material issue for many organizations. Historically professional organizations have been able to service their membership by focusing on the markets that make up their region or local community, rightly gearing their product toward what would be of benefit to their direct membership. Now, however, their membership is being impacted by the rapid globalization of the financial marketplace and the instantaneous availability of information made possible by, among other sources, the Internet.

Global issues are now finding their way into virtually any financial risk-management conversation. The result is that many professional organizations' members are seeking information on how to compete and/or respond to local or regional issues by using globally accepted approaches. Local professional organizations are finding it difficult or impossible to respond. They have to consider expanding their mandates beyond their local coverage and/or choosing a subject area they can respond to given the broad financial risk marketplace. Unfortunately many local professional organizations will not have the legal ability to expand their activities, personnel well versed in international issues, or the financial ability to develop globally based programs.

The above issues are causing professional financial risk organizations to think of developing alliances with other professional organizations that will complement their business model and respond to their members' needs. However, these alliances are fraught with their own issues. Some issues are simply financial related while others are directly affected by the need to manage these relationships and the time that it requires. With many professional organizations working with limited budgets, the challenges presented are daunting.

This article is not intended to suggest that the time has come to minimize "local" professional organizations. What is being suggested is that it will become increasingly difficult for a professional organization without a global mandate to ultimately develop one. There are barriers to entry for professional organizations as well as for corporations. There are few compelling reasons for one professional organization to compete with another professional organization, especially if the organization that has the global experience is doing a good job. In fact, there are

material reasons for this not to occur given the general not-for-profit or charitable business models followed by most professional organizations.

The Role of the Professional Organization

The role of the professional risk-management organization is evolving, as is the profession of financial risk management. However, it is clear that a professional risk-management organization should possess a number of attributes.

Any professional organization should be independent. Independence is more than just legal separation from various parties. It means that it should not be influenced directly or indirectly by third parties who are attempting to market their products or promote their own agendas. Products and/or services offered as a means to ensure an individual has gained a certain professional standing should not be the subject of outright paid endorsements by third parties, or if they are the relationship(s) must be fully disclosed. Anything less should be considered a conflict of interest.

An organization's committee structure and board should not consist of members who are "interested parties", meaning whose vote on a decision could result in positive benefits accruing to them or their firm, unless there is full disclosure either directly or indirectly through the rationale behind the organization's formation.

Corporate governance is just as important to professional organizations as it would be to any other corporate entity. There are strong arguments that a sound corporate governance structure is even more important to a professional organization as by its nature it is supposed to represent leadership in its chosen field. The governance structure of all professional organizations should be open, transparent, and subject to review by its board and/or executive committee, with its membership kept regularly informed of material financial and structural issues.

Conflicts of interest must be carefully considered and monitored. For example, organizations offering professional designations should be careful that the certifying activity is not compromised by the offer of a product or service that would directly or indirectly benefit individuals who are charged with setting the standards for that designation. In many instances, apparent conflicts of interest can be just as damaging as actual conflicts. Professional organizations should

be held to a higher standard as they are effectively role models to their members and the industry they represent.

A professional organization should seek excellence in all it does. It exists to serve its membership. Its members are relying on the organization as a forum to share ideas and learn about new products and/or approaches to issues. In most cases these members rely on the organization to be the expert in its chosen field, presenting well-developed and unbiased views of issues and concepts.

An industry representative professional organization should take a practice-orientated approach to issues and concepts. It needs to develop and share information that will assist its members in doing their everyday job better. Professional organizations in risk management play a major role in educating their members. This education must take into consideration the fact that its members are seeking to not only keep up with advances but also that they are working full time in their roles. Presenting succinct programs that are practical and thought provoking or broadening in scope should be high on the organization's agenda.

Bridging the gap between academia and the everyday working world is an agenda item that should also rate high in most organizations. Academics through their research and teaching activities offer to the professional practitioner community a valuable resource of new ideas and approaches, concepts that, for the most part, risk management or other professional practitioners will not have the time to develop owing to their own daily activities. This academic/practitioner sharing can benefit both groups of professionals and should be encouraged through the professional association's activities.

Conclusion

Professional organizations in risk management are playing an increasingly important role in educating professionals and industry regulators. They also serve as a bridge between geographic locations, allowing for a sharing of resources and programs to assist financial service practitioners having a need to gain knowledge and understanding of how markets work in different parts of the world. The networking opportunities offered to their membership are just as important as these will allow for an informal and free sharing of ideas that should help

any practitioner in his or her daily activities. Most importantly, professional organizations must operate as any other business developing programs that are of the highest quality, innovative, and informative. It is these organizations that will independently provide

the foundation for the establishment of a culture of risk awareness throughout the world.

RICHARD APOSTOLIK

Protection of Infrastructure

Critical infrastructure (*see* **Managing Infrastructure Reliability, Safety, and Security; Game Theoretic Methods**) consists of physical and information technology facilities, networks, services, and assets that are essential to the well-being, operations, and continuity of operation of a country. They include, but are not limited to, agriculture and food, public health, telecommunications, banking and finance, transportation, water and power systems, emergency services, government, waste removal, and the industrial base.

Risk assessment for critical infrastructure has two components: assessment of the likelihood of particular disruptions from either natural or man-made disasters, and the evaluation of the consequence from that disruption. There are many metrics by which decision makers assess the consequences arising from the failure of a critical infrastructure: (a) human health and safety (e.g., in terms of the number of lives lost or placed at risk), (b) economic (e.g., in terms of the recovery cost, or the cost of lost business, or both), and (c) environmental (e.g., cleanup costs). There are also vaguer metrics, relating to sociopolitical damages and national security perception.

Although the consequence assessment can be cast as in a multiattribute utility function (*see* **Multiattribute Modeling; Multiattribute Value Functions**) framework, this is not usually the primary concern. Policy makers have good intuition about outcomes that are highly problematic, and delicate trade-offs among these do not drive their decisions. The main problem concerns the probability of such outcomes, and that is the focus in this article: the methodology of risk assessment for critical infrastructure.

Often one can rely on historical data to assess the likelihood of occurrence of natural disasters, such as earthquakes, hurricanes, floods, or failure in the power grid (assuming that the secular trends in these are slow or negligible, which is sometimes questionable). In contrast, assessing the likelihood of a terrorist attack or a complex engineering failure is harder because only a few incidences have occurred and because the causal factors are poorly understood. Often, risk analysts distinguish “aleatory” probabilities from “epistemic” probabilities. The former refer to events that are well

described by laws of probability, such as the actuarial rate of automobile accidents. The latter refer to events that have probability mechanisms that are poorly understood, such as the probability that a specific driver will have an accident (this is clearly affected by many unknowns, such as the experience, sobriety, and amount of driving that person will do).

For epistemic applications, it is often only possible to rank order the relative likelihood of potential scenarios, as informed by such tools as scientific models, psychological reasoning, and game theory. All three contribute to estimating the risk of a bioterrorist attack; medical science gives guidance on how easily a specific pathogen could be used, psychological reasoning indicates what kinds of people might be likely to mount such an attack, and game theory describes the interplay of countermeasures and attack strategies in terms of reducing the level of risk.

Understanding the likelihood and consequence of disruptions are essential for rational allocation of finite resources to protect high value infrastructure and to distribute assets to help speed recovery after an event. In terrorist situations, making these estimates is the responsibility of the Office of Infrastructure Protection (OIP) of the Department of Homeland Security (DHS). For nonterrorist threats, responsibility is typically shared among various federal and state agencies, with the Federal Emergency Management Agency (FEMA) having a special role in disaster response.

Modeling the Risk of Rare Events

There is a large and straightforward literature on modeling the occurrence of small to medium natural disasters using standard extreme value theory (*see* **Large Insurance Losses Distributions; Extreme Value Theory in Finance; Mathematics of Risk and Reliability: A Select History**). The approach relies upon fitting the tail behavior of distributions (*cf.* [1], for a discussion in the context of insurance). There are a number of standard procedures for specific situations, largely growing from the work of Pickands [2] on extreme order statistics. Of course, the entire field is strongly influenced by the work of Gumbel [3] and Fisher and Tippett [4], who established that there are only three possible asymptotic distributions for the maxima or minima of a set of independent random variables.

However, the risk of extreme natural disasters (sometimes called *low-probability, high-consequence events*) are harder to model. The same holds true for failures of complex technological systems that should not fail, such as a nuclear power plant and/or both space shuttle disasters. The Boxing Day tsunami, the Three Miles Island nuclear accident, and the failures of the Columbia and Challenger, all support the view that the statistical and scientific models used at that time grossly underestimated the actual risk.

To do risk assessment for low-probability, high-consequence events, most analysts attempt to combine expert judgment with a decomposition of the problem into constituent elements. Different sets of experts may be consulted about the different elements (which can create problems, since there is a heuristic in the field that most failures occur at the interfaces between the elements in a complex system). The hope is that the judgment of domain experts can partially overcome the lack of historical data. However, this hope often founders because experts tend to be overconfident of their opinions (and thus understate the true uncertainty) because elicitation techniques (see **Imprecise Reliability; Bayesian Statistics in Quantitative Risk Assessment; Expert Judgment; Multiattribute Modeling**) are technically difficult to implement correctly (cf. [5]), and because the problem of combining expert judgments is hard (the best answer may come from a single expert in the pool who has correctly apprehended something that the others overlooked). So use of expert judgment is difficult, but in general there is no practical alternative.

Similarly, it can be problematic to divide a complex system into its functionally distinct elements. A common strategy decomposes the system into its subsystems and components tied together by a fault tree (see **Systems Reliability; Fault Detection and Diagnosis; Probabilistic Risk Assessment**) (cf. [6]), or more generally, by a Bayes net (cf. [7]). In the absence of pertinent historical information, the Bayes net is populated using expert knowledge. In this context, the problem of combining expert knowledge has prompted the exploration of several alternative uncertainty quantification paradigms, including fuzzy logic (see **Early Warning Systems (EWSs) for Predicting Financial Crisis**), Dempster–Shaffer belief functions, and probability intervals (the authors of this article deliberately give no citations to this literature, since we do not endorse these approaches for any application – the methods do not provide

the coherent framework for uncertainty quantification that is provided by probability theory). A more appropriate strategy for quantifying the uncertainty in the experts’ judgments of the conditional probabilities in a Bayesian network is to model these probabilities as random variables (but this is not simple to implement). The advantage of this approach is that inference relies on the rules of probability while, at the same time, one can accommodate the uncertainty inherent from the expert elicitation.

The problem of determining such things as the risk of a terrorist event is even harder than predicting the failure of a complex engineered system because of the need to model political forces (causes of origin, politics, and social-cultural basis) and human behavior (terrorist organizational structure, activity, and terrorism effects). The risk assessment is further compounded by the evolving nature of terrorism, and the fact that terrorist groups range from lone wolves to rogue nations. As a result, any risk assessment in this area relies heavily upon judgments from many kinds of social scientists.

The Terrorism Risk Insurance Act (TRIA) of 2002 created a federal “backstop” for insurance claims related to acts of terrorism. The TRIA was intended as a temporary measure to allow time for the insurance industry to develop their own solutions and market new products to insure against acts of terrorism. The renewal of the TRIA act in 2005 is an indication of the difficulty the insurance companies have had in evaluating the risk associated with terrorism.

Infrastructure Interdependencies

Historically, individual infrastructures have been modeled as physically and logically separate systems, and so was the modeling of the impact of disruptions to each infrastructure. But in reality, as a result of advances in information technology and the necessity of improved efficiency, the various infrastructures have become increasingly automated and interconnected, linked by people, energy, and information. For example, a disruption in the transportation sector impacts all other sectors by limiting the ability of individuals to get to work. Or a disruption of the power grid can impact the telecommunication sector, which in turn adversely affects the banking sector. The spatial–temporal scale of the interaction depends on the type of interconnection (human, energy, and information) between the infrastructures.

These interdependencies also are responsible for new vulnerabilities, ranging from equipment failure, human error, weather and other natural causes, to physical and cyberattacks that were not previously taken into account. Furthermore, such interdependencies have the potential of generating feedback loops that can exacerbate the impact of the original disruption. As a result, modern risk assessment of critical infrastructures must model these interdependencies to effectively assess the consequence of disruption.

These models for interdependency simulate the dynamics of individual infrastructures and couple separate infrastructures to each other according to known connections. This tool enables analysts to model the cascading effect associated with a particular failure. For example, repairing damage to the electric power grid in a city requires transportation to failure sites and delivery of parts, fuel for repair vehicles, telecommunications for problem diagnosis and coordination of repairs, and the availability of labor. The repair itself involves diagnosis, ordering parts, dispatching crews, and performing repairs. The electric power grid responds to the initial damage and to the completion of repairs with changes in its operating characteristics, such as the number of megawatts that can be distributed to customers. The dynamic within and between infrastructure sectors are modeled by differential equations, discrete events, and codified rules of operation, and may involve as many as 5000 variables, many of which are output metrics estimating the human health, economic, or environmental effects of disturbances to the infrastructures.

Agent-based modeling and simulation (ABMS) is a tool of increasing interest for the analysis of large-scale complex systems that consist of a large number of subsystems (agents) which interact with each other and with their environment (*cf.* [8]). Sociotechnical simulations form a particular set of simulation tools that models human interaction with one another and their surroundings. The overall dynamic behavior of the resulting complex system typically emerges in a nonobvious manner from the simpler interactions among subsystems of the system. Many users feel these models are more robust and transparent than the systems of differential equations that are also popular.

As an example of a sociotechnical simulation, consider TRANSIMS, a transportation model (*see Homeland Security and Transportation Risk*) developed at the Los Alamos National Laboratory. For that simulation, one starts out by creating a

synthetic population of households whose demographic characteristics match the census data. Using data from the National Household Travel Survey [9], one assigns a set of activities to each household, together with the location for the assigned activities. The agents (humans) interact with one another when traveling from one location to another on the road network. Such simulation has been shown to give realistic traffic patterns and has been used to predict the impact of adding public transportation or new highways to an existing road network.

In the context of infrastructure protection, TRANSIMS has been used to predict the effect of road and bridge closures. Since other sectors require workers (e.g., the health sector needs doctors and nurses), variations of this tool can model the interdependence between the transportation infrastructure and other infrastructures. The strength of sociotechnical models is that they can describe the demand for resources imposed by the various infrastructure sectors. For example, TRANSIMS allows the tracking of all the cell phones in a city, thereby representing realistic spatial-temporal demand for mobile telecommunications.

Recent work on infrastructure interdependence has identified two new areas that are attracting attention. Both concern optimal investment strategies for protecting infrastructure.

Consider a decision maker who wants to protect a city and has a limited budget to do so. Some investments, such as hardening the city's levee system, will protect only against hurricanes and floods. Other investments, say in hospitals and emergency housing, provide a smaller degree of protection against a larger set of threats, not only hurricanes and floods, but also fires, explosions, and pandemic flu. Managers have to balance these trade-offs, but often the decision-making structures prevent effective investment. Decisions about police department funding are handled by people different from those who harden levees or fund hospitals [10].

The second new issue concerns how competing corporations that control critical infrastructure should invest in counterterrorism protection. For example, air traffic is critical to the national economy, and it is a shared responsibility among different firms. Each firm needs to decide how much it will invest in security operations. One strategy says that a company should spend just enough to be the second most vulnerable target – that way terrorists will exploit the

most vulnerable and the company has not overspent on protection. Another strategy is to encourage all of one's competitors to invest in security, and claim to do so too, but then to secretly defect, thereby getting a free ride from the expenditures of others. This line of work points up the need for public sector regulation to ensure uniform levels of protection, as has, in fact, happened with airline security [11].

Adversarial Risk

Adversarial risk is an emerging area of risk research that involves game theory (*see Risk Measures and Economic Capital for (Re)insurers; Group Decision; Game Theoretic Methods*). It treats situations in which threats are posed by intelligent opponents, who choose actions nonrandomly to maximize damage to infrastructure. The classic example is terrorism, but the issues also arise in corporate competition and cybersecurity.

In conjunction with game theory, one also needs to use risk analysis. Classical game theory assumes that the costs to attack and defend a particular target are known, but in fact these are often uncertain. Statistical risk analysis supports game theory by determining the joint distribution of the unknown cost matrix used in the game-theoretic formulation.

In many situations, particularly those in which the game involves the possibility of repeated play, it may be also appropriate to consider some optimization procedure, such as dynamic programming (*see Optimal Stopping and Dynamic Programming*), to determine how the opponents allocate their resources for attack and defense to achieve acceptable levels of risk and reward. For example, some terrorist organizations plan multiple attacks staggered over multiple years (though other terrorists, such as Timothy McVeigh, invest in a single, all-or-nothing attack). Similarly, competing corporations rarely bet the firm in an all-out attack; rather they use more cautious strategies that ensure they will still have operating capital and market share in subsequent years of potential competition.

The combination of these three techniques is difficult, and in many applications decision makers use various shortcuts. At the Department of Homeland Security, people often use red-teaming to inform decisions; here one asks a group of smart people with diverse skills to plan attacks as if they were specific groups of terrorists, and then analysts decide

what kinds of target hardening are most protective against threats developed by the red team. Corporate approaches to adversarial risk seem even less formal.

The remainder of this section discusses how the three components of the problem (game theory, risk analysis, and optimization) can be combined. At the same time, we emphasize that in practical applications one must often make gross approximations.

Game Theory

Game theory is a mature field of mathematics; Myerson [12] offers a general overview of the technical aspects. But the main idea, from the standpoint of an infrastructure protection example, is simple. Suppose that terrorists are considering an attempt to bomb a subway station. In broad strokes, they have three options: do not make the attempt, try something cheap and simple, or invest in a major attack. Correspondingly, again in broad strokes, counterterrorists can choose to ignore the threat, or provide inexpensive hardening, or to invest in significant protection.

From a game theory standpoint, this situation can be represented by a cost matrix (see Table 1).

The entries in each cell of the matrix represent the costs for a particular attack/defense combination (this assumes a zero-sum game, which is approximately reasonable in this example, and our convention is that the entries are interpreted as the cost to the defender).

If one knows what the values of the C_{ij} terms are, then there are several approaches to finding the optimal attack strategy and the optimal defense strategy. The classical solution is to use the minimax theorem; here the terrorist looks for the column that maximizes the minimum cost, and the counterterrorist looks for the row that minimizes the maximum cost. If this row and column intersect in the cell that achieves both criteria, then the solution to the game is unique and the best plan for both parties is determined. If not, then a related but more complex

Table 1 Illustration of a payoff table in counterdefense for a normal form two-person zero-sum game with random payoffs

	No attack	Minor attack	Major attack
No defense	C_{11}	C_{12}	C_{13}
Weak defense	C_{21}	C_{22}	C_{23}
Strong defense	C_{31}	C_{32}	C_{33}

“mixed” strategy, based on randomization and the solution of a system of linear equations, is optimal.

There are alternative approaches. A Bayesian version of game theory (*see* **Near-Miss Management: A Participative Approach to Improving System Reliability**) developed by Kadane and Larkey [13] uses a distribution over the actions of one’s opponents to reflect uncertainty about his or her choice. The distribution is subjective, but presumably incorporates minimax calculation, knowledge of psychology, and a truncated “I know that he knows that I know that he knows . . .” kind of argument. Given this subjective distribution, the player chooses the action that minimizes his or her expected loss.

Many people criticize game theory as an unreasonable model for human behavior. In particular, real decision makers often make choices using intuition rather than calculation. Kahneman and Tversky [14] describe many experiments that show that decision makers are often technically irrational, in the sense that they do not select the optimal choices in a particular situation. However, this does not mean that game theory ought not guide real-life solutions, and if one’s opponent makes choices that are suboptimal in terms of game theory, then optimal play on one’s own part is an advantage.

Risk Analysis

Traditional game theory assumes that the entries C_{ij} are known precisely. In fact, for most situations, such as the motivating example of a subway bombing, one has only a vague understanding of the costs.

For example, suppose the terrorists decided upon a weak attack (perhaps something similar to the London Tube bombing on July 7, 2005, undertaken by a small number of amateurs) and the counterterrorists decided upon a weak defense (again, perhaps similar to the routine police surveillance that was in place before the London tube bombing). Then it could happen that the terrorist plot is entirely thwarted, or that the terrorists had to detonate their bomb prematurely and caused minimal damage, or that the terrorists were lucky and the subway was crowded and the damage was devastating. These outcomes are random events whose probabilities can be assessed through classical risk analysis techniques.

To estimate the random costs to Great Britain in this scenario is not easy, but it is susceptible to the kinds of nonadversarial risk analysis techniques

described previously in this article (because we have conditioned on the adversarial component of the problem in fixing the kind of attack and the kind of defense). Components of this analysis would include the following:

- finding the distribution of the number of people at risk by consulting transit ridership data from the previous year;
- assessing probabilities from explosives experts about the likely size of a blast and the number of people who would be killed or injured as a function of the timing and placement of the explosive;
- assessing the probability of preemptive discovery from police records on similar previous attempts, for example, by Sinn Féin or from suicide bombers in Israel;
- consulting with civil engineers to estimate the distribution of the repair cost under different scenarios of destruction;
- consulting with health care economists to determine the medical costs of caring for the injured;
- consulting with social psychologists and market economists to estimate how many people would avoid using the subway, and the financial impact this would have.

This is not an exhaustive list, but it indicates the range of expertise that would be needed and the kinds of data sources that might support the expert judgments.

From these inputs, the risk analyst would develop an equation that represents the random cost as a mathematical function of a number of random variables, where the distribution of each random variable is determined by elicitation of expert opinion or previous data (as with the transit ridership records). Elicitation of expert opinion is not always reliable (*cf.* [5]), but there are no real alternatives when attempting to do risk analysis in situations for which there is little historical data. Fortunately, it is often sufficient to have an equation that is order-of-magnitude correct; the decision that is ultimately obtained from this kind of analysis is often insensitive to minor inaccuracies and approximations. Usually it is more important that the equation be transparent and reasonable than that every single little detail be micromodeled.

To put things on a common footing, it is probably unavoidable to monetize very different quantities, such as the value of a life or the loss in quality of life

associated with fear to travel by subway. Monetizing the value of life is always controversial; some people use expected income or expected productivity, whereas others take account of noneconomic value.

This kind of analysis would be repeated for each cell in the game theory cost matrix. Many of the random variables would be the same from cell to cell (e.g., the transit ridership distribution or the medical cost of treating an injured person) but some (such as the chance of success or the number of bombs planted) might be very different, depending on the attack–defense pair that define the cell.

The risk analyst views the cost matrix as a random matrix with the joint distribution of the entries determined by the mathematical functions of the random variables involved in estimating the component costs. This enables the analyst to use repeated sampling to determine the probability that a particular defense is the one that minimizes the cost. Specifically, the analyst draws many realizations of the random cost matrix, and for each matrix determines the game-theoretic solution (either by applying the minimax criterion, or the minimum expected loss criterion from the Kadane–Larkey approach, or by some other procedure). Thus the analyst can estimate the proportion of matrices for which a particular choice of defense is optimal, and advise decision makers accordingly. (See [15], for a detailed example in the context of the decision about whether to inoculate the general US population against smallpox.)

For some applications, it can happen that the outcome is very sensitive to the assumptions made in the risk analysis or the game-theoretic solution technique that is chosen. However, often there is little sensitivity, typically because a few cells with large costs dominate the decision. Additionally, this methodology allows decision makers to explore “what-if” scenarios, to see how changes in assumptions or resourcing affect the outcome.

Portfolio Analysis

The preceding analysis of adversarial risk is incomplete because it does not capture resource constraints on the adversaries and long-range planning objectives that shape their investments in attacks and defenses. The branch of mathematics that addresses these elements is *portfolio theory*.

It is reasonable to imagine that some kinds of terrorists expect to operate for the foreseeable future. They would need be unlikely to sink all their resources into a single attack, and they would want to stage their different attacks so as to ensure a steady record of high-profile successes, in order to encourage recruitment and donations. On the other hand, different kinds of terrorists might gamble all their resources on one effort (if history is a guide, these tend to be small groups of amateurs). And, from the standpoint of the counterterrorists, nations cannot bankrupt their future to purchase a defense, and they will want to spread security investments among different programs so as to maximize the net protection. (The same issues apply to corporate competition; businesses need to plan their attacks and defenses within fixed budgets and with long-term planning horizons.)

From this perspective, one can describe the problem in terms of investment in a portfolio (*see Role of Alternative Assets in Portfolio Construction; Risk in Credit Granting and Lending Decisions: Credit Scoring*) of attacks (or defenses). A terrorist decision maker must allocate a certain amount of money, time and people among truck bombs, nuclear devices, biological pathogens, and so forth; similarly, the counterterrorists can spread their investments among border security, airline passenger screening, intelligence operations, cargo inspection, and so forth. And, as the game is played repeatedly over time, some of these investments may be redirected, just as one can reshuffle money among one’s stock market portfolio, or buy new stocks.

Portfolio analysis is hard and relies upon dynamic programming. The goal is to maximize return under constraints on the amount of risk one is willing to accept. As one observes market history, one adaptively changes the portfolio to best achieve one’s goals in a stochastic environment. The analogy to the decisions of terrorists and counterterrorists, or competing firms, is immediate.

Currently, no formal portfolio analysis has been undertaken in the context of infrastructure protection. But informal analyses appear in many places; the United States has made decisions about which kinds of defenses will be supported, and the amount of risk it is willing to tolerate. Nonetheless, one imagines that a rigorous analysis might find more cost-effective investments.

References

- [1] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, Berlin.
- [2] Pickands, J. (1975). Statistical inference using extreme order statistics, *Annals of Statistics* **3**, 119–131.
- [3] Gumbel, E. (1935). Les valeurs extremes des distributions statistiques, *Annals de l'Institut de Henri Poincaré* **5**, 115–158.
- [4] Fisher, R. & Tippett, L. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society* **24**, 180–190.
- [5] Wolfson, L. (1999). Elicitation, in *Encyclopedia of Statistical Sciences*, Update Volume 3, S. Kotz, C. Read & D. Banks, eds, John Wiley & Sons, New York, pp. 199–206.
- [6] Vesely, W. (1987). *Fault Tree Handbook*, Nuclear Regulatory Commission.
- [7] Ben-Gal, I. (2007). Bayesian networks, in *Encyclopedia of Statistics in Quality and Reliability*, F. Ruggeri, F. Faltin & R. Kenett, eds, John Wiley & Sons, New York.
- [8] Deguchi, H. (2004). *Economics as an Agent-Based Complex System*, Springer, New York.
- [9] U.S. Department of Transportation, Bureau of Transportation Statistics (2002). *National Household Travel Survey* at <http://www.bts.gov/programs/national-household-travel-survey>.
- [10] Bier, V., Nagaraj, A. & Abhichandani, V. (2005). Optimal allocation of resources for defense of simple series and parallel systems from determined adversaries, *Reliability Engineering and System Safety* **87**, 313–323.
- [11] Heal, G. & Kunreuther, H. (2005). IDS models of airline security, *Journal of Conflict Resolution* **49**, 210–217.
- [12] Myerson, R. (1991). *Game Theory: Analysis of Conflict*, Harvard University Press, Cambridge.
- [13] Kadane, J. & Larkey, P. (1982). Subjective probability and the theory of games, *Management Science* **28**, 113–120.
- [14] Kahneman, D. & Tversky, A. (1972). Subjective probability: a judgment of representativeness, *Cognitive Psychology* **3**, 430–454.
- [15] Banks, D. & Anderson, S. (2006). Combining game theory and risk analysis in counterterrorism: a smallpox example, in *Statistical Methods in Counterterrorism*, A. Wilson, G. Wilson & D. Olwell, eds, Springer, New York, pp. 9–22.

DAVID L. BANKS AND NICHOLAS
HENGARTNER

Public Health Surveillance

Public health surveillance (see **Environmental Exposure Monitoring; Digital Governance, Hotspot Detection, and Homeland Security**) has been described as the “ongoing, systematic collection, analysis, and interpretation of health-related data essential to the planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those responsible for prevention and control” [1]. Recently, terms such as *syndromic surveillance* (see **Syndromic Surveillance**), *biological surveillance*, and *biosurveillance* (see **Sampling and Inspection for Monitoring Threats to Homeland Security**) have been coined to describe somewhat more time-sensitive versions of public health surveillance.

The broad topic of public health surveillance includes two very different activities. *Case surveillance* (see **Dynamic Financial Analysis**) focuses on individuals, or sometimes groups of individuals, to identify individuals with certain diseases and take action to stop spread of disease. *Statistical surveillance* (see **Managing Infrastructure Reliability, Safety, and Security**), on the other hand, focuses on populations to identify differentials and trends that can inform public health policymaking, including the allocation of resources. Both approaches have roots going back at least into the nineteenth century and beyond, but it was not until 1963 that Alexander Langmuir of Centers for Disease Control and Prevention CDC combined them in his classic definition of public health surveillance [2], which formed the basis for that quoted above. Case surveillance and statistical surveillance, however, involve different goals and objectives, data sources, and methods. Indeed, the contrast between them seems responsible for some of the current controversies about modern biological and syndromic surveillance, which sits uncomfortably in both worlds.

This paper begins by identifying the goals and summarizing the history of public health surveillance, and describes some of the key data sources used. This is followed by a discussion of recent developments in biological and syndromic surveillance from a public health perspective. Many of the examples are from the United States, but the situation is similar in other countries.

The Goals and History of Public Health Surveillance

Surveillance is regarded as an essential public health service and is part of the “assessment” function of public health. The Institute of Medicine defined assessment as collecting, assembling, analyzing, and making information available on the health status and health needs of the community [3]. Traditional public health surveillance, a critical part of the assessment function, seeks to assess and track the health of the public, define public health priorities, evaluate programs, and stimulate research [4]. CDC also lists a number of “uses” of surveillance: estimating the magnitude of a health problem, determining the geographic distribution of illness, portraying the natural history of a disease, detecting epidemics, generating hypotheses, stimulating research, evaluating control measures, monitoring changes in infectious agents, monitoring emerging and reemerging infections with laboratory data, detecting changes in health practices, and facilitating planning [4].

The first use of public health surveillance was in the Republic of Venice in the fourteenth century. At that time, public health authorities boarded ships to identify persons with symptoms of bubonic plague and to prevent them from disembarking. In 1741, Rhode Island required tavern keepers to report patrons with smallpox, yellow fever, and cholera to local authorities. Statewide efforts began in Massachusetts in 1874. At that time, a voluntary, post-card reporting format was used to provide weekly reports on prevalent diseases. In 1878, Congress authorized the forerunner of the United States Public Health Service (PHS) to collect morbidity data for use in quarantine measures against “pestilential diseases” such as cholera, smallpox, plague, and yellow fever [5]. Compulsory reporting of infectious diseases began on a national basis in Italy in 1881, and in other European countries shortly afterwards. At the global level, the World Health Organization (WHO) (see **Polychlorinated Biphenyls; Cohort Studies**) modified the International Health Regulations in 2005 to require that all countries notify WHO of all events “that may constitute public health emergencies of international concern”. This requires that countries have the core surveillance and response capacities needed to fulfill the international reporting requirements [6].

From the start, surveillance was focused on detecting individual cases and taking action regarding the affected individuals. Control strategies include monitoring, treatment, quarantine, and contact tracing. Indeed, smallpox was eradicated in the 1970s using a surveillance-based approach. Earlier efforts to immunize nearly 100% of the population failed because of logistical difficulties. The successful “ring strategy” relied on intensive surveillance for identification of cases, isolation of all known cases, and immunization of individuals who may have come in contact with cases [7].

There are few cases these days of “pestilential diseases” for which immediate action is necessary, but the need for quick action to prevent the spread of infectious diseases does remain. For tuberculosis and sexually transmitted diseases, for instance, one of the main goals of surveillance is to identify infectious individuals before they infect others and thus to prevent an exponentially growing epidemic. Case surveillance has received additional prominence along with the increasing interest in emerging infections and, since 9/11, in bioterrorism. Enhancing health departments’ ability to receive and respond effectively to case reports is a critical component of the efforts of public health preparedness [8].

Surveillance also has a history as a statistical activity, although the term “surveillance” was not applied until the mid-twentieth century. In this perspective, surveillance focuses on populations rather than on individuals. The earliest statistical surveillance activities made use of vital statistics, primarily derived from death certificates. The analysis of vital statistics began with von Leibnitz and Graunt in the seventeenth century and was extended to examine cause of death and applied to public health in the nineteenth century by Farr in England and Shattuck in the United States [5].

Statistical surveillance also includes health surveys based on scientifically chosen sample surveys. The first national health survey was conducted in the United States in 1935, and the National Health Interview Survey (NHIS) has been in continuous operation since 1957 [9]. In the 1990s, the CDC developed the behavioral risk factor surveillance system (BRFSS), a model for state population health surveys, which is now used by all states and the District of Columbia [10]. In addition, many special-purpose surveys are now available and commonly used.

Registries are another source of data for statistical surveillance. The National Cancer Institute (NCI) surveillance, epidemiology, and end results (SEER) system, which operates in 11 population-based cancer registries and 3 supplemental registries covering approximately 14% of the US population, uses active surveillance methods to record all incident cases of cancer as well as their treatments and outcomes. As a result, NCI is able to estimate cancer incidence and survival rates (*see Risk Classification/Life; Reliability Data; Lifetime Models and Risk Assessment*), such estimation not being possible for most chronic diseases [11].

The extension of the original focus of surveillance from infectious diseases to other aspects of public health paralleled developments in our understanding of the determinants of health and of public health itself. The original focus in vital statistics on all-cause mortality in the seventeenth century expanded to cause-specific mortality in the nineteenth century as more became known about the causes of disease [12]. In the second half of the twentieth century, public health and surveillance began to focus on chronic diseases, and simultaneously on health risk factors as well as on health care utilization, costs, outcomes, and so on. In 1979, *Healthy People* created awareness community health indicators, performance measures, and so on [13], and surveillance efforts of this type have increased with publication of *Healthy People 2000* and *Healthy People 2010* [14, 15]. Etches and colleagues review similar efforts in Canada and other countries and suggest a framework for using indicators to report on all the domains of population health [16].

Surveillance of occupational morbidity and mortality, developed in concert with new regulations on workplace safety regulation, and surveillance of injury became more common in the 1990s as public health turned its attention to intentional and unintentional violence. A growing focus on health care quality in the early twenty first century and attendant concerns about medical errors and iatrogenic injuries in recent years [17, 18], have led to intensified surveillance efforts, along with postmarketing surveillance for adverse effects of drugs and vaccines [19].

Surveillance Data

In order to support the different purposes and uses summarized above, surveillance systems rely

on a diversity of data sources. Infectious disease surveillance, for instance, still relies primarily on the system of notifiable diseases put in place in the nineteenth century. Health care providers are required to report all cases of a given set of diseases to their local health department. Reports are generally made by name and with a variety of information relating to possible risk factors to the local health department in which the patient resides. Local health departments report to state health departments, usually by mail, and states compile reports and transmit them to the CDC for national tabulation and publication. Disease surveillance systems have generally been developed for specific diseases and differ from one another in what information is requested, how it is reported, and so on.

Not surprisingly, given this fragmented system and the demands of the modern health care system, physicians' compliance with this *passive surveillance* system is limited. Moreover, some individuals may not experience symptoms or have good access to health care, and thus do not seek care. As a result, health officials acknowledge that the number of reported cases of a disease can be far below the number of actual cases. In addition, because most surveillance systems require laboratory confirmation of a diagnosis before it is reported, there can be substantial delays between the development of symptoms in an individual and reporting to health officials.

Two recent developments may help to improve the completeness and accuracy of case reports. First, health departments are developing systems for clinical laboratories to report cases that test positive directly to state or local health departments rather than through physicians. While this may make reporting both more complete and timely, laboratories generally do not have all the clinical and demographic information on the cases that is necessary. Moreover, managed care organizations and other organized health care delivery systems are increasingly using out-of-state laboratories, and linking the information back to the proper local health department can be difficult. Second, CDC is developing a National Electronic Disease Surveillance System (NEDSS) for transmitting surveillance information electronically but it has only been implemented in a few states [20]. Current implementations, however, are focused on communication between local and state agencies and with the CDC, and direct links to providers are not generally available. With the

increased emphasis on surveillance after September 11, linkages to physicians, hospitals, and laboratories are being developed [20].

Vital records, based on individual birth and death certificates, are another important source of surveillance data. Unlike disease case reports, however, birth and death certificates are legal documents that are essential to the persons concerned and their families, and thus there is a strong incentive for the reports to be complete. Timeliness has generally not been an issue with vital statistics and the emphasis has been on completeness and accuracy in their compilation and analysis.

Registries for cancer and other diseases, as described above, provide data on disease incidence, treatments, case fatality, and health outcomes. They generally rely on *active surveillance* efforts, where registry personnel regularly contact physicians, hospitals, and other health care providers to ascertain new cases and learn about developments on existing cases. Because of the cost of these efforts, registries only exist for selected diseases and in limited geographic areas. As with vital records, registry data tend to be analyzed and reported in statistical terms, and the focus is on completeness and accuracy rather than timeliness.

Sentinel surveillance efforts are specially focused, intensive efforts to identify disease incidence and case characteristics in special settings. Because influenza case surveillance reports are very incomplete, for instance, CDC has organized the Sentinel Provider Network (SPN), which consists of approximately 1200 healthcare providers nationwide, who report the number of weekly outpatient visits for influenza-like illness (ILI) [21].

Population-based surveys using scientifically selected random samples provide useful data for statistical surveillance activities. The NHIS and BRFSS, as discussed above, are general-purpose surveys at the national and state levels, respectively, and many other special-purpose surveys have been developed and used at the national, state, and local levels, and for special populations. Population-based surveys are designed to make estimates of the total number of cases using statistical techniques, but not to identify all cases of a particular disease or health condition. The accuracy of these estimates depends on the selection of the sample, complete and accurate reporting by those in the sample, and use of proper statistical methods to analyze the data. Because the

individuals providing data are only a sample of the target population, attempts are generally not made to respond to cases that are identified in the survey.

The final major source of surveillance data is medical records and administrative systems. Such records have existed for many years, but they are increasingly used for both research and surveillance purposes now that they are available in computerized form. Because medical decisions and payments are based on such documents, they are generally regarded as nearly complete. Their administrative and clinical use, however, can introduce biases. An individual's condition, for instance, may be "upcoded" to increase a hospital's or a physician's reimbursement. Such records may also not include information such as risk factors or race and socioeconomic status, which can be useful for surveillance purposes. New federal regulations established in response to the Health Insurance Portability and Accountability Act (HIPAA) restrict the use of patient information to protect privacy and confidentiality. Hospitals and physicians can, however, disclose protected health information for public health activities authorized by law [22].

Syndromic Surveillance

Heightened awareness of the risks of bioterrorism since 9/11, coupled with a growing concern about naturally emerging and re-emerging diseases such as West Nile, severe acute respiratory syndrome (SARS), and pandemic influenza have led public health policymakers to realize the need for early warning systems and, more generally, improved surveillance. The sooner health officials know about an attack or a natural disease outbreak, for instance, the sooner they can treat those who have already been exposed to the pathogen to minimize the health consequences, vaccinate some or all of the population to prevent further infection, and identify and isolate cases to prevent further transmission. In addition, it is felt that improved surveillance systems should allow for better "situational awareness" and thus help to manage the response to public health emergencies.

Traditional public health surveillance approaches monitor disease using prespecified case definitions and use manual data collection, human decision-making, and manual data entry. In contrast, current electronic biosurveillance systems use sophisticated information technology and statistical methods to

gather, process, and analyze large amounts of data and display the information for decision makers in a timely way. For instance, syndromic surveillance systems assume that during an attack or a disease outbreak people first develop symptoms, then stay home from work or school, attempt to self-treat with over-the-counter (OTC) products, and eventually see a physician with nonspecific symptoms days before they are formally diagnosed and reported to the health department. To identify such behaviors, syndromic surveillance systems regularly monitor existing data for sudden changes or anomalies that might signal a disease outbreak. Syndromic surveillance systems have been developed to include data on school and work absenteeism, sales of OTC products, calls to nurse hotlines, and counts of hospital emergency room (ER) admissions, or reports from primary physicians for certain symptoms or complaints [23]. Fricker offers a more detailed discussion on the various statistical methods and algorithms used in syndromic surveillance.

CDC's system for electronic biosurveillance, BioSense, was initiated in 2003 and currently receives data from national healthcare organizations including the Department of Defense (DoD), Veterans' Affairs (VA), and other sources. The ICD-9 diagnosis codes and current procedural terminology (CPT) procedure codes from these sources are binned into 11 syndromes. Data analyses are displayed in the BioSense application, which can be viewed through the CDC secure data network by CDC, state and local health departments, and DoD and VA personnel [24]. In mid-2005, CDC began to develop the capacity, through BioSenseRT, to receive real-time data from hospital facilities across the United States. This initiative emphasizes access to real-time clinically rich data. Individual patient records are stripped of obvious identifiers such as names, addresses, and telephone numbers at the data source before being sent to CDC. Patient populations include those treated in the emergency department (ED), hospital inpatient, and hospital-affiliated outpatient settings. Hospital-level data includes the daily hospital census. Patient-level data of interest includes patient's chief complaint, physician diagnosis, selected ED-specific data (e.g., vital signs), microbiology laboratory orders and results, radiology order and results, and pharmacy orders. As of December 2005, chief complaint, diagnosis, and selected demographic data were being received from 32 hospitals in 10 cities

distributed across the United States; and plans call for the number of hospitals to increase to 350 [25]. BioSense's vision is to (a) provide local, state, and nationwide health situational awareness for suspected illness and cases of disease before, during, and after a health event; (b) help confirm or refute the existence of an event; and (c) monitor the size, location, and rate of spread of disease outbreaks [26].

Recognizing that the "ability to gather and analyze information quickly and accurately would improve the nation's ability to recognize natural disease outbreaks, track emerging infections, identify intentional biological attacks, and monitor disease trends," the Institute of Medicine has recently called for more research on syndromic surveillance and other "innovative systems of surveillance that capitalize on advances in information technology". However, since surveillance systems in the United States "remain fragmented and have not evolved at the same rate as ... electronic technological advances," the Institute of Medicine calls for these systems to be "carefully evaluated for their usefulness in detection of infectious disease epidemics, including their potential for detection of major biothreat agents, their ability to monitor the spread of epidemics, and their cost effectiveness" before widespread implementation [27].

Practical Concerns about Syndromic Surveillance in Public Health Practice

Despite the generally recognized promise of syndromic surveillance and other electronic biosurveillance systems, there are practical concerns about the use of these systems, and BioSense in particular, in state and local public health practice.

Reingold provocatively noted the challenges confronting syndromic surveillance: characterizing the types of bioterrorist events that it is likely to detect, determining whether it can help identify the population at risk in a more timely way during a bioterrorist event, determining the appropriate response to apparent increases in illnesses signaled by syndromic surveillance, and ultimately demonstrating that it can reduce morbidity and/or mortality following a bioterrorist event. Reingold also notes the likelihood that syndromic surveillance will produce useful information about naturally occurring diseases and the importance of identifying the circumstances under which it is likely to strengthen local and state public health departments [28].

The possibility of earlier detection and more rapid response to a bioterrorist event has tremendous intuitive appeal, but its success depends on local health departments' ability to respond effectively. When a syndromic surveillance system sounds an alarm, health departments typically wait a day or two to see if the number of cases continues to remain high or if a similar signal is found in other data sources. Doing so, of course, reduces both the timeliness and sensitivity of the original system. If the health department decides that an epidemiological investigation is warranted, it may begin by identifying those who are ill and talking to their physicians. If this does not resolve the matter, additional tests must be ordered and clinical specimens gathered for laboratory analysis. Health departments might also choose to initiate active surveillance by contacting physicians to see if they have seen similar cases.

Additionally, a syndromic surveillance system that only says "there have been five excess cases of flu-like illness at hospital X" is not of much use unless the five cases can be identified and reported to health officials. For example, if there are 55 rather than 50 cases expected, syndromic surveillance systems cannot say which 5 are the "excess" ones, and all 55 must be investigated. Health departments cannot act simply on the basis of a suspicion. Even when the cause and route of exposure are known, the available control strategies – quarantine of suspected cases, mass vaccination, and so on – are expensive and controversial, and often their efficacy is unknown. Coupled with the confusion that is likely during a terrorist attack or even a natural disease outbreak, making decisions could take days to weeks.

Syndromic surveillance systems are intended to raise an alarm, and all alarm systems have intrinsic statistical trade-offs. The most well known is between sensitivity, the ability to detect an attack when it occurs, and the false-positive rate, the probability of sounding an alarm when there is in fact no attack. The costs of excessive false alarms are both monetary, in terms of resources needed to respond to phantom events, and operational, as too many false events desensitize responders to real events. Taking into account the different data types and multiple jurisdictions, thousands of syndromic surveillance systems soon will be running simultaneously in cities and counties throughout the United States. If 1000 data streams are being monitored, each with a 0.1%

false-positive rate (which is very low), there will be approximately one false alarm per day.

The timeliness of a surveillance system depends on the time it takes to generate and acquire data, analyze it, and take action [29]. With syndromic surveillance, an additional component is the time that it takes to accumulate enough evidence of an outbreak to trigger a detection algorithm. To illustrate this point, Stoto and colleagues used a simulation approach to analyze ILI-related ER admissions data from a typical urban hospital. A hypothetical number of extra cases spread over a number of days were added to actual baseline data to mimic the pattern of a potential bioterror attack (*see Digital Governance, Hotspot Detection, and Homeland Security*). The results indicate the size and speed that outbreaks must attain before they are detectable. These results are sobering: even with an excess of 9 cases over 2 days (the first 2 days of the “fast” outbreak), three times the daily average, there was only about a 50% chance that the alarm would go off. When 18 cases were spread over 9 days, chances were still no better than 50–50 that the alarm would sound by the ninth day. Moreover, this finding holds true only outside of the winter flu season. In the winter, the detection threshold (*see Hazardous Waste Site(s); Early Warning Systems (EWSs) for Predicting Financial Crisis*) must be set high so that the flu itself does not sound an alarm, but a terrorist attack that appeared with flulike symptoms would be harder to detect (Figure 1) [30]. Fricker and Rolka discuss other statistical challenges to the development and effective use of electronic surveillance systems in public health practice [31].

Ultimately, the value of syndromic surveillance as an early detection system depends on how well it is integrated into public health systems. The detection of a sudden increase in cases of flulike illness – the kind of event that syndromic surveillance can detect – can have varied significance. It could be a bioterrorist attack but is more likely a natural occurrence, perhaps even the beginning of the annual flu season. An increase in sales of flu medication might simply mean that pharmacies are having a promotion. A surge in absenteeism could reflect natural causes, or even a period of particularly pleasant spring weather. Similar problems can occur even with the more complete dataset in the real-time implementation of BioSenseRT. For example, changes in local hospital systems, or even in coding practices, can result in

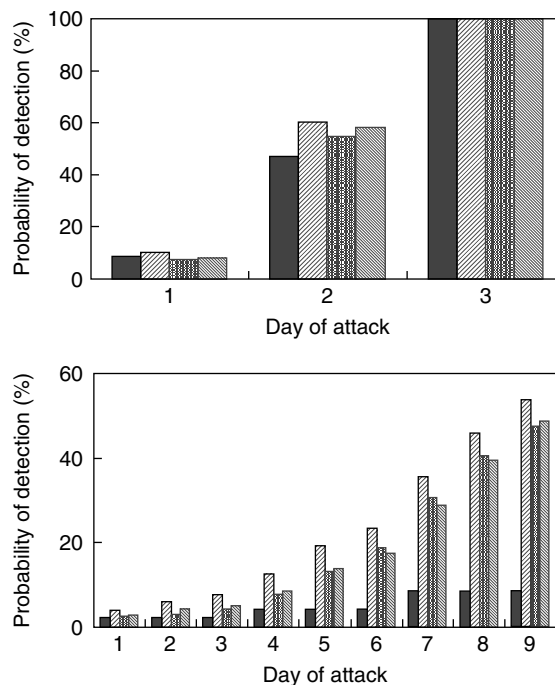


Figure 1 Sensitivity of syndromic surveillance (probability of detection by day) for influenza-like illness at a typical urban hospital emergency room using four detection algorithms, (a) fast outbreak: 18 cases over 3 days, (b) slow outbreak: 18 extra cases over 9 days [Reproduced from [30] © Springer New York LLC, 2004.]

substantial changes that could raise concern if they are not understood.

Alternative Applications of Syndromic Surveillance

Since 9/11, the focus of syndromic surveillance efforts has been on early detection of bioterrorist events. The most value, however, may ultimately come from its use in the detection of natural disease outbreaks. More generally, if twenty-first century bio-surveillance means effective use of health information technology in identifying cases before they are formally diagnosed, it can supplement traditional public health approaches and improve their effectiveness.

One potential use is in detecting influenza outbreaks. In an “ordinary” year, influenza results in approximately 36 000 deaths and more than 200 000 hospitalizations in the United States. In addition to

this human toll, influenza is annually responsible for a total cost of over \$10 billion in the United States. A pandemic, or worldwide outbreak of a new influenza virus, perhaps evolving from the H5N1 avian flu virus circulating in Asia, could dwarf this impact by overwhelming our health and medical capabilities, potentially resulting in hundreds of thousands of deaths, millions of hospitalizations, and direct and indirect costs of hundreds of billions of dollars [32]. Electronic biosurveillance systems, syndromic surveillance, and BioSense feature prominently in federal, state, and local plans to prepare the United States for pandemic flu.

CDC has a number of influenza surveillance systems in place [21], yet they do not provide population-based rates of incidence or prevalence rates on a national level because many infected persons are asymptomatic or experience only mild illness and do not seek medical care. Also, laboratory testing is rare in less severe cases, and testing occurs late in the course of illness (e.g., in cases with severe complications) and can yield false-negative results because the patient is no longer shedding virus.

Epidemiological characteristics (*see Molecular Markers and Models; Meta-Analysis in Nonclinical Risk Assessment*) of both seasonal and pandemic influenza, however, suggest that syndromic surveillance and other biosurveillance systems are likely to make an important contribution beyond the capabilities of existing surveillance systems, and thus enable a more effective public health response [33]. Simulation studies have shown that unless a bioterrorism outlook is exceptionally large, syndromic surveillance detection algorithms take on the order of days to detect them [30, 33, 34]. This time frame is longer than some proponents of syndromic surveillance feel is needed to detect bioterrorism [35]. Compared to the current influenza surveillance systems, however, a 1-week lead time would provide valuable information, and this is likely to be achievable for syndromic surveillance.

Furthermore, a number of studies have demonstrated the potential that syndromic surveillance of ILI offers at the national, state, and local level. Sebastiani and colleagues show that children and infants presenting to the pediatric ED with respiratory syndromes are an early indicator of impending influenza morbidity and mortality, sometimes by as much as 3 weeks [36]. Using data from New York

City, Lu and colleagues have shown that monitoring both outpatient and ED data can enhance detection of ILI outbreaks [37]. With similar data, Olson and colleagues note that age-stratified analyses of ED visits for fever and respiratory complaints offers the potential for more precise quantification of the burden of illness, earlier warning of the arrival of epidemic influenza, and greater sensitivity for detecting the characteristic age shift of pandemic influenza [38]. Comparing unspecified infection cases in DC hospitals using optimal detection algorithms to CDC's sentinel physician data for the South Atlantic states for 4 years in which there was a discernable influenza outbreak, Stoto and colleagues found that in two of these years, the DC syndromic surveillance based on hospital ER data outperformed the other two systems, and in 1 year it flagged only 2 days after the CDC system [39]. Given a built-in delay of about 2 weeks in the CDC system, this is a substantial advantage.

Even in normal flu seasons, laboratory analysis to determine whether a case is truly influenza, or to identify the viral strain, is rarely done. Pandemic influenza, in which an antigenic shift causes an outbreak that could be more contagious and/or more virulent and to which few people are immune by virtue of previous exposure, is a growing concern. Syndromic surveillance of flulike symptoms might trigger more laboratory analysis than is typically done, and in this way hasten the public health response. In a normal flu season, Labus has reported that early identification of the start of the influenza season using syndromic surveillance in Clark County Nevada enabled the notification of the medical community. Physicians were encouraged to submit specimens for culture, and the county health department provided kits to help them do this, which allowed for rapid identification of the major circulating strain. In 2003–2004 (a period with a marked increase of early season influenza and deaths in children in other parts of the country), this syndromic surveillance system allowed for better tracking, and provided data for daily reports to decision makers and the media [40].

Because of their focus on the early detection of bioterrorist events, syndromic surveillance systems are intended to detect large increases in the number of people with common symptoms such as ILI. As such, they cannot be expected to detect small numbers of cases, even if very unusual. One reason is that in a small disease outbreak or the early stages of a larger

one, each case will be seen by only one physician. The natural tendency of physicians who see only one case, however suspicious it may be, is to discount it. After all, physicians are appropriately taught “when you hear hoofbeats, think horse not zebra”. And some may fear the embarrassment of reporting a case that may turn out to be a false alarm.

Modern health informatics systems provide the potential to identify the presence of small numbers of cases of concern before they are formally diagnosed. Automated systems can, for instance, aggregate data for a metropolitan area spanning local reporting jurisdictions to identify, say, cases of rash and fever, which would suggest smallpox. Systems can also be set up to enable and encourage early reporting of cases on the basis of symptoms only. For instance, the syndrome reporting information system (SYRIS), now operating in Lubbock Texas and elsewhere, enables physicians to report suspicious cases to the local health department without waiting for laboratory confirmation, and encourages them to do so by providing feedback in the form of information about practice guidelines and other similar cases [41]. This can be thought of as a kind of active syndromic surveillance or as IT support for astute physicians.

Real-time access to prediagnostic data can also help health authorities respond to public health threats. If person-to-person transmission of avian flu virus is documented in Asia, for instance, health department in Europe and the United States might want to identify and follow-up on local cases of people hospitalized with flulike symptoms, and syndromic surveillance systems could be designed to identify them. If an environmental sensor detects signs of possible terrorist agent tularemia, syndromic surveillance systems can be checked for cases with appropriate symptoms. This actually occurred in Washington DC in 2005, and the lack of cases in area ERs reassured local officials that the alarm was false. Syndromic surveillance systems can also be queried to determine background rates when it is not clear whether a reported cluster of cases is out of the ordinary.

The *Escherichia coli* O157:H57 outbreak in the New York City area in late 2006 provides an example of how syndromic surveillance could have been used for case finding. The outbreak came to light on November 17 when the first case was reported to a local health department in New Jersey. By November 27, 11 cases were reported in that jurisdiction and

3 days later the Taco Bell restaurant, where 9 of the 11 cases had eaten, closed voluntarily. On December 1, a similar case (originally attributed to another cause) was reported to a local health department in New York State, and it turned out that this person and three more cases in that jurisdiction had eaten at a different Taco Bell restaurant. By December 4, all Taco Bells in the New York metropolitan area were closed, and 2 days later a particular food item, green onions, was identified as the likely source of contamination. By December 9, more than 61 *E. coli* O157:H57 cases in at least four states were reported [42].

Although a number of syndromic surveillance systems were operating at this time in New York City and the surrounding jurisdictions, there were too few cases in any location to detect. However, once the outbreak was identified in New Jersey, an advanced syndromic surveillance system could have searched ED admissions for cases of bloody diarrhea and abdominal cramps in the entire metropolitan area. Cases so identified could have been interviewed to obtain a food history and lab samples could have been obtained to test for *E. coli* O157:H57. In addition, health departments could have initiated active surveillance by physicians in the area, searched data from surrounding states to identify additional cases for follow-up and to confirm lack of cases elsewhere. If these things had been done, it is possible that the restaurant chain and green onions could have been identified and remedial steps taken earlier – either closing the restaurant or removing the green onions. It is also likely that the additional data that syndromic surveillance systems with these capabilities could have provided would have generated information to resolve the uncertainty about what was happening, and thus diminish public concerns.

Conclusions

Disease surveillance, combining case-finding efforts with statistical approaches to track population health trends, is an essential core function of public health. In some respects, however, its methods have changed little since the nineteenth century. The introduction and growth of syndromic surveillance following 9/11 has focused on a statistical approach to detecting bioterrorist attacks, and much impressive research and development has been done with respect to the real-time integration of many disparate data streams

and other aspects of information technology, and to the development of statistical detection algorithms. Despite these accomplishments, the value of syndromic surveillance for detecting bioterrorist attacks has not yet been demonstrated. This is because of the relatively narrow window between what can be detected in the first few days and what is obvious, and the need for better integration with public health systems.

Ultimately, the most important public health contribution of syndromic surveillance systems may be for natural disease outbreaks, such as seasonal and pandemic influenza, and for other public health purposes beyond simple detection. For this to be achieved, however, syndromic surveillance system designers will have to change their focus from statistical systems and from data appropriate for early detection of large-scale events, and develop tools for public health uses other than outbreak detection.

References

- [1] Thacker, S.B. & Berkelman, R.L. (1988). Public health surveillance in the United States, *Epidemiology Review* **10**, 164–190.
- [2] Langmuir, A.D. (1963). The surveillance of communicable diseases of national importance, *New England Journal of Medicine* **268**, 182–192.
- [3] Institute of Medicine (1988). *The Future of Public Health*, National Academy Press, Washington, DC.
- [4] Centers for Disease Control and Prevention (2001). *Updated Guidelines for Evaluating Public Health Surveillance Systems*, at <http://www.cdc.gov/mmwr/pre-view/mmwrhtml/rr5013a1.htm> (accessed Nov 2007).
- [5] Thacker, S.B. (2000). Historical development, in *Principles and Practice of Public Health Surveillance*, 2nd Edition, S.M. Teutsch & R.E. Churchill, eds, Oxford, New York, pp. 1–16.
- [6] World Health Organization (2005). *International Health Regulations (IHR)*, at <http://www.who.int/csr/ihr/en/> (accessed Nov 2007).
- [7] Henderson, D.A. (1999). Smallpox: clinical and epidemiologic features, *Emerging Infectious Diseases* **5**, 537–539.
- [8] Centers for Disease Control and Prevention (2006). *Cooperative Agreement Guidance for Public Health Emergency Preparedness*, at <http://www.bt.cdc.gov/planning/coopagreement/#fy06> (accessed Nov 2007).
- [9] Centers for Disease Control and Prevention (2007). *National Health Interview Survey (NHIS)*, at <http://www.cdc.gov/nchs/about/major/nhis/hisdesc.htm> (accessed Jan 2007).
- [10] Centers for Disease Control and Prevention (2006). *About the BRFSS*, at <http://www.cdc.gov/brfss/about.htm> (accessed Nov 2007).
- [11] National Cancer Institute (2003). *Overview of the SEER Program*, at <http://seer.cancer.gov/about/> (accessed Jan 2007).
- [12] Stoto, M.A. & Durch, J.S. (1993). Forecasting survival, health, and disability: report on a workshop, *Population and Development Review* **19**, 557–582.
- [13] U.S. Department of Health, Education and Welfare (1979). *Healthy People: The Surgeon General's Report on Health Promotion and Disease Prevention*, Government Printing Office, Washington, DC.
- [14] U.S. Department of Health and Human Services (1991). *Healthy People 2000: National Health Promotion and Disease Prevention Objectives*, Office of the Assistant Secretary for Health, Washington, DC.
- [15] U.S. Department of Health and Human Services (2000). *Healthy People 2010*, 2nd Edition, Government Printing Office, Washington, DC.
- [16] Etches, V., Frank, J., Di Ruggiero, E. & Manuel, D. (2006). Measuring population health: a review of indicators, *Annual Review of Public Health* **27**, 29–55.
- [17] Institute of Medicine (1999). *To Err is Human: Building a Safer Health System*, National Academy Press, Washington, DC.
- [18] Institute of Medicine (2001). *Crossing the Quality Chasm: A New Health System for the 21st Century*, National Academy Press, Washington, DC.
- [19] Strom, B.L. (2006). How the US drug safety system should be changed, *Journal of the American Medical Association* **295**, 2072–2075.
- [20] Centers for Disease Control and Prevention (2007). *About PHIN: Public Health Information Network*, at <http://www.cdc.gov/phn/resources/phn-facts.html> (accessed Nov 2007).
- [21] Centers for Disease Control and Prevention (2007). *Overview of Influenza Surveillance in the United States*, at <http://www.cdc.gov/flu/weekly/pdf/flu-surveillance-overview.pdf> (accessed Nov 2007).
- [22] Gostin, L.O. (2001). National health information privacy: regulations under the health insurance portability and accountability act, *Journal of the American Medical Association* **285**, 3015–3019.
- [23] Mandl, K.D., Overhage, J.M., Wagner, M.M., Lobner, W.B., Sebastiani, P., Mostashari, D., Pavlin, J.A., Gesteland, P.H., Treadwell, T., Koski, E., Hutwagner, L., Bukeridge, D.L., Aller, R.D. & Grannis, S. (2004). Implementing Syndromic surveillance: a practical guide informed by the early experience, *Journal of the American Medical Informatics Association* **11**, 141–150.
- [24] Loonsk, J.W. (2004). BioSense: a national initiative for early detection and quantification of public health emergencies, *Morbidity and Mortality Weekly Report* **53**, 53–55.
- [25] Centers for Disease Control and Prevention (2007). *BioSense For Public Health Departments*, at <http://www.cdc.gov/biosense/publichealth.htm> (accessed Nov 2007).

- [26] Steele, L. (2006). Biosense: using health data for early event detection and situational awareness, *Presented at the BioSense Science Input Meeting*, Atlanta.
- [27] Institute of Medicine (2003). *Microbial Threats to Health: Emergence, Detection, and Response*, National Academy Press, Washington, DC.
- [28] Reingold, A. (2003). If syndromic surveillance is the answer, what is the question? *Biosecurity and Bioterrorism* **1**, 1–5.
- [29] Buehler, J.W., Berkelman, R.L., Hartley, D.M. & Peters, C.J. (2003). Syndromic surveillance and bioterrorism-related epidemics, *Emerging Infectious Diseases* **9**, 1197–1204.
- [30] Stoto, M.A., Schonlau, M. & Mariano, L.T. (2004). Syndromic surveillance: is it worth the effort? *Chance* **17**, 19–24.
- [31] Fricker, R.D. & Rolka, H.R. (2006). Protecting against biological terrorism: statistical issues in electronic bio-surveillance, *Chance* **19**, 4–13.
- [32] Homeland Security Council (2005). *National Strategy for Pandemic Influenza*, The White House, Washington, DC.
- [33] Stoto, M.A., Fricker, R.D., Jain, A., Diamond, A., Davies-Cole, J.O., Glymph, C., Kidane, G., Lum, G., Jones, L., Dehan, K. & Yuan, C. (2006). Evaluating statistical methods for syndromic surveillance, in *Statistical Methods in Counter-Terrorism*, D. Olwell, A.G. Wilson & G. Wilson, eds, Springer, New York, pp. 141–172.
- [34] Jackson, M.L., Baer, A., Painter, I. & Duchin, J. (2005). A method for evaluating algorithms used in syndromic surveillance, *Presented at the National Syndromic Surveillance Conference*, Seattle.
- [35] Wagner, M.M., Tsui, F.C., Espino, J.U., Dato, V.M., Sittig, D.F., Caruana, R.A., McGinnis, L.F., Deerfield, D.W., Druzdel, M.J. & Fridsma, D.B. (2001). The emerging science of very early detection of disease outbreaks, *Journal of Public Health Management and Practice* **7**, 50–58.
- [36] Sebastiani, P., Mandl, K.D., Szolovits, P., Kohane, I.S. & Ramey, M.F. (2006). A Bayesian dynamic model for influenza surveillance, *Statistics in Medicine* **25**, 1803–1816.
- [37] Lu, J., Konty, K., Mostashari, F. & Metzger, K. (2006). Could outpatient visits enhance our ability of early detecting influenza-like illness outbreaks? *Advances in Disease Surveillance* **1**, 44.
- [38] Olson, D.R., Heffernan, R.T. & Mostashari, F. (2005). Age matters: emergency department syndromic surveillance for epidemic influenza, *Presented at the National Syndromic Surveillance Conference*, Seattle.
- [39] Stoto, M.A., Griffin, B.A., Jain, A., Davies-Cole, J.O., Lum, G., Kidane, G. & Washington, S.C. (2006). Fine-tuning and evaluation of detection algorithms for syndromic surveillance, *Presented at the National Syndromic Surveillance Conference*, Baltimore. [Paper has been submitted to a journal, and this reference will be updated when accepted].
- [40] Labus, B. (2005). *Practical Applications of Syndromic Surveillance: 2003–2004 Influenza Season. Emergency Preparedness E-Newsletter*, NACCHO, at <http://www.naccho.org/topics/emergency/APC/E-Link/documents/PracticalApplicationsofSyndromicSurveillance.pdf> (accessed Jan 2007).
- [41] Lindley, C. & Ward, T. (2006). Experience with clinician-based syndromic surveillance in West Texas, *Presented at the National Syndromic Surveillance Conference*, Baltimore. Abstract available at [http://thci.org/_documents/temp/Syndromic Surveillance Conf2006.doc](http://thci.org/_documents/temp/Syndromic%20Surveillance%20Conf2006.doc) (accessed Jan 2007). [Abstract will be published in *Advances in Disease Surveillance*].
- [42] Centers for Disease Control and Prevention. *Multistate Outbreak of E. coli O157 Infections Linked to Taco Bell*, at <http://www.cdc.gov/ecoli/2006/december/index.htm> (accessed Jan 2007).

Related Articles

Syndromic Surveillance

MICHAEL A. STOTO

Public Participation

Public participation processes and stakeholder involvement strategies are adopted to engage people in a decision-making process that they have an interest in, or that may affect them in some way. They may be experts, representatives of different interest groups, or the general public (*see* **Expert Judgment; Group Decision**). Involving different stakeholder groups gives decision makers a clear idea of what the public wants, and involving local people in local decisions may also provide information that may have otherwise been overlooked. By these means, the decision-making process is strengthened and the quality of decision is improved. Participation of different groups with opposing views also helps to resolve conflict. Compromise may be reached with negotiation, and when different viewpoints and values are heard, participants may become more sympathetic to the opposing cause. By opening up the decision-making process to the public, granting them some control over their own environment, and so increasing transparency, trust can be built. Participatory processes also seek to educate and inform the public about matters they may otherwise have been fearful of. Thus, participation processes can be adopted to involve stakeholders in the management of risk and the development of risk mitigation strategies.

Objectives of Participation

There are a number of objectives one may wish to achieve when involving the public in a participatory process. Here the objectives of public participatory processes that have been identified from an extensive literature search are described. Figure 1 gives an overview of the possible objectives of participation. Alongside and leading the practical developments, there have been many academic studies covering issues such as the philosophical underpinnings, categorization of different types of activity, and case studies [1–25]. Beyond the literature on participation and deliberative democracy, there is also other relevant material on discussions of public engagement with science and more generally risk communication [26–31]. Aside from a social science literature, there is also relevant material in the management literature [32–37]. Information systems approaches to

participatory design are also very relevant [38]. In the following sections the objectives of participation are discussed.

Information Sharing

A vital purpose of public participation processes is the exchange of information in order to inform participants and explore issues thoroughly. Scientific and technical information communicated should be made easier to understand, but not oversimplified. Information sharing can be one-way or two-way between stakeholders and authorities and/or decision makers. One-way information sharing from the authority to stakeholders may relate to an event, or be educational to explain “science” and “expert views”, for example, an education campaign highlighting the dangers of drink driving or information dissemination in the event of environmental flooding. Conversely, an authority may seek local knowledge about an issue or an area where complete information is not available. Authorities may wish to determine public perceptions of an issue or understand lay views. Consideration of the way in which the public perceives risks and other information is important to any communication or involvement process. It is well understood in risk perception literature that “lay” judgments and “expert” opinions on the nature of various risks and the hazards they pose differ to a large extent. Neither is necessarily right or wrong but the differences need to be appreciated. It may be feasible to change public perceptions and viewpoints with information provision and, with reassurance, the public are more likely to accept a decision it perceives to involve a high degree of risk. The 1998 Aarhus convention, ratified by 40, Asian and European countries, recognizes that the public have a right to participate in environmental decision making. Incorporation of public values and preferences is an important aspect of public participation processes; the process will have more credibility if the public feel they are being listened to; the decision will be more widely accepted if the public view has been taken into consideration; and the quality of decisions will be improved. Different stakeholders and members of the public hold different values and so it is important to involve a diverse array of people in the process. A two-way information sharing consultation process may be employed for a decision-making

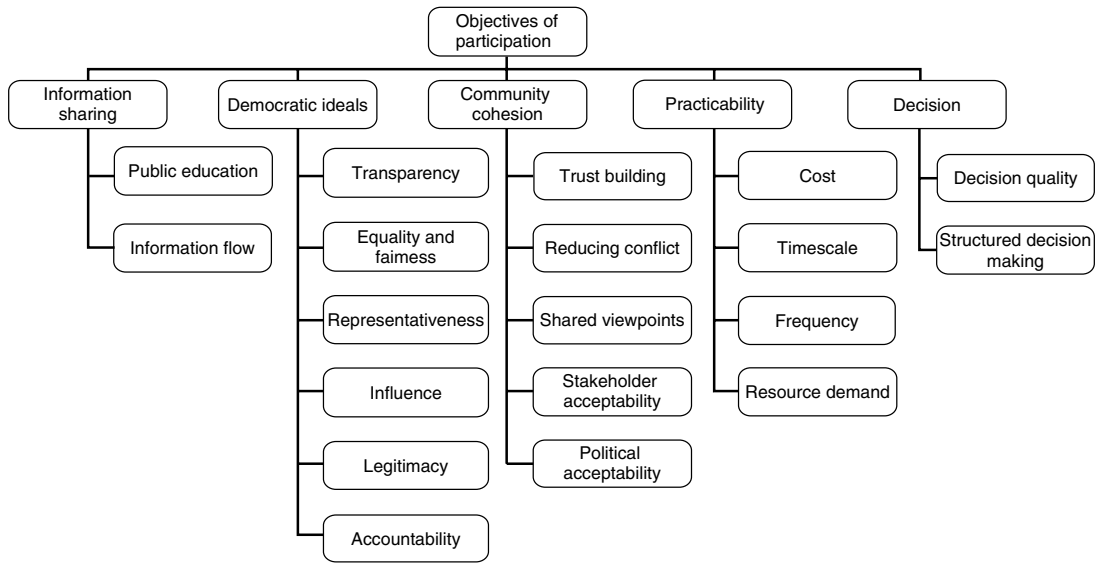


Figure 1 Objectives of participation

process related to a local environment that meet the needs of all stakeholders.

Decision/Risk-Management Quality

More creative alternatives, better identification of risks, and problem formulation can all be achieved by drawing in public and stakeholder ideas early in the process. The success of a public participation process may be determined by the active involvement of participants and their responsiveness to information they have received. In one study [9] of the quality of stakeholder-based decisions, one of the criteria examined was whether participants contributed innovative ideas, useful analysis, or new information. In the majority of cases it was found that the processes did meet this criterion, therefore improving decision quality. The public can provide valuable information and an alternative insight into the decision problem.

Decision quality may be improved by introducing more flexibility in problem formulation and broader framing of issues; in order for the public to understand the decision situation fully, the processes should be adaptable enough to allow participants to raise issues outside the remit of the dialogue process – the process should not predetermine the way in which the decision is discussed [39]. Participants may also widen the framing of a problem by introducing social and ethical issues alongside scientific

and technical ones. The openness of problem framing may be constrained by the process, which may jeopardize the benefits of dialogue. Decision quality may be improved by testing assumptions. As a starting point to any process, assumptions are made in order to guide the process and to be prepared for any information that may surface in discussion, particularly information that may be sensitive. These assumptions should not be constraining and can be tested.

Community Cohesion

A well-designed participatory process may contribute to social cohesion, it can smooth tensions between those closely involved in the issues and who may have conflicting objectives, and it can create a more general feeling of community involvement. Trust is an essential part of any process; if there is no trust then people will become suspicious of the process and of the people involved. Trust is important to break down communication barriers and promote open and honest discussion rather than discussion that is defensive and self-interested. It is particularly important to build and maintain trust between different stakeholder groups and the authority. The decline of public trust in government and government-related institutions in the management of risk situations has been apparent in recent years, especially since public relations disasters such as the BSE crisis in the United Kingdom

where information about this disease in cattle and links between the human form, vCJD, were undisclosed. The decline of public trust in government can be attributed to a number of factors; in particular, public expectations have risen over time – people expect more from organizations and the individuals representing them. Also, there has been a decline in deference to authority so that people are less willing to accept government or “expert” advice without question [40]. Opening up the decision-making process to the public and granting them some control over their own environment aim to increase transparency and build trust. It is important to build trust between different stakeholder groups. Improving trust and understanding between participants can help to encourage more thorough and in-depth discussion.

Democratic Ideals

There are a whole range of democratic issues that need to be considered when designing a participatory process: *Transparency*, to build trust in the authorities conducting the process and in the process itself. Transparency is essential not only for participants but also for the wider public in order for them to accept the decision or the outcome deliberation; the *legitimacy* of the process is important: the reasons for the process, its purpose, and the motivation of the body conducting the process need to be made clear as otherwise participants may feel skeptical and disillusioned with the process; *equality and fairness*, the rights of individuals to participate and express their views fairly and equally; *representativeness* of all stakeholders: to ensure that all viewpoints are represented; and *accountability*, to ensure that decision makers are acting in the public's best interest. However, there may be occasions where participatory processes are not intended to be democratic, for example, the handling of incidents may require too much urgency to debate issues with the public and involve them in the ultimate decisions. Participation in such cases is much more like to be focused on explaining what is being done and why, and perhaps gathering information on the full scale of the likely consequences.

Practicability

Full participation can be very costly and time consuming, requiring complex logistics. The practicalities of participation need to be considered, such as

the costs to hold a consultation process, the demands on resources and on participants' time, and whether the chosen process will produce acceptable outcomes given the constraints.

Characteristics of Different Participatory Processes

The characteristics of the participatory process chosen will reflect the aims and objectives of the task in hand. The number and composition of different stakeholders as participants in a decision-making process may affect the outcome. Different groupings exhibit altered behavior, and group members interact differently depending on the dynamics of the group. Some of the characteristics to consider when designing a participatory process are covered here.

Participants

The participants involved in the process are crucial to its outcome. The way in which participants interact with one another cannot be easily predicted, as different characters bring unique values and insight to the discussion. While participants should be representative of the population affected by the decision, considering qualities such as age, sex, and cultural background may add to the difficulty of the process. There are several ways in which participants may be recruited into the process. Some, such as public hearings, are open to all members of the public. They are publicized and interested parties are invited to attend. In other processes, such as consensus conferences and citizen advisory committees, participants are carefully selected to represent the population concerned. The number of participants within a process has a significant impact on group dynamics. If a group is too small, there may not be enough input of information and ideas to explore the problem thoroughly and create a range of potential solutions. If a group is too large, there is less opportunity for all participants to contribute, and reticent participants may not get an opening to be involved at all. Large groups also restrict the level of interaction – for example, public hearings are most likely to be one-way information provision rather than two-way deliberation. In larger groups it is more difficult for close relationships to form, which may affect the depth of discussion (see the section titled “Group Dynamics”).

Structure

The structure of the participatory process will depend on the decision situation, and what works in one situation might not work in another. Flexibility is often required to allow for a change in direction or to incorporate new information. General information can be communicated through the press and media. It is a good way of informing a large population in a relatively short time frame. It should also be noted that often the media report what they want to report and this may not be a balanced view of the situation and may have a detrimental impact on public perception. Other methods of communicating information include advertisements, postal letters, and E-mail. Face to face interaction ensures that the information communicated is understood, and body language provides the more subtle benefits of the information. Discussion of a problem is the best way to resolve conflict, and brainstorming generates solutions to a problem. The way in which a meeting is conducted may have an impact on the outcome of the meeting. Web-based participation tools are increasingly being used to involve the public in the decision-making process. As with most things they have their advantages and disadvantages. On the positive side, web-based tools can reach a much larger and geographically diverse group; information is available 24/7; they allow people to comment and express their views anonymously and in a nonconfrontational manner. On the negative side, people may be overwhelmed by the tools and have insufficient technical knowledge to use them; there is no mechanism to determine if the participants have understood and interpreted the given information correctly; although more people are using the Internet, access to the tools is not guaranteed. Web-based tools should be used alongside more traditional techniques to provide full participation.

The structure of the decision process should be considered when thinking about the structure of the participatory processes to be adopted. One method of structuring a decision process is to break it into three main phases: *issue formulation*, where data and material are gathered to give a clear and comprehensive picture of the issues involved in the decision. This stage lays the foundations for reaching a solution to the decision problem and is very important to the decision-making process as a whole; the *analysis* phase, where further investigation is made into the issues that are brought up in the first phase. A number of alternative courses of action to solve the problem can be postulated. These alternatives may be explored further with models that can predict an outcome given certain conditions; finally, the *decision* or *evaluation* phase, where the alternatives can be considered by the decision makers and a choice between alternatives can be made. This three-stage process can be repeated to ensure that all of the issues and alternatives relevant to the decision are explored to ensure that the decision is requisite (see Figure 2).

Stakeholder participation may be involved in one or more of these phases of the decision process, depending on the objectives of the decision maker.

Group Dynamics

It is important to appreciate the dynamics of any group of people that are assembled together and the effect this might have on the process outcome. There are a number of factors to consider [36, 42]. The group size and the number of participants involved in the process have a significant impact on the outcome. The balance between individuality and group behavior varies depending on the size of the group.

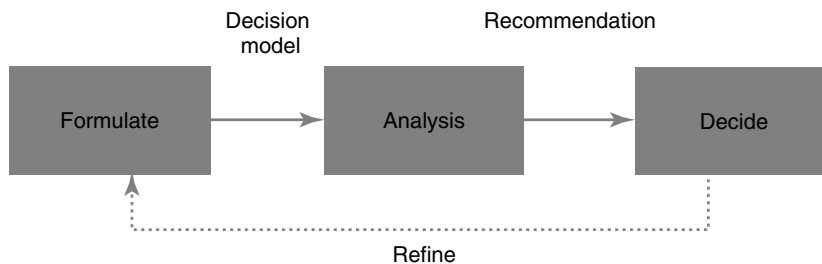


Figure 2 Three-stage methodology of decision [Reproduced from [41]. © The Institute for Operations Research and the Management Sciences (INFORMS), 1989.]

A group of 2–6 people can be classed as an “intimate group”, where individuality of participants is maintained. A group of 7–15 constitutes a “small group” in which individuality is preserved, and yet group processes emerge and exert considerable influence. A group of this size is small enough to work toward consensus but large enough to be representative. All participants have good opportunity to express their views, and differences of opinion can be used constructively to generate new perspectives. With groups of more than 15 participants, group processes dominate and individuality is submerged. The larger the group, the more difficult it becomes to reach consensus. Group cohesion is important, but maintaining some individuality will help the group remain creative. When the group is together, it has a character or personality as well as an emotional life. The emotional life affects how the task in hand gets done. An individual’s behavior may be influenced by the role that he/she has assumed, knowingly or unknowingly. Typical roles include the expert, the skeptic, the clown, the savior, the prophet, the critic, the mother, the protector, the warrior, and the leader. All these have an impact on the group life and interaction between participants. As participants begin to interact, differences between them (such as status, influence, role, and ability) may appear. As participants become more familiar with the decision situation, and the information they are provided with, barriers between them start to break down. McKenzie and Hunt describe the flow of conversation changes in a typical dialogue [39]. People begin by being polite with one another, both as a courtesy and as a way of gauging the reactions of others. Politeness breaks down into disagreements and individuals affirm their positions. As discussion progresses, participants question their positions, and begin to distance themselves from them. It is then possible to have dialogue that is open and not influenced by other agendas, where participants gain an understanding of others and the discussion can move forward. The process can be smoothed out and rendered more effective by the use of a neutral facilitator or moderator, at least in small groups [32, 36, 43]. Factors in group dynamics, such as flow of communication, perception of status, and emergence of leadership, are influenced by the physical position of group members, their distance apart, and their body orientations. These are affected by the shape and size of the room where a group meets,

and the spatial arrangement of chairs and tables. A lack of tables may be threatening to some but usually encourages openness and informality. Equally spaced chairs in a circle create a level playing field for all participants. When the environment and the arrangements for using it are focused, the group focuses more readily.

Other Important Characteristics

Participants may only be willing to contribute if the information they supply is kept anonymous and/or confidential. It should be established early in the introductory phase that such wishes will be respected. In some cases, the source of information may add weight to an argument. Permission must be granted before a source is identified. Ownership of the process and who runs it are important. The process is more transparent and legitimate if it is run by an independent facilitator, rather than someone who has some involvement or interest in its outcome. However, the authority usually has ultimate responsibility for the outcome and its implementation. For transparency, it should be clarified to all participants as to what will be done with the information they provide. Publishing a summary of the process can help people reflect, help with understanding of the issues, and help build consensus [44]. If comments are to be summarized and published, it should be considered which body is the appropriate body to handle the information. An authority may face criticism for protecting its own interest if they publish an outcome that is favourable to themselves. For example, in an online Environmental Protection Agency (EPA) (of the USA) participation process, the EPA was criticized for downplaying negative comments [4, 8]. The summary would be more legitimate if information is handled by a neutral facilitator.

Public Participation Processes

There is a wealth of literature describing various public participation processes. Table 1 describes some examples of conventional participation processes, and Table 2 gives some examples of some electronic (e-) participation tools. To achieve a set of objectives for the participation exercise, a combination of both conventional and e-participation tools can be

Table 1 Some examples of conventional participatory processes [3, 8, 11, 14, 23, 36, 42]

Participatory process	Description	Objectives
Referenda	A vote. Voters choose from available options and each vote cast has the same weight – all participants have equal influence. The vote is open to all members of the eligible voting population. Can be used in a number of situations to voting in a new government to community-based decision making.	Referenda should be representative, democratic, and acceptable to the population. The demand on resources will vary depending upon the context. Referenda do not necessarily produce structured decision making or a balance or risks or benefits as a vote can be cast without deliberation.
Public hearings/enquiries	A forum to inform the public about a particular situation; this could be a consultation on changes to regulation or a local development.	Public hearings are for sharing information. They are often viewed as a public relations exercise and have little influence on the outcome of decision making.
Consensus conference	A group of 10–16 participants representative of the population is brought together with experts in the area of the topic being discussed. The discussion is facilitated and participants discuss and try to reach consensus. The meetings are open to the public. The outcome can directly influence the development of legislation.	Difficult to represent the population with a group of 10–16 participants. It goes some way to involving the public in decision making that is well structured and of good quality; therefore, the outcome can be trusted.
Citizens jury	A group of 12–24 participants representative of the population is selected to meet over several days as part of a jury. The lay panel question expert witnesses mediated by a facilitator. Juries are not open to the public. The purpose is to reach a decision or formulate a set of recommendations.	Difficult to be representative of the population with 12–24 participants. The outcome is well-structured, quality decision making. It has a relatively high demand on resources.
Workshops	A facilitated workshop is an intensive session or working meeting attended by 7–15 stakeholders with different fields of expertise to discuss and share information.	Good for determining what the issues are related to the discussion. Good for developing a shared understanding. Not representative.
Focus groups	Used to gather opinions and attitudes about general topics. Groups of 5–12 people have free discussion. No decision-making role for participants.	Representative of specific groups; different groups can bring valuable information on the topic. All about information gathering.

complimentary. For example, an in-depth discussion to formulate the issues with a small group of individuals followed by a wider discussion over the web will increase representation of different groups. It is not expected that all objectives will be achieved with one process or that all objectives are important in all contexts. While there are a large number of studies focused on the performance of individual

mechanisms, there has been very little comparison of different participatory processes in the literature to determine which participatory tool works best for a given scenario. Bayley and French [47] attempted to make a comparison between processes in terms of how they met a list of objectives. However, as noted in the paper, the judgments of effectiveness are based on what little evidence could be gathered

Table 2 Some examples of electronic participatory tools and processes [21, 45, 46]

Participatory process	Description	Objectives
Online discussion forums	These enable an unlimited number of people to discuss one or more topics asynchronously.	Are anonymous so the discussant cannot be identified. Inappropriate additions to the forum can be censored, however this might bring into question the democratic values of the process.
E-voting	Online voting	Technology still being developed to ensure that the votes are cast by those that are eligible, and that the vote is secure.
E-polling	Online opinion polling	Can be useful to gauge opinions of the online community on particular matters.
Geographical information systems (GIS)	Provides an ideal infrastructure to debate geographically referenced issues. Such systems can be web enabled to allow the public and stakeholders to take part in a planning process and enhance participation in the decision-making process.	High resource demand. Possibility that people will interpret the information in different ways. Needs to be combined with an effective human–computer interface. Can be very useful for people to picture the geographics of a situation and have an input.

from the literature, from experience, and occasionally little more than common sense – the comparison was based on judgment rather than empirical evidence.

The task of the designers of a participatory process is to consider the situation and identify the objectives that matter to them in that context along with their relative importance. The appropriateness of such a process is context specific. For instance, in the case such as the withdrawal of an exemption license for a pesticide, there would be little urgency and the timescale of the participatory process could be relatively relaxed, whereas dealing with a situation that was potentially a threat to health would have more urgency. In the former emphasis would surely be on exploring public and stakeholder values and checking against a wider set of perceptions for possible unanticipated consequences, whereas information provision and advice would have a higher importance in the latter. Thus the design process can be seen essentially as a standard decision analytic process where the analysis is conducted using an additive multiattribute value function. A design process may follow these steps:

1. discuss and identify the issues and concerns, exploring the context from a variety of perspectives, for instance, using soft modeling techniques to facilitate discussion [48, 49];
2. discuss and identify the objectives' hierarchy describing the authority's goals for the context [48];

3. develop a number of alternative processes, which draw on a variety of mechanisms to meet the objectives;
4. score the alternative processes against each objective [50];
5. evaluate the relative importance of the objectives through appropriate weights, e.g., using swing weighting [50];
6. form overall weighted scores for each alternative process, thus ranking them;
7. explore the sensitivity of the resulting ranking to changes in scores and weights [51–53].

More complex evaluations introducing key uncertainties are also possible [51, 53, 54].

Case Studies Involving Stakeholder Participation

The Lake Päijänne Case

Lake Päijänne is the second largest lake in Finland and has been regulated since 1964 to increase hydropower production and decrease agricultural flood damages [5]. The lake has extensive recreational housing developments along the shores and there are thousands of recreational users and fishermen on the lake. Problems currently recognized on Lake Päijänne include the low water levels during spring, changes in the littoral zone vegetation, and

negative impacts of the regulation on the reproduction of fish stocks.

An extensive multidisciplinary research project was carried out between 1995 and 1999 to reevaluate the regulation policy of Lake Päijänne. The aims of the project were to assess the ecological, economic, and social impacts of the regulation. Stakeholder's opinions were sought about the current regulation and its development, comparison of new regulation policy options, and recommendations to diminish the harmful impacts of the regulation. At the beginning of the project there was wide distrust, especially among fishermen, toward the project and the organization responsible for it. Therefore, an open and participatory planning process was considered to be necessary in order to gain public support for the project and to find a consensus solution for further regulation strategy. The overall objective was to achieve consensus on the recommendations for the regulation policy through improved understanding of the complex nature of the regulation problem and of the other interest groups' preferences.

A number of participatory mechanisms were utilized to involve stakeholders in deliberation. A steering group consisting of 18 representatives of different stakeholders was set by the Ministry of Agriculture and Forestry. Four working groups were established to improve the communication between the water resource authorities, local stakeholders, experts on regulation, and researchers. To inform the public, a local press conference was arranged after almost every steering group meeting. Five public hearings were arranged at the beginning of the project and five at the end of the project. A mail questionnaire was sent to over 2000 users of the watercourse to determine the quantity and quality of recreational use of the area, problems related to water use, and objectives for future regulation practice. In addition to this, web tools were used to support the process and enable people not attending the workshops to access the results. An essential part of the participation process was decision analysis interviews of the extended steering group.

The results of the interviews were applied to prioritize the objectives of regulation in the different water conditions. This information was further used when the final consensus recommendation for the sustainable regulation strategy for Lake Päijänne was created. The draft recommendations were compiled by the experts and authorities responsible for the

management of Lake Päijänne. These were presented, discussed, and revised in the steering group meetings and finally unanimously accepted by all representatives. The implementation of the recommendations was relatively easy because all major stakeholders committed to the outcome of the process.

The Case of UK Energy Policy

In 2002 the Department for Trade and Industry (DTI) carried out a wide consultation on UK energy policy [55]. The evaluation was to determine the energy-related goals that matter most to citizens in the United Kingdom, what their attitudes toward different energy sources are, and what measure they would like to see implemented to enhance these goals. The consultation process was organized in four streams: web-based expert consultations; survey-based consultation of stakeholders with special interests; focus groups; and deliberative workshops with members of the general public. The aim of the former three processes was to gather information about the main concerns from the respective groups. The aims of the deliberative workshops were to gain an in-depth understanding of the public's perspectives about future national energy policy; to determine the objectives that the public considers crucial for the provision of electricity, and the trade-offs between the objectives; to establish the public's attitude toward nuclear energy, renewable energy technologies, and energy efficiency enhancement; to determine ways to achieve these objectives, i.e., the roles of government, companies, and individuals; and to evaluate the processes of public debate, public information, and policy making in the future and how information changes the participants' attitudes to electricity generation.

The outcome of the deliberation was strong support for enhancing the use of renewable energy sources to produce electricity and for increasing energy efficiency while keeping an eye on energy costs. All participants wanted a change to current trends. Attitudes to nuclear power were split: some participants thought that it should be part of a mix of energy sources utilized and others wanted to end it as a matter of principle. The majority of participants could not foresee a solution to the nuclear waste problem. Diversity of energy sources was perceived as important for security of supply. Over the course of the workshops, participants changed their minds

to some extent about which renewable energy technologies they preferred, given the information they received about costs and benefits of each option. Wave and wind power tended to rise in comparison with solar energy, probably reflecting the higher costs. Participants saw a major role for the government in changing behaviors; policy suggestions included measures to sell only goods that were energy efficient. Education of young people and adults was considered to be the most effective way to encourage households to use less energy. There was also clear support for the introduction of higher taxes to encourage people to purchase energy-efficient goods.

Public Participation Using Web-Based Geographical Information Systems (GIS): Local Planning in Slaithwaite

The Neighbourhood Initiatives Foundation (NIF) in the area of Slaithwaite in West Yorkshire, UK, aimed to involve local people in local environmental planning problems and decision making. The local community created a large-scale map of the area onto which local people put flags with their ideas. An event “Shaping Slaithwaite” was held at the village fair, where people were given the opportunity to have their say on their community. Alongside this, a web-based geographical information systems (GIS) map was created to allow wider participation [45]. GIS is used to help people visualize problems that involve a special element using electronic maps.

The web-based system was found to be popular; however, the user group was not representative of the population. Many more males than females used the system; occupation information collected suggested a strong weighting toward professional/managerial and educational positions; and the age range showed a heavy skew toward school children (due to educational uses of the web tool). Other problems encountered include the lack of familiarity with the technology involved and the lack of Internet access to all sectors of the population.

Summary

Participation processes can be adopted to involve stakeholders in decision-making processes, including the management of risk and the development of risk mitigation strategies, as demonstrated in the case studies.

References

- [1] Arnstein, S.R. (1969). A ladder of citizen participation, *American Institute of Planners Journal* **35**, 216–224.
- [2] Asaro, P.M. (2000). Transforming society by transforming technology: the science and politics of participatory design, *Accounting, Management and Information Technologies* **10**, 257–290.
- [3] Beirele, T. (1999). Using social goals to evaluate public participation in environmental decisions, *Policy Studies Review* **16**(3/4), 75–103.
- [4] Beirele, T. & Cayford, J. (2002). *Democracy in Practice: Public Participation in Environmental Decisions*, Resources for the Future, p. 208.
- [5] Mustajoki, J., Hämäläinen, R.P. & Marttunen, M. (2002). *Participatory Multi-criteria Decision Analysis with Web-HIPRE: A Case of Lake Regulation Policy*, Systems Analysis Laboratory, Helsinki University of Technology, Helsinki, at <http://www.sal.hut.fi/Publications/pdf-files/mmusb.pdf>.
- [6] Regan, P.J. & Holtzman, S. (1995). R&D advisor: an interactive approach to normative decision system model construction, *European Journal of Operational Research* **84**, 116–133.
- [7] Sheppard, S.R.J. & Meitner, M. (2005). Using multi-criteria analysis and visualisation for sustainable forest management planning with stakeholder groups, *Forest Ecology and Management* **207**(1–2), 171–187.
- [8] Abelson, J., Forest, P.G., Eyles, J., Smith, P., Martin, E. & Gauvin, F.P. (2003). Deliberations about deliberative methods: issues in the design and evaluation of public participation processes, *Social Science and Medicine* **57**, 239–251.
- [9] Beirele, T. (2002). The quality of stakeholder-based decisions, *Risk Analysis* **22**(4), 739–749.
- [10] Beirele, T. & Konisky, D. (2000). Values, conflict, and trust in participatory environmental planning, *Journal of Policy Analysis and Management* **19**, 587–602.
- [11] Chess, C. & Purcell, K. (1999). Public participation and the environment: do we know what works? *Environmental Science and Technology* **33**(16), 2685–2692.
- [12] Renn, O. (1998). The role of risk communication and public dialogue for improving risk management, *Risk Decision and Policy* **3**(1), 5–30.
- [13] Renn, O. (1999). A model for an analytic-deliberative process in risk management, *Environmental Science and Technology* **33**(18), 3049–3055.
- [14] Rowe, G. & Frewer, L. (2000). Public participation methods: a framework for evaluation, *Science Technology and Human Values* **25**(1), 3–29.
- [15] Rowe, G. & Frewer, L.J. (2005). A typology of public engagement mechanisms, *Science Technology and Human Values* **30**(2), 251–290.
- [16] Webler, T. (1995). Right discourse in citizen participation: an evaluative yardstick, in *Fairness and Competence in Citizen Participation: Evaluating Models for*

- Environmental Discourse*, O. Renn, T. Webler & P.M. Wiedemann, eds, Kluwer Academic, Dordrecht, Boston.
- [17] Webler, T. (1999). The craft and theory of public participation: a dialectical process, *Journal of Risk Research* 2(1), 55–71.
- [18] Webler, T., Tuler, S. & Kruger, R. (2001). What is a good public participatory process? Five perspectives from the public, *Environmental Management* 27(3), 435–450.
- [19] Irvin, R. & Stansbury, J. (2004). Citizen participating in decision-making: is it worth the effort? *Public Administration Review* 64(1), 55–65.
- [20] Rowe, G. & Frewer, L.J. (2004). Evaluating public-participation exercises: a research agenda, *Science Technology and Human Values* 29(4), 512–557.
- [21] Chappelet, J.-L. & Kilchenmann, P. (2005). Interactive tools for e-Democracy: examples from Switzerland, in *E-Government, Towards Electronic Democracy, International Conference, TCGOV 2005 Proceedings*, Bolzano, March 2–4, 2005, M. Böhlen, J. Gamper, W. Polasek & M.A. Wimmer, eds, Springer-Verlag GmbH, pp. 36–47.
- [22] Fischhoff, B. (1995). Risk perception and communication unplugged: twenty years of process, *Risk Analysis* 15, 137–145.
- [23] Renn, O., Webler, T. & Wiedemann, P.M. (1995). Fairness and competence in citizen participation: evaluating models for environmental discourse, *Technology, Risk, and Society*, Kluwer Academic, Dordrecht, Boston, Vol. 10.
- [24] Slovic, P. (1993). Perceived risk, trust and democracy, *Risk Analysis* 13, 675–682.
- [25] Susskind, L. & Field, P. (1996). *Dealing with an Angry Public: The Mutual Gains Approach*, Free Press, New York.
- [26] Poortinga, W., Bickerstaff, K., Langford, I., Niewöhner, J. & Pidgeon, N.F. (2004). The British 2001 foot and mouth crisis: a comparative study of public risk perceptions, trust, and beliefs about government policy in two communities, *Journal of Risk Research* 7(1), 73–90.
- [27] Bennett, P.G. & Calman, K.C. (eds) (1999). *Risk Communication and Public Health: Policy Science and Participation*, Oxford University Press, Oxford.
- [28] House of Lords (2000). *Science and Society*, Select Committee on Science and Technology, Third Report, London.
- [29] Langford, I., Marris, C. & O’Riordan, T. (1999). Public reactions to risk: social structures, images of science and the role of trust, in *Risk Communication and Public Health Policy, Science and Participation*, P.G. Bennett & K.C. Calman, eds, Oxford University Press, Oxford, Vol. 33–50.
- [30] Leach, M., Scoones, I. & Wynne, B. (2005). *Science and Citizens*, Zed Books, London.
- [31] Berry, D.C. (2004). *Risk Communication and Health Psychology*, Open University Press, Maidenhead.
- [32] Eden, C. & Radford, J. (eds) (1990). *Tackling Strategic Problems: The Role of Group Decision Support*, Sage, London.
- [33] Monplaisir, L. (2002). Enhancing CSCW with advanced decision making tools for an agile manufacturing system design application, *Group Decision and Negotiation* 11, 45–63.
- [34] Mumford, E. (2003). *Redesigning Human Systems*, Idea Group Publishing.
- [35] Nunamaker, J.F., Applegate, L.M. & Konsynski, B.R. (1988). Computer-aided deliberation: model management and group decision support, *Operations Research* 36, 826–848.
- [36] Phillips, L.D. & Phillips, M.C. (1993). Facilitated work groups – theory and practice, *Journal of the Operational Research Society* 44(6), 533–549.
- [37] Winn, M.I. & Keller, L.R. (2001). A modelling methodology for multi-objective multi-stakeholder decisions: implications for research, *Journal of Management Inquiry* 10(2), 166–181.
- [38] Duggan, E.W. (2003). Generating systems requirements with facilitated group techniques, *Human-Computer Interaction* 18, 373–394.
- [39] McKenzie, D. & Hunt, J. (2001). *Riscom II Deliverable 4.5: Designing Dialogue*, in *Riscom II*, University of Lancaster, Lancaster.
- [40] MORI (2003). *Exploring Trust in Public Institutions*, Report for the Audit Commission.
- [41] Holtzman, S.S. (1989). *Intelligent Decision Systems*, Addison-Wesley, Reading.
- [42] Sinkko, K. & Hamalainen, R.P. (2005). *Facilitated Workshops – A Participatory Method for Planning of Countermeasures in Case of a Nuclear Accident*, Helsinki University of Technology.
- [43] Ackermann, F. (1996). Participants perceptions on the role of facilitators using group decision support systems, *Group Decision and Negotiation* 5, 93–112.
- [44] DETR (2000). *UK Strategy for Radioactive Discharges 2001–2020 (Consultation Document)*.
- [45] Carver, S., Evans, A., Kingston, R. & Turton, I. (2001). Public participation, GIS, and cyberdemocracy: evaluating online spatial decision support systems, *Environment and Planning B: Planning and Design* 28, 907–921.
- [46] Prosser, A. & Krimmer, R. (eds) (2004). *Electronic Voting in Europe – Technology, Law, Politics and Society: Proceedings of the Workshop of the ESF TED Programme Together with GI and OCG, Lecture Notes in Informatics*, Gesellschaft für Informatik, Bonn.
- [47] Bayley, C. & French, S. (2007). Designing a participatory process for stakeholder involvement in a societal decision, *Group Decision and Negotiation*, ISSN: 0926–2644 (Print) 1572–9907 (Online) (Currently online).
- [48] French, S., Carter, E. & Niculae, C. (2007). Decision support in nuclear and radiological emergency situations: are we too focused on models and technology? *International Journal of Risk Assessment and Management* 4(3), 421–441.

-
- [49] Rosenhead, J. & Mingers, J. (eds) (2001). *Rational Analysis for a Problematic World Revisited*, John Wiley & Sons, Chichester.
- [50] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*, John Wiley & Sons, New York.
- [51] Belton, V. & Stewart, T.J. (2002). *Multiple Criteria Decision Analysis: An Integrated Approach*, Kluwer Academic Press, Boston.
- [52] French, S. (2003). Modelling, making inferences and making decisions: the role of sensitivity analysis, *Sociedad de Estadística e Investigación Operativa* **11**(2), 229–251.
- [53] Goodwin, P. & Wright, G. (2003). *Decision Analysis for Management Judgement*, 3rd Edition, John Wiley & Sons, Chichester.
- [54] Keeney, R.L. (1992). *Value-Focused Thinking: A Path to Creative Decision Making*, Harvard University Press.
- [55] Stagl, S. (2006). Multicriteria evaluation and public participation: the case of UK energy policy, *Land Use Policy* **23**, 53–62.

Related Articles

Players in a Decision

Risk Attitude

Scientific Uncertainty in Social Debates Around Risk

Stakeholder Participation in Risk Management Decision Making

CLARE L. BAYLEY

Quantitative Reliability Assessment of Electricity Supply

The basic function of an electric power system is to satisfy its customer load demands as economically as possible and with a reasonable assurance of quality and continuity. Evaluation of the level of assurance of quality and continuity is known as *reliability assessment* [1]. Figure 1 shows the three functional zones of generation, transmission, and distribution that constitute an electric power system [2].

In a vertically integrated system, these facilities are owned and operated by a single entity. In the deregulated environment, there could be many owners and subsystem operators, and therefore, overall long-range planning has some difficult challenges. Power systems normally contain redundant facilities in the various functional zones to provide a reasonable level of reliability of supply to the customer, and the economic and reliability constraints can be quite competitive. This is not a new problem and has existed since the creation of electric power supply systems.

The functional zones shown in Figure 1 can be combined to create hierarchical levels to facilitate system analysis, planning, and operation. Hierarchical level one (HLI) involves the generation facilities, and reliability analysis at this level is concerned with the ability of the system to generate sufficient power to meet the overall electrical power and energy demands. Hierarchical level two (HLII) involves the composite generation and transmission system (bulk electric system), and analysis at this level is concerned with the ability of the system to deliver the required energy to the bulk supply points. The hierarchical level three (HLIII) includes all the system facilities. Detailed quantitative analysis at HLIII is not normally conducted because of the enormity of the problem in a practical system. Detailed analysis is, however, usually conducted within the distribution functional zone. The actual level of redundancy in the physical facilities decreases as the focus moves through the generation, transmission, and distribution functional zones, and most load points are served by radial networks at the actual customer level. This could include extensive

manual or automatic switching provisions to decrease the duration of supply failures but the primary supply is still radial in nature. Statistics produced by the Canadian Electricity Association (CEA) indicate that in Canada, 82% of the interruptions seen by a customer in 2005 occurred because of failures in the distribution functional zone [3]. The overall level of reliability is quite high and any proposed improvement in reliability must be balanced with the commensurate economic cost of attaining the improvement. Considerable work has been done in the area of reliability cost/worth analysis in attempts to determine how reliable should electricity supply be in a modern society [1].

Quantitative reliability assessment of electricity supply is an important area of activity in modern power systems. The criteria and the resulting techniques to assess and incorporate reliability concerns in the planning and operation of electric power systems were, initially, all deterministically based and some of them are still in use today. Percent reserve margin and the loss of the largest unit criteria were used to determine the required generating capacity at HLI. Transmission system analysis is still largely based on the ability of the system to meet the peak demand with any one transmission element out of service. This is commonly known throughout the world as the $n - 1$ criterion and is sometimes extended to include additional elements, i.e. $(n - m)$. The system investment costs and the system reliability, however, increase as m increases, and again it becomes a question of how reliable should the system be. Deterministic techniques are usually relatively simple to appreciate and to apply. They are, however, quite inconsistent, as not all elements have the same likelihood of failure or have the same impact on the system. Deterministic techniques, in their basic form, do not and cannot incorporate the inherent probabilistic or stochastic nature of system behavior, such as component failures, adverse weather, and future customer load demands. This has been understood and appreciated for many years and considerable effort has been devoted to the development of probabilistic techniques that respond to the actual factors that influence the reliability of electric power supply. A considerable number of publications available on this subject are illustrated in [4–10].

Most electric power utilities collect considerable system and component data to monitor past reliability performance and to provide the ability to predict

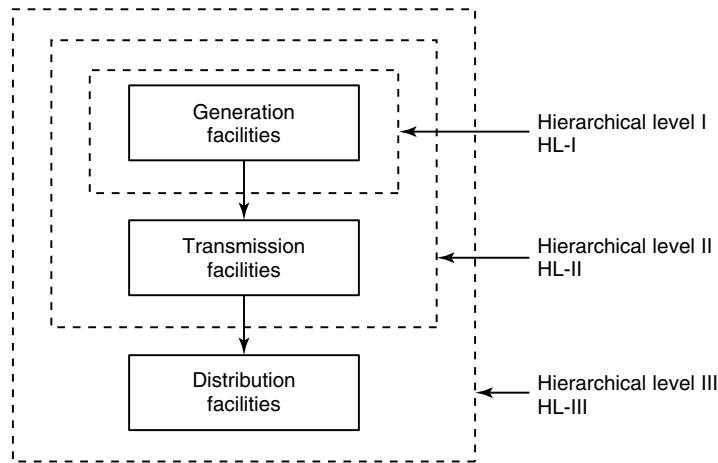


Figure 1 Functional zones and hierarchical levels

the future performance. Collecting data is expensive but necessary, as these data are the foundation for past assessments and future estimates of the system's ability to serve its customers. These data are used in the probabilistic analysis of existing and proposed system facilities and provide valuable input to system expansion planning and the resulting decision-making processes. In Canada, generation and transmission component data are collected, analyzed, and disseminated by the CEA using comprehensive and consistent protocols [11, 12]. Similar systems exist in other jurisdictions.

Quantitative power system reliability assessment can be conducted using a variety of methods. These methods are generally categorized as being either analytical or simulation techniques. Most of the earlier work was done using analytical approaches. The increasing availability of high-speed computing facilities has, however, made simulation techniques a powerful and practical option for many studies. Analytical techniques represent the system and its components by mathematical models, which are used to compute the required reliability indices using direct numerical analysis. The obtained reliability indices are usually expected values. Simulation methods are usually designated as Monte Carlo simulation (MCS) (*see Numerical Schemes for Stochastic Differential Equation Models; Systems Reliability; Cross-Species Extrapolation; Bayesian Statistics in Quantitative Risk Assessment*) [13] and are used to estimate the reliability indices by simulating the random behavior of the individual system

components and the overall system. MCS can be divided into two general approaches. The state sampling approach can be used to determine the state (available or unavailable) of each component and the resulting state of the system. A suitably long series of trials is conducted and the resulting reliability indices are estimated. The sequential sampling approach examines the behavior and history of each component and the system in chronological order for a suitably long period of time. The reliability indices are then estimated from the system-simulated history. Durations of the up and down state of the component are determined using the probability density functions of failure and repair time. Theoretically, sequential MCS is a very powerful technique that can incorporate virtually all the aspects and contingencies associated with an actual system. It can be used to generate reliability index probability distributions in addition to the expected values. The required simulation time to analyze a large practical system can, however, be quite lengthy even with modern high-speed computing facilities.

Quantitative Reliability Assessment at HLI

A basic requirement in achieving acceptable reliability in a modern electric power system is that the system generating capacity exceeds the customer load requirements by a suitable margin. Reliability studies can be generally divided into the two categories of

adequacy and security analyses. The North American Electric Reliability Council (NERC) defines adequacy as “the ability of the electric systems to supply the aggregate electrical demand and energy requirements of their customers at all times, taking into account scheduled and reasonably expected unscheduled outages of system elements” [14]. Adequacy evaluation involves static or steady-state analysis, and dynamic conditions are normally incorporated in security evaluations. Generating capacity reliability evaluation is generally focused on the adequacy domain.

The most commonly used probabilistic generating capacity adequacy indices are the loss of load expectation (LOLE) and the loss of energy expectation (LOEE) [1]. The LOLE is the most popular index and is the expected number of hours or days in a given year that the system load will exceed the available capacity. The predicted risk indices are not applicable to individual customer load points and apply to the total load in the system under study. Most applications involve a daily peak load model rather than an hourly load model, and an annual period. The risk index in this case is, therefore, the expected number of days per year that the daily peak load exceeds the available capacity. A common criterion is 0.1 days yr^{-1} . The reciprocal of this index is “1 day in 10 years” and is often used to express the criterion. The risk can also be expressed as the loss of load probability (LOLP). An index of 1 day in 10 years does not imply that there will be a generating capacity shortage on average once in a 10-year period. The LOLE is not a frequency index. This is a common misconception and one that seems to exist in general application. Techniques are available to estimate the average frequency of capacity deficiencies using either analytical models or Monte Carlo simulation [1, 13]. These techniques are well developed but require considerably more system data than does the basic loss of load approach. The basic procedure used to assess generating system adequacy is relatively standard in concept and is based on the creation of suitable generation and load models [1]. The generation and load models are combined (convolved) to create a risk model.

There are certain basic factors that should be included in a comprehensive assessment of generation system adequacy. The following is a quotation taken from a recent NERC regional standard.

“Among the factors to be considered in the calculation of the probability are the characteristics of the loads, the probability of error in load forecast, the scheduled maintenance requirements for generating units, the forced outage rates of generating units, limited energy capacity, the effects of connections to the pools, and network transfer capabilities.”

The relatively recent rapid growth in renewable energy related generating facilities and the anticipated considerable future growth clearly dictate the need to accurately incorporate these facilities in generating capacity adequacy evaluation. Considerable attention is being given to wind turbine units at the present time. These units can be considered to be capacity limited owing to the diffuse and random nature of the wind. Conventional generating units, other than perhaps hydro units on a controlled single river system, are usually considered to be independent. This is not the case with wind turbine units in a wind farm. These units are basically all dependent on the same wind regime. This is an important element in the adequacy assessment of a wind generator farm. The degree of correlation between multiple wind farms is also an important factor in adequacy assessment of overall generating capacity.

The basic techniques used to conduct conventional generating capacity evaluation are extremely flexible and can be modified and extended to include new constraints, criteria, and requirements associated with the new utility environment. A key requirement, however, in quantitative reliability evaluation is the availability of consistent and comprehensive data on relevant equipment, load characteristics, and system constraints. Databases such as the equipment reliability information system (ERIS) maintained by the CEA and the generation availability data system (GADS) operated by NERC are valuable repositories of utility data and equipment reliability parameters. Similar systems exist throughout the world.

Quantitative Reliability Assessment at HLII

As noted earlier, the determination of how much generating capacity is required to provide a reasonable level of reliability is a fundamental element in power system planning. An equally important planning requirement is to develop a suitable transmission network to transport the required energy to the bulk

electric system supply points. The energy is then conveyed through the connected distribution facilities to the customer load points. The bulk transmission facilities must be capable of maintaining acceptable voltage levels, loadings within the thermal limits of the transmission facilities, and the system stability limits. The static assessment of the composite generation and transmission system's ability to satisfy the load requirements is designated as adequacy assessment and has been the subject of considerable development [4–10]. A wide range of probabilistic techniques have been developed and applied using both analytical methods and MCS. The bulk of the computation time is used in the solution and resulting examination of the system state under the outage states of various components. A wide range of load point and system indices can be calculated at HLII. The most popular indices are the probability of load curtailment (PLC) and the expected energy not supplied (EENS). These indices are similar to the LOLP and LOEE indices used in HLI assessment, respectively.

The energy-based indices (LOEE and EENS) can be extended to assess the expected customer interruption costs (ECOST) at individual load points and at the system level using an interrupted energy assessment rate (IEAR) expressed in dollars per kilowatt-hour of unserved energy. The ability to quantify the reliability of a bulk electric system provides the opportunity to estimate the reliability cost associated with a particular system configuration. The monetary cost of system unreliability can be included with the capital, operating, and maintenance costs in the examination of system planning options.

Quantitative Reliability Assessment at HLIII

There is relatively little practical application of quantitative predictive reliability assessment at HLIII, although some research has been conducted in this area. The basic approach is to utilize the indices obtained at a bulk system load point as supply indices in series with the actual distribution system. Quantitative assessment of past reliability performance at HLIII is conducted by many distribution utilities throughout the world. The resulting indices are known as *service continuity statistics* in North America, and are collected through the CEA [3] in Canada. The basic service continuity indices used throughout

North America and in many other places in the world are as follows:

SAIFI: system average interruption frequency index;

SAIDI: system average interruption duration index;

CAIDI: customer average interruption duration index;

IOR: index of reliability.

The aggregated statistics for Canadian electric power utilities are typical of the service continuity levels in most developed countries around the world. The following indices are the annual values for 2005:

$SAIFI = 2.13$ interruptions/system customer;

$SAIDI = 4.80$ h/system customer;

$CAIDI = 2.26$ h/customer interruption;

$IOR = 0.999452$.

Quantitative Reliability Assessment – Distribution

Predictive techniques for reliability assessment in the distribution functional zone are highly developed [1]. Both analytical and MCS techniques have been applied to estimate the reliability at individual load points and the service continuity statistics used to measure past performance. The focus at the customer level is on physically based indices that can be readily interpreted by the customer. The most common indices are the frequency of failure, the duration of failure when it occurs, and the unavailability of supply in hours per year. Analytical techniques are used to provide the expected value of these parameters. Sequential MCS can be used to determine the expected values and the probability distributions associated with the annual performance. These techniques have been extended to include the customer interruption costs associated with power supply failures. The ECOST values are used in reliability cost/worth analyses associated with system modifications, reinforcements, operating considerations, and long term planning. Many countries around the world have conducted intensive surveys of their customers to determine the cost of power failures [15]. The resulting data have been used in a wide range of tasks including reliability cost/worth evaluation, optimum reliability level determination, and penalty payments for supply failures.

References

- [1] Billinton, R. & Allan, R.N. (1996). *Reliability Evaluation of Power Systems – Second Edition*, Plenum Press, New York.
- [2] Billinton, R. & Tatla, J.S. (1983). Hierarchical indices in system adequacy assessment, *CEA Transactions* **22**, 1–14.
- [3] Canadian Electricity Association (2006). *Annual Service Continuity Report on Distribution System Performance in Electrical Utilities – 2005*, Canadian Electricity Association.
- [4] Billinton, R. (1972). Bibliography on the application of probability methods in power system reliability evaluation, *IEEE Transactions* **PAS-91**, 549–560.
- [5] IEEE Subcommittee Report (1978). Bibliography on the application of probability methods in power system reliability evaluation, 1971–77, *IEEE Transactions* **PAS-97**, 2235–2242.
- [6] Allan, R.N., Billinton, R. & Lee, S.H. (1984). Bibliography on the application of probability methods in power system reliability evaluation, 1977–82, *IEEE Transactions* **PAS-103**, 275–282.
- [7] Allan, R.N., Billinton, R., Shahidehpour, S.M. & Singh, C. (1988). Bibliography on the application of probability methods in power system reliability evaluation 1982–87, *IEEE Transactions on Power Systems* **3**(4), 1555–1564.
- [8] Allan, R.N., Billinton, R., Briepohl, A.M. & Grigg, C.H. (1994). Bibliography on the application of probability methods in power system reliability evaluation, 1987–1991, *IEEE Transactions on Power Systems* **9**(4), 275–282.
- [9] Allan, R.N., Billinton, R., Briepohl, A.M. & Grigg, C.H. (1999). Bibliography on the application of probability methods in power system reliability evaluation, 1992–1996, *IEEE Transactions on Power Systems* **14**(1), 51–57.
- [10] Billinton, R., Fotuhi-Firuzabad, M. & Bertling, L. (2001). Bibliography on the application of probability methods in power system reliability evaluation, 1996–1999, *IEEE Transactions on Power Systems* **16**(4), 595–602.
- [11] Canadian Electricity Association (2005). *Generation Equipment Status, 2004 Annual Report*, Ottawa.
- [12] Canadian Electricity Association (2005). *Forced Outage Performance of Transmission Equipment, 2004 Annual Report*, Ottawa.
- [13] Billinton, R. & Li, W. (1994). *Reliability Assessment of Electric Power Systems Using Monte Carlo Methods*, Plenum Press, New York.
- [14] North American Electric Reliability Council (2005). *NERC Planning Standards*, at www.nerc.com/glossary/glossary-body.html.
- [15] CIGRE Task Force 38.06.01 (2001). *Methods To Consider Customer Interruption Costs In Power System Analysis*, CIGRE Publications, Paris.

ROY BILLINTON

R&D Planning and Risk Management

Research and development (R&D) projects are inherently risky. Investors in R&D projects are often uncertain about the probability of eventual technical success, the costs and time needed to succeed (if success is possible), competitor activities, and market acceptance and financial rewards if a technical success occurs. Quantitative risk models can help characterize and manage R&D risks and answer such practical risk-management decision questions the following:

- *R&D project scheduling and management*: what part(s) of a project to work on first, how to order the activities within a project, decision rules for when to abandon a project
- *R&D project selection*: which projects to fund?
- *R&D investment strategy*: how intensely or quickly to pursue R&D?
- *R&D and intellectual property*: how to manage the results of R&D (e.g., through licensing and patent systems) to create incentives that will benefit society?

This article surveys methods for quantifying and managing R&D risks.

Minimizing R&D Project Risks by Optimally Ordering Activities

A very simple model of an R&D project divides it into multiple activities, each having a known cost to attempt it and a known probability that it can be successfully completed if it is attempted. An R&D manager may wish to minimize the expected cost to either successfully complete the whole project or to learn (by trying activities and learning that they fail) that it cannot be completed. The problem of sequencing the project activities to achieve this goal can be solved easily in certain cases.

Example: Optimal Sequencing of Activities in an R&D Project

Setting: Suppose that an R&D project can be completed successfully only by successfully completing

both of two activities, A and B. The cost to attempt activity A is $c_A = 1$ cost unit; the cost to attempt B is $c_B = 2$ cost units; and the conditional probabilities of successfully completing these activities, if they are attempted, are $p_A = 0.8$ and $p_B = 0.5$, respectively. The two activities may be attempted in either order. The value of successfully completing the project (i.e., of successfully completing both A and B) is $V = 10$ benefit units (on a scale commensurable with the cost units).

Problem: (i) What investment strategy maximizes the expected value of investment in this set of R&D activities? (ii) What is the expected value of this R&D project?

Solution: (i) Attempting A first, and then B if A succeeds, yields an expected net value of (cost to attempt A) + (probability A succeeds) $\times$ (cost to attempt B after A succeeds + probability that B succeeds $\times$ value obtained if B and A succeed) which is given by

$$\begin{aligned} & -c_A + p_A \times (-c_B + p_B \times V) \\ & = -1 + 0.8 \times (-2 + 0.5 \times 10) = 1.4 \text{ units} \quad (1) \end{aligned}$$

Symmetrically, attempting B first, and then A if B succeeds, yields an expected net value of

$$\begin{aligned} & -c_B + p_B \times (-c_A + p_A \times V) \\ & = -2 + 0.5 \times (-1 + 0.8 \times 10) = 1.5 \text{ units} \quad (2) \end{aligned}$$

Since 1.5 is the higher value (and is positive), the optimal strategy is to attempt B before A, and the expected monetary value of the R&D project (if the optimal strategy is used) is 1.5.

More generally, trying A before B yields a higher expected value than trying B before A if and only if

$$\begin{aligned} & -c_A + p_A \times (-c_B + p_B \times V) > -c_B + p_B \\ & \quad \times (-c_A + p_A \times V), \text{ or, equivalently,} \\ & \text{if } c_B \times (1 - p_A) > c_A \times (1 - p_B) \quad (3) \end{aligned}$$

That is, for a project with positive expected value, activities should be attempted in order of decreasing failure probability per unit cost ratio (i.e., in order of decreasing $c_i/(1 - p_i)$, where $i = A$ or B in this example); this decision rule holds for any number of activities. [Projects with negative expected value should not be attempted at all, if the goal is to maximize expected value. The expected value of

a sequence of N consecutive activities is found by summing over all activities 1 through N , the probability of failing for the first time at that activity times the net financial loss from doing so, and then subtracting that number from the overall success probability (which is the product of the success probabilities of all activities in the sequence) times the net benefit ($= V - \text{sum of activity costs}$), if the project succeeds.]

The preceding example is especially simple because completing the project required all of the activities in it to be completed. The only decision was the order in which to attempt them. More generally, there may be several alternative ways to successfully complete an R&D project. A subset of activities is called a *path set* if completing all of the activities in it suffices to successfully complete the project. A *minimal path set* is one in which no proper subset of activities is a path set. The project succeeds if and only if all of the activities in at least one minimal path set are successfully completed. If minimal path sets are disjoint, then an optimal (expected value maximizing) investment strategy is to attempt alternative minimal path sets in decreasing order of success probability per unit expected cost (where the success probability for a minimal path set is the product of the success probabilities of the activities in it and the expected cost of the minimal path set is calculated assuming that activities *within* each minimal path set are sequenced in decreasing order of failure probability per unit cost ratio, as in the preceding example). Investment in a project should cease as soon as no minimal path sets with positive expected value remain.

These ideas can be extended to projects in which activities cannot be attempted in any arbitrary order. Suppose that precedence constraints among activities restrict the order in which they may be attempted: it may be necessary to complete some before others can be attempted. Suppose that these constraints are represented by some acyclic graph, where each edge represents an activity and each activity can be attempted if and only if all of its predecessors have been successfully completed. The project succeeds and its benefit is obtained, if and only if at least one path is successfully completed from a starting activity (one with no predecessors) to an ending activity (one with no successors). In this quite general framework, the optimal (expected utility maximizing) investment strategy for a decision

maker with linear or exponential utility can still be found by assigning to each activity a number called its *index* (generalizing the failure probability per unit cost ratio in the preceding example) that can be computed quickly from the cost to attempt, success probability, and precedence constraint data. The optimal decision rule is to *attempt activities in order of decreasing index until either success is achieved or no activities with positive indices remain* [1]. Many other extensions of index policies (also called *Gittins index* policies) have been developed for situations with discount rates, randomly evolving opportunity sets, and relatively sophisticated models of risky projects, such as Markov decision processes and multiarm bandit decision problems [2, 3].

Optimizing R&D Project Portfolios

Optimally sequencing the activities within an R&D project can minimize the expected cost to complete it or to discover that it cannot be completed; more generally, it can maximize the expected utility of investing in the project. But even after such within-project optimization, the problem of choosing a subset of available R&D projects to fund – often from many proposals – remains. Quantitative approaches to optimizing investment in proposed risky R&D projects to maximize resulting net benefits over time include the following:

1. Combinatorial optimization models

These search for combinations of select and reject decisions to maximize the net value of the selected portfolio of projects, given budget constraints and possible interactions and dependencies among the projects. Large-scale, nonlinear, 0–1 programming, and mixed integer programming models have been developed for this purpose and are practical with current optimization technology and computational resources [4, 5].

2. Improving an existing portfolio

Ringuest *et al.* [6] consider the financial impacts of changing an existing portfolio by adding or deleting projects, taking into account how such additions and deletions shift the predicted (simulated) probability distribution of net return from the entire project portfolio and taking into account interactions among projects.

3. Continuous dynamic optimization

Loch and Kavadias [7] propose a dynamic model of resource allocation for projects that may have statistically dependent, dynamically evolving market payoffs, assuming that partial investment in a project can generate partial returns. In this setting, explicit optimal investment formulas can be developed and interpreted in terms of equating appropriately defined marginal returns across investments in different projects.

In general, optimization of R&D project portfolios when there are strong interactions among projects (e.g., because of shared subsets of activities, or interdependencies among their technical success probabilities, and/or market values if successfully completed) provides a rich set of challenging problems for sophisticated, computationally intensive, optimization-based approaches to risk management. Promising approaches include simulation optimization and robust and multicriteria optimization methods for portfolio selection, together with financial evaluation models that recognize the real option value of risky projects that can be abandoned if they become unattractive. The pharmaceutical and energy industries have developed and applied many such models [8]. From the practical point of view, R&D portfolio optimization models can require potentially burdensome input data on project investment opportunities and their interactions. To ease this burden, some recent work has explored the use of approximate (fuzzy) descriptions of potential future cash flows in assessing project portfolio values in a real options framework, leading to a fuzzy mixed integer programming model for optimization of R&D portfolios [5].

Owner versus Manager Incentives in R&D Project Management

In the simplest case of a single minimal path set and no precedence constraints, as we have seen, expected value is maximized by attempting the “most risky” or “most difficult” activities first, meaning those with the highest failure probability per unit cost ratios. This strategy minimizes the expected cost of discovering that success is impossible, if that is the case. (Since project success requires completing all activities, the expected net benefit from success

does not depend on the order in which activities are attempted; thus, focusing on minimizing the expected cost to discover failure, if success is not possible, is an appropriate goal.) However, rewarding a manager for successfully completed activities may create an incentive to postpone high-risk activities as long as possible, increasing the manager’s expected reward but increasing the expected cost of unsuccessful projects. R&D projects that should be abandoned may be continued too long if managers postpone attempting high-risk activities until ones that are more likely to succeed have been completed. Conversely, R&D managers may abandon projects too soon and/or may invest too little effort in making them succeed (from the owner’s perspective) when effort and success probability are the managers’ private information.

Such incentive problems (which are instances of moral hazard and principal–agent conflict problems) can be largely overcome by evaluating partially completed R&D projects using a real options framework and then compensating R&D managers according to these evaluations. More specifically, compensating R&D managers using appropriate adjusted residual income measures (with capital charge rates that are contingent on manager reports of estimated profitability, recognizing that these may be distorted by managers in an attempt to obtain greater compensation) allows an effective separation of R&D investment decisions by owners and effort allocation decisions by R&D managers. Such compensation rules optimally balance incentives to invest in risky R&D against incentives to abandon unfavorable projects in some specific models of owner and manager information and incentives [9].

Optimal Intensity of Investment in Competitive R&D Projects

When a company undertakes R&D to obtain a competitive advantage, time pressure from competition (especially in a “winner takes all” patent and intellectual property environment) may make it worthwhile to invest simultaneously in multiple alternative approaches (e.g., in multiple minimal path sets). This stochastically reduces the time, but stochastically increases the cumulative expenditure, until either success is achieved or no attractive paths forward remain.

Exactly how intensely a company should invest in R&D to maximize expected profit depends on

how intensely competitors invest. Hence, quantitative models of optimal investment intensity are often formulated as differential games in continuous time, with the competing firms deciding at each moment how intensely to invest in R&D based on information available so far. In such competitive R&D models, details of R&D project activities are typically suppressed, and R&D projects are represented by relatively simple models, e.g., functions showing how the cumulative probability (or, equivalently, the hazard function) for successful completion of a project varies with the cumulative amount invested. In these models, incremental investment purchases incremental success probability, and more rapid expenditures hasten the time until either success occurs, it becomes optimal to abandon the project, or no investment opportunity remains. Typical conclusions from such models are as follows:

- Competitive rivalry increases R&D spending by each competitor when the benefits of the first breakthrough are captured by the firm that achieves it [10].
- If patent protection is incomplete (e.g., reflecting positive externalities from increased knowledge) and/or there are opportunities for licensing, so that the benefits of innovation may be shared even by noninnovators, then the effects of competition on the overall pace of innovation are ambiguous. Only some firms may invest in R&D, and social welfare may be either increased or decreased compared to the “winner takes all” situation [11].

Such strategic models of competitive R&D investment and resulting project value complement *real options* approaches that treat competitor actions as exogenously specified and then optimize a firm’s abandonment decisions (i.e., choice of when and whether to exercise the “option” to abandon). Such models typically assume that remaining cost to complete is uncertain and evolves according to a random process as R&D proceeds [12].

Conclusions

R&D portfolio selection, project management, and incentive design problems pose challenging practical applications for quantitative risk-management methods. The stakes for effective risk-management

methods in R&D are high, as the same set of risky R&D project opportunities can lead to either profits or losses, depending on how well the risks are managed through project selection and management.

Although sophisticated quantitative methods for modeling and managing R&D risks have been developed, as summarized in this article and pursued further in the references, applying such methods in practice, with realistically limited and uncertain data and estimates of potential future R&D costs and benefits, continues to be challenging. Yet, it is of great importance in areas such as pharmaceuticals, high technology, and energy, where portfolios of risky projects often drive company value. Therefore, further development of practical methods for R&D risk characterization and risk-management decision making will probably remain an active and rewarding area of both theoretical and applied risk management in the foreseeable future.

References

- [1] Denardo, E.V., Rothblum, U.G. & van der Heyden, L. (2004). Index policies for stochastic search in a forest with an application to R&D project management, *Mathematics of Operations Research* **29**(1), 162–181, DOI: 10.1287/moor.1030.0072.
- [2] Tsitsiklis, J.N. (1994). A short proof of the Gittins index theorem, *The Annals of Applied Probability* **4**(1), 194–199.
- [3] Kaspi, H. & Mandelbaum, A. (1998). Multi-armed bandits in discrete and continuous time, *Annals of Applied Probability* **8**(4), 1270–1290.
- [4] Ghasemzadeh, F., Archer, N. & Iyogun, P. (1999). A zero-one model for project portfolio selection and scheduling, *The Journal of the Operational Research Society* **50**(7), 745–755.
- [5] Carlsson, C., Fullér, R., Heikkilä, M. & Majlender, P. (2007). A fuzzy approach to R&D project portfolio selection, *International Journal of Approximate Reasoning* **44**(2), 93–105.
- [6] Ringuest, J.L., Graves, S.B. & Case, R.H. (2000). Conditional stochastic dominance in R&D portfolio selection, *IEEE Transactions on Engineering Management* **47**(4), 478–484.
- [7] Loch, C.H. & Kavadias, S. (2002). Dynamic portfolio selection of NPD programs using marginal returns, *Management Science* **48**, 1227–1241.
- [8] Blau, G.E., Pekny, J.F., Varma, V.A. & Bunch, P.R. (2004). Managing a portfolio of interdependent new product candidates in the pharmaceutical industry, *Journal of Product Innovation Management* **21**(4), 227–245.

- [9] Pfeiffer, T. & Schneider, G. (2007). Residual income-based compensation plans for controlling investment decisions under sequential private information, *Management Science* **53**(3), 495–507.
- [10] Dockner, E.J., Feichtinger, G. & Mehlman, A. (1993). Dynamic R&D competition with memory, *Journal of Evolutionary Economics* **3**(2), 145–152.
- [11] Mukherjee, A. (2007). *R&D, Licensing and Patent Protection*, <http://citeseer.ist.psu.edu/545430.html> (accessed Apr 2007).
- [12] Schwartz, E.S. (2004). Patents and R&D as real options, *Economic Notes* **33**(1), 23–54.

Related Articles

Enterprise Risk Management (ERM)

Simulation in Risk Management

LOUIS A. COX JR

Radioactive Chemicals/ Radioactive Compounds

Materials are made up of atoms, and key subatomic particles are called *neutrons*, *protons*, and *electrons*. The central nucleus of each atom is made up of protons and neutrons, and electrons orbit the nucleus like planets orbit the sun. Some combinations of the subatomic particles are unstable. Radioactivity is the spontaneous transformation of an unstable atom to a more stable form, often with the emission of particles and energy that together are called *radiation*. Radioactive forms of elements are often called *radioactive isotopes* or *radionuclides*.

In common notation for radionuclides, the element symbol (such as Co for cobalt or N for nitrogen) is shown with a superscript number and sometimes a subscript number. The subscript is the atomic number, the number of protons in the nucleus of an atom. For a given element, the atomic number never changes. For example, all isotopes of uranium have an atomic number of 92. The superscript number is the atomic weight, which equals the number of protons and neutrons in the nucleus. For example, $^{235}\text{U}_{92}$ has 92 protons and $235 - 92 = 143$ neutrons in its nucleus. The atomic number is sometimes omitted when radionuclides are named in text or tables with just the element name or symbol and the atomic weight, such as sodium-24 or ^{24}Na .

The traditional unit of radioactivity, the curie, is equal to 3.7×10^{10} disintegrations per second, historically based on the radioactivity in one gram of radium-226. In the International System of Units (SI), 1 Becquerel (1 Bq) equals 1 disintegration per second.

Sources of Radioactive Compounds

Over 60 radionuclides are found in nature, and hundreds more can be made by man in devices such as reactors or accelerators. Compounds that include radionuclides can occur in nature and can also be synthesized by man. Radionuclides normally encountered by humans can be categorized by their origin as terrestrial, cosmogenic, or man-made [1].

Terrestrial radionuclides are found in the earth's crust. Several dozen naturally occurring radionuclides

decay very slowly, and are left over from when the earth was created. Some of these "primordial" radionuclides occur in series that decay through chains of radionuclides until a stable isotope is reached. There are three naturally occurring decay series:

- the uranium series – headed by uranium-238 (^{238}U);
- the actinium series – headed by uranium-235 (^{235}U); and
- the thorium series – headed by the thorium-232 (^{232}Th).

Cosmic radiation refers to the radiation from space, coming primarily from stars outside our solar system. Cosmic radiation can cause cosmogenic radionuclides to be formed through interaction with the earth's atmosphere and crust. Cosmogenic radionuclides include carbon-14, hydrogen-3 (also called *tritium*), sodium-22, and beryllium-7.

Anthropogenic radioactivity comes from human activities such as nuclear weapon tests and use, reactor operation and accidents (such as Chernobyl), and use of particle accelerators. Examples include tritium from weapons testing and fission reactors, reactor fuel reprocessing facilities, and nuclear weapons manufacturing; iodine-129, iodine-131, strontium-90, and cesium-137 from weapons testing and reactors; technetium-99, a decay product of fission product molybdenum-99; and plutonium-239 from reactors and particle accelerators.

Types of Health Effects from Radiation Exposure

Health effects from exposure to radiation fall into two categories. Deterministic effects occur only when a minimum dose or threshold is exceeded. Examples include reddening of the skin (erythema) and formation of cataracts. Once the threshold has been exceeded, the *severity* of the effect increases with dose. Ionizing radiation can damage genetically important molecules in cells, such as DNA, either directly (direct hit) or indirectly through chemical means. The damage to the DNA of a cell can result in the cell losing its ability to reproduce. The cell might eventually die, generally when division is attempted. If enough cells suffer lethal damage in an organ or tissue, there will be a loss of organ function.

Units of radiation dose in current use include the gray (Gy), the SI unit of absorbed dose in tissue; 1 Gy corresponds to absorption of 1 J of energy/kg of body tissue, or 100 rad in traditional units. The sievert is the SI unit of dose equivalent, which is the absorbed dose in gray multiplied by a quality factor that reflects the relative effectiveness of the different types of radiation in producing biological effects (1 Sv = 100 rem in traditional units). For a point of reference, the current occupational dose limit for adults in the United States is 50 mSv (0.05 Sv) per year total effective dose equivalent from external exposures to the whole body and from intakes of radioactive material [2].

Without quality medical care, the dose at which 50% of people will die within 60 days (the LD_{50/60}) is about 3–5 Gy (300–500 rad). If a small area of the body is acutely exposed to high doses, death may not occur but other effects may be seen. An acute dose of approximately 6 Gy to the skin from β radiation, such as from skin contamination, will cause sunburn-like reddening of the skin (erythema) within a week [3]. The threshold for temporary sterility in males is a single absorbed dose in the testes of about 0.15 Gy. Permanent sterility will occur in men if doses exceed 3.5–6 Gy, and in women when doses exceed 2.5–6 Gy [4]. Radiation-induced cataract formation has been observed following doses of over 0.5–1.5 Gy to the lens of the eye [5]. The thresholds discussed above are for acute exposures: in other words, doses delivered in a single, brief exposure. If the dose is spread out over a longer period of time, the effect will be less, as cells can repair much damage inflicted on them, if given time to do so.

For the other category of radiation effects, stochastic effects, the *probability* rather than the severity of the effect increases with dose. With stochastic effects, the nature of the harm arising from a low dose of radiation, if it occurs, is the same as that arising from a higher dose. While there is controversy regarding whether or not there are thresholds below which radiation exposures do not increase the likelihood of stochastic effects, for radiation safety purposes it is commonly assumed that any exposure to radiation, no matter how small, increases the probability that stochastic effects will occur.

If an irradiated cell is damaged, it might survive and be able to reproduce. However, if repair is not perfect, then mutations can result in the formation of

cancer cells or in the passing of faulty information to succeeding generations of cells. These are stochastic effects, and it is possible that the harm from a low dose of radiation will be just as great as that arising from a higher dose. However, the probability of that harm occurring is lower with the lower dose.

Heritable effects of ionizing radiation exposure have not been clearly demonstrated in humans. On the basis of studies in mice, it appears that the radiation dose in humans that would double the natural incidence of a genetic or somatic abnormality is not likely to be smaller than approximately 1 Sv (100 rem) [5]. The International Commission on Radiological Protection (ICRP) risk factors for hereditary effects are 0.01 Sv⁻¹ for all future generations for the whole population and 0.06 Sv⁻¹ for the population of adult workers.

The lifetime risk for a fatal cancer as determined by the ICRP is 0.05 Sv⁻¹ for the general population and 0.04 Sv⁻¹ for adult workers [4]. These are risk factors for cancer deaths, and not for cancer incidence. Some cancers have a relatively low mortality rate, so that the probabilities of cancer induction are generally higher by a factor of about 1.5. If we assume that the excess fatal cancer incidence is proportional to the radiation dose received, we can estimate the number of cancer deaths that might occur in an exposed group of people by multiplying the collective dose equivalent (person-Sv) by the cancer risk factor. For example, if a group of 100 000 people is exposed to 0.02 Sv, the estimated number of cancer deaths over the life span of the group would be 100 000 persons $\times$ 0.02 Sv $\times$ 0.05 Sv⁻¹ = 100 deaths. This number may be compared with about 25 000 deaths from cancer that would arise “naturally”. In practice, it would be impossible to distinguish cancer deaths that may have resulted from radiation exposure from those that would have otherwise occurred in the population.

Factors Affecting Degree of Hazard from Radioactive Compounds

Health risk from radioactive compounds can result from intake of the material, contact with the material, or being within the range of the radiation that the material gives off. The degree of health risk from a radioactive compound varies with the following factors.

The Concentration and Strength of the Radionuclide in the Material

Radionuclides can be found in trace quantities in materials that are encountered, or as purified sources or major constituents of materials. The activity of a radionuclide per unit weight of element present is called its *specific activity*. Specific activity can be expressed in units such as curies per gram or terabecquerel per gram. Specific activities of common radionuclides included in Table 1 range (in their pure forms) over almost 14 orders of magnitude, from 0.00000000405 for thorium-232 to 322 000 TBq g⁻¹ for sodium-24 (0.000000109 to 8 710 000 Ci g⁻¹ in traditional units).

The amount of material necessary to assemble 1 Ci of radioactivity can vary from a quantity too small to be seen (less than a microgram for sodium-24, for example) to tons of the material (more than 9 MT for thorium-232). The specific γ -ray dose constants presented in Table 1 illustrate how some radionuclides are sources of more intense radiation than others. These constants, which represent the approximate dose rate at a distance of 1 m from a γ radiation source for each megabecquerel of activity that it contains, vary by more than a factor of 100 000 from 0.000000001424 mSv h⁻¹ MBq⁻¹ for polonium-210 to 0.000524 mSv h⁻¹ MBq⁻¹ (1.8 rem h⁻¹ Ci⁻¹) for sodium-24.

The Type of Radiation That is Emitted

There are several more common forms of radiation that have enough energy to damage cells or other atoms. These “ionizing” radiations include α particles, β particles, and γ rays.

α Particles are relatively large subatomic fragments consisting of two protons and two neutrons that are ejected from the nucleus of some unstable atoms. They are the least penetrating form of radiation, as depicted in Figure 1, as they can be stopped by an inch or two of air or thin layers of light materials, such as a sheet of paper or the outer layer of one’s skin.

β Particles are subatomic particles (essentially electrons) that are ejected from the nucleus of some radioactive atoms. Over 7000 times smaller than α particles, β particles are more penetrating than α particles, but they are stopped by several feet of air. They are less ionizing and thus do less cellular

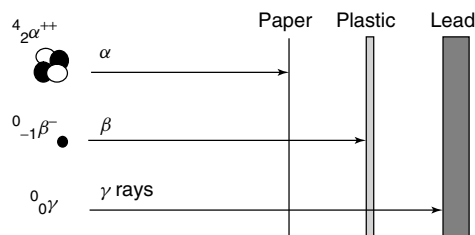


Figure 1 Penetrating abilities of the main types of radiation

damage, but can still present a substantial risk. Most β rays would be absorbed by one’s clothing, a few sheets of cardboard, or an acrylic sheet.

γ Rays are not particles, but rather electromagnetic radiations like the visible light from the sun and ultraviolet light, but with higher energies. γ Rays are essentially the same as X rays, but γ rays originate in the nucleus and X rays originate in the electron fields surrounding the nucleus. γ Rays are highly penetrating – it takes about 1 in. of solid steel to absorb 50% of them.

The Route by Which Exposure Occurs

Radioactive materials can appear as solids, powders, dusts, liquids, gases, vapors, or solutions. Radioactive compounds can enter the body by inhalation, ingestion, or absorption through the unbroken skin or through wounds. Unlike chemical toxins, radioactive materials can present health hazards even if a person does not touch the material or take any into the body. If a person is within the range of radiations that are emitted from the material, even if the material itself is well contained, exposure can occur as radiation is absorbed by body tissues in its path.

A source of α radiation kept outside your body is not a significant hazard. Radioactive material that emits α particles can pose a serious health threat when it is taken into the body. While α particles will not travel far into the body, they are “highly ionizing” and can do a lot of cellular damage along their short paths.

Direct exposure to β particles is a hazard, because emissions from strong sources can redden or even burn the skin. However, emissions from inhaled or ingested β -particle emitters are the greatest concern. β Particles released directly to living tissue can cause damage at the molecular level, which can disrupt cell

function. As they are much smaller and have less charge than α particles, β particles generally travel further into tissues. As a result, the cellular damage is more dispersed.

Both direct (external) and internal exposures to γ rays are of concern. γ Rays can travel much farther than α or β particles and have enough energy to pass entirely through the body, potentially exposing all organs. A large portion of γ radiation largely passes through the body without interacting with tissue. Most of the radiation dose received by the internal organs of the body from an external source would be from γ rays.

How the Material Behaves in the Human Body

The hazard from a radionuclide taken into the body depends on numerous factors, including its physical form, chemical form, solubility, where it is retained in the body, and how quickly the levels in body tissues decrease through radioactive decay and biological elimination. The particle sizes of a material determine whether it will make it into the lungs or whether it will be collected in the nasal passages or the upper respiratory tract. If a radioactive material is soluble in body fluids, it is called *transportable* and will enter the blood stream from the lungs or the digestive tract. It will normally deposit in various body organs at different levels depending on its physical and chemical properties, and will be eliminated *via* the urine and/or feces. If a radioactive material is not soluble in body fluids, it will typically be retained in the lungs for some period of time, with a fraction transported from the lungs by ciliary action leading to eventual elimination in the feces.

The time required for a radionuclide to lose 50% of its activity by decay is called its *radioactive (or physical) half-life* ($T_{1/2}$). While the amount present in the body decreases owing to radioactive decay, it also decreases owing to biological elimination. The biological half-life (T_b) is the time taken for the amount of a particular element in the body to decrease to half its initial value owing to elimination by biological processes alone. An effective half-life (T_{eff}), the time taken for the amount of a specified radionuclide in the body to decrease to half its initial value as a result of both radioactive decay and biological elimination, can be calculated as follows:

$$T_{\text{eff}} = \frac{(T_{1/2} \times T_b)}{(T_{1/2} + T_b)} \quad (1)$$

As can be seen in Table 1, a short biological half-life can result in a short effective half-life even when the physical half-life is very long in comparison (see ^{60}Co , ^{235}U , and ^{238}U).

Much research has been conducted on humans and animals to describe how radioactive materials are retained and transported in the human body and to describe the kinetics of retention and excretion. For many radionuclides, detailed mathematical models have been developed to support prediction and/or reconstruction of concentrations of radionuclides in the different body tissues over time so that doses to specific organs can be estimated and health risks evaluated. For example, models have been developed that represent in the following manner [4, 10, 11]:

- The respiratory system in four regions and 10 compartments, with compartmental fractions and removal halftime specified for three classes of retained material [Class D = minimum retention (halftime less than 10 days), Class W = moderate retention (10–100 days), and Class Y = avoid retention (greater than 100 days)] (see Figure 2).
- The gastrointestinal (GI) tract in four sections and five compartments, with mean residence times specified for each section, halftime specified for transport from one section to the next, and a half-time specified for transport of the radionuclide from the small intestine to the body fluids compartment. Data are provided for up to three different classes of material based on the fraction of ingested material that is transferred to the blood.
- The skeleton in six tissue types, with factors describing irradiation of cells near bone surfaces and red bone marrow from each tissue type, for radionuclides that bind to bone surfaces or are distributed uniformly through the bone volume.
- Clearance of radionuclides into organs and tissues after translocation from the GI tract or the respiratory tract into body fluids, with excretion also represented with specified halftime.

On the basis of these models, ingestion and inhalation dose coefficients are published, which allow one to estimate doses that will be received by 21 organs and a weighted total effective dose equivalent to age 70 following intakes at various ages [12].

Table 1 Data relevant to radiological hazard from selected radionuclides^(a)

Radionuclide	Key radiations ^(b)	Specific activity of isotope ^(c) (TBq g ⁻¹)	Specific γ -ray dose constant ^(d) (mSv h ⁻¹ at 1 m MBq ⁻¹)	Half-life		Important organ ^(e)
				Physical	Effective	
Americium-241	α, γ	0.127	0.000085	432.2 years	139 years	Bone
Barium-140	β, γ, D	2710	0.000044	12.8 days	11 days	Bone
Californium-252	γ, α , neutron, D	19.9	0.000011	2.6 years	2.2 years	Bone
Carbon-14	β	0.165	—	5730 years	12 days	Total body
Cesium-137	β, γ, D	3.22	0.000103	30.17 years	70 days	Total body
Chromium-51	γ	3420	0.000063	27.7 days	27 days	Total body
Cobalt-60	β, γ	41.8	0.000370	5.3 years	10 days	Total body
Europium-152	β, γ, D	7.03	0.000201	13.6 years	3 years	kidney
Hydrogen-3 (tritium)	β	357	—	12 years	12 days	Total body
Iodine-131	β, γ, D	4590	0.000076	8.0 days	8 days	Thyroid
Iron-59	β, γ	1840	0.000179	44.6 days	42 days	Spleen
Lanthanum-140	β, γ	20600	0.000305	40.22 hours	40.22 hours	GI tract
Lead-210	β, γ, D	2.83	0.000068	22.3 years	1.3 years	Kidney
Manganese-54	β, γ	286	0.000138	312.7 days	6.6 days	GI tract
Molybdenum-99	β, γ, D	17700	0.000031	2.8 days	1.5 days	Kidney
Neptunium-237	α, γ, D	0.0000255	0.000125	2.1 million years	200 years	Bone
Phosphorus-32	β	10600	—	14 days	14 days	Bone
Plutonium-238	α, γ	0.634	0.000021	87.8 years	63 years	Bone
Plutonium-239	α, γ	0.0023	0.0000081	24 100 years	197 years	Bone
Polonium-210	α	166	0.000000014	138 days	46 days	Spleen
Radium-226	α, γ, D	0.0366	0.0000033	1600 years	44 years	Bone
Ruthenium-106	β, D	124	0.000037	368 days	2.5 days	Kidney
Sodium-24	β, γ	322000	0.000524	15 hours	14 hours	Total body
Sr-90	β, D	5.05	—	28 years	15 years	Bone
Technetium-99m	γ	194000	0.000033	6 hours	5 hours	Total body
Thorium-232	α, γ, D	0.00000000405	0.000018	14 billion years	200 years	Bone
Uranium-235	α, γ, D	0.00000008	0.000092	704 million years	15 days	Kidney
Uranium-238	α, γ, D	0.0000000124	0.000018	4.47 billion years	15 days	Kidney
Yttrium-90	β	20100	—	64 hours	64 hours	Bone
Zinc-65	β, γ	305	0.000089	244.4 days	194 days	Total body
Zirconium-95	β, γ, D	795	0.000126	64.0 days	56 days	Total body

^(a) References [2, 3, 6–9]^(b) ‘D’ indicates the possible presence of decay product ‘daughters’ with half-life less than 25 years^(c) A terabecquerel (TBq) is a million million (10¹²) transformations per second, or 27.03 Ci in traditional units^(d) Value for cesium-137 includes a radiologically significant decay product, barium-137m^(e) Different chemical forms and modes of exposure will determine what organs are most affected. This table is intended to give a limited presentation of one principal organ at risk until details can be obtained to support assessment of which organ receives the highest dose from an intake of the radionuclide, or has the most significant biological effect

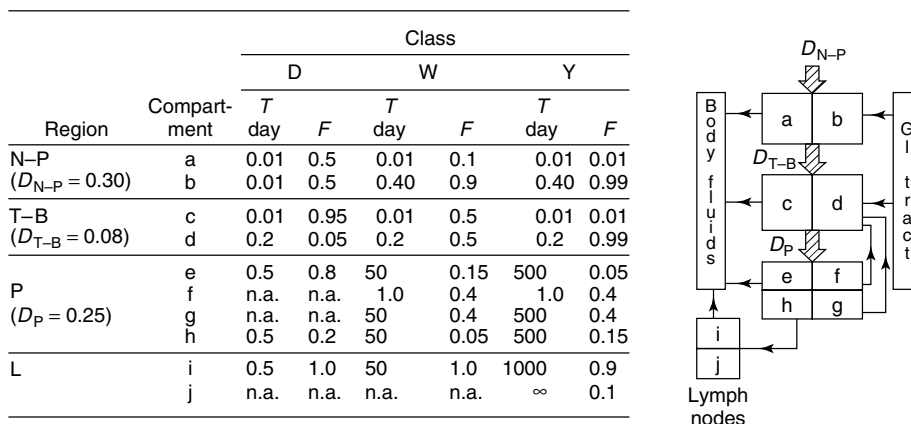


Figure 2 Mathematical model used to describe clearance of radionuclides from the respiratory system. The values for the removal half-time, T_{a-j} , and compartmental fractions, F_{a-j} , are given in the tabular portion of the figure for each of the three classes of retained materials. The values given for D_{N-P} , D_{T-B} , and D_P (left column) are the regional depositions for an aerosol with an activity median aerodynamic diameter (AMAD) of $1\ \mu\text{m}$. The schematic drawing identifies the various clearance pathways from compartments a-j in the four respiratory regions, N-P, T-B, P, and L. n.a.: not applicable [Reproduced from [10]. © Pergamon Press, 1979.]

The Chemical Toxicity of the Material

Radionuclides and compounds that contain chemical toxicity do not behave noticeably different chemically than the stable forms of the same elements or compounds. Compounds that exhibit chemical toxicity in their stable forms will retain that toxicity when they contain radioactive isotopes. In some cases, chemical toxicity becomes secondary in importance to radiological hazard, but that is not always the case. For example, uranium is used in various military, aviation, nuclear power, and nuclear weapon applications. As a heavy metal, uranium in soluble forms damages the kidneys. Natural uranium normally contains 99.27% uranium-238, 0.72% uranium-235, and 0.0057% uranium-234 [3]. For some uses, uranium is “enriched” so that its fissionable ^{235}U isotope is more abundant. For individuals exposed to airborne uranium that is relatively soluble in body fluids, the chemical toxicity of uranium is believed to be predominant over the radiological toxicity for ^{235}U enrichments below 7 to 20% [13].

References

- [1] Eisenbud, M. & Gesell, T. (1997). *Environmental Radioactivity*, 4th Edition, Academic Press, San Diego.
- [2] U.S. Nuclear Regulatory Commission (1996). *Instruction Concerning Risks from Occupational Radiation Exposure*, Regulatory Guide 8.29, Washington, DC.
- [3] Shleien, B., Slaback Jr, L.A. & Birky, B.K. (1998). *Handbook of Health Physics and Radiological Health*, 3rd Edition, Williams & Wilkins, Baltimore.
- [4] International Commission on Radiological Protection (1990). *1990 Recommendations of the International Commission on Radiological Protection*, ICRP Publication No. 60, Annals of the ICRP, Pergamon Press, New York.
- [5] National Research Council, Committee on the Biological Effects of Ionizing Radiation (1990). *Health Effects of Exposure to Low Levels of Ionizing Radiation, BEIR V*, National Academy Press, Washington, DC.
- [6] U.S. Department of Health, Education, and Welfare, Bureau of Radiological Health (1970). *Radiological Health Handbook, Revised Edition*, Public Health Service, Rockville.
- [7] International Commission on Radiological Protection (1960). *Report of Committee II on Permissible Dose for Internal Radiation*, ICRP Publication No. 2, Pergamon Press, New York.
- [8] International Commission on Radiological Protection (1972). *The Metabolism of Compounds of Plutonium and Other Actinides*, ICRP Publication No. 19, Annals of the ICRP, Pergamon Press, New York.
- [9] National Council on Radiation Protection and Measurements (1979). *Management of Persons Accidentally Contaminated with Radionuclides*, NCRP Report No. 65, Bethesda.
- [10] International Commission on Radiological Protection (1979). *Limits for Intakes of Radionuclides by Workers*.

- ICRP Publication No. 30, Annals of the ICRP, Pergamon Press, New York.
- [11] International Commission on Radiological Protection (1986). *The Metabolism of Plutonium and Related Elements*, ICRP Publication No. 49, Annals of the ICRP, Pergamon Press, New York.
- [12] International Commission on Radiological Protection (1989). *Age-dependant Doses to Members of the Public from Intake of Radionuclides: Part 1*, ICRP Publication No. 56, Annals of the ICRP, Pergamon Press, New York.
- [13] Brodsky, A. (1996). *Review of Radiation Risks & Uranium Toxicity with Applications to Decisions Associated with Decommissioning and Clean-up Criteria*, RSA Publications, Hebron.

Related Articles

Chernobyl Nuclear Disaster

Radon Risk

What are Hazardous Materials?

THOMAS E. WIDNER

Radon

Radon, or more properly, radon-222, is a colorless, odorless, inert gas, which arises from the radioactive decay of radium-226. Radon is found, at varying levels, in all indoor and outdoor environments. It can diffuse through cracks in foundations, can be transported in water, and can move into basements through sewer lines or drain pipes [1]. It is of interest to the risk analysis community because inhaling radon, or more accurately its decay products or “progeny”, can expose lung tissue to significant doses of ionizing radiation in the form of α particles, which may in turn cause lung cancer [2, 3]. The United States Environmental Protection Agency (USEPA) [4] has estimated that, in the United States, exposure to radon progeny causes between 15 400 and 21 800 excess lung-cancer cases each year (depending on the risk model selected), but acknowledges that the range of uncertainty is between 3000 and 33 000 cases per year. Even the lowest of these figures make radon the second leading cause of lung cancer after cigarette smoking (*see Environmental Tobacco Smoke*).

The 10-fold range given by EPA suggests substantial uncertainty in radon risk assessments. One major uncertainty is the radiation exposure term (often expressed in working level months (WLM)) that will result from exposure to a given level of radon progeny (measured in pico curies or pCi l^{-1} of air). This issue is of such complexity that it was considered in a National Research Council report [5]. Nonetheless, it remains a major uncertainty [6].

Another uncertainty is form of the dose–response function (*see Dose–Response Analysis*) that translates radon dose into lung-cancer risk. Recent risk assessments [3, 4] have used a relative risk model in which the effect of radon exposure is to multiply baseline risk, and it is generally accepted that radon exposure and cigarette smoking interact multiplicatively to increase lung-cancer risk. Thus the baseline lung-cancer risk in an exposed population is an important uncertainty in any radon risk assessment.

Uncertainty also arises because many of the excess lung cancers predicted necessarily result from very large populations exposed to very low levels of radon. Since much of our data for estimation of radon progeny risks comes from uranium miner cohorts, which have high radon exposure levels and many confounding factors like exposure to ore dust and

diesel exhaust, there is a question of how these data apply to low-level environmental exposures. However, two large recent meta-analyses of studies of residential radon exposure and lung-cancer risk [7, 8] suggest that risk estimates per unit dose are fairly comparable between miner and residential studies.

Lung-cancer risks from living in radon contaminated homes can be substantial. The USEPA [1] estimates that people living in homes at the “action level” of 4 pCi l^{-1} of air have a lifetime risk of about 6 in 100 (6%) if they are smokers and about 7 in 1000 (0.7%) if they are nonsmokers. Lifetime risk is approximately proportional to exposure level. That is, 8 pCi l^{-1} has about twice the risk of 4 pCi l^{-1} and 2 pCi l^{-1} has about half the risk of 4 pCi l^{-1} . Reducing radon risk is a matter of reducing radon exposure levels. EPA recommends that radon reduction be undertaken when levels exceed 4 pCi l^{-1} and should be considered when levels exceed 2 pCi l^{-1} . While risks at the lower level are still substantial compared with allowable risks from things like food additives (generally in the range of 10^{-5} – 10^{-6}) reducing levels much below 2 pCi l^{-1} is difficult because of the ubiquitous nature of radon.

Since a major route of infiltration is through foundation cracks, sealing cracks may substantially reduce levels. More aggressive reduction strategies focus on systems that use pipes and fans to suck radon out of the soil spaces beneath the foundation floor (often referred to as *subslab* ventilation), and in some areas houses are constructed to be “radon proof” in the sense that subslab ventilation is a built-in feature [1]. Exposure can also result from radon in drinking water but is much less common and is of concern primarily for homes using groundwater as their water supply source. If radon in drinking water is of concern, aeration can substantially reduce levels (*see Environmental Risk Assessment of Water Pollution*).

References

- [1] EPA (2005). *A Citizens Guide To Radon: The Guide to Protecting Yourself and Your Family from Radon*, U.S. Environmental Protection Agency, Document Number 402-K02-006.
- [2] NCRP (1984). *Exposures from the Uranium Series with Emphasis on Radon and its Daughters*, Report Number 77, National Council on Radiation Protection and Measurements, Bethesda, p. 131 + vi.

- [3] NRC (1999). *Health Effects of Exposure to Radon: BEIR VI*, National Academy Press, Washington, DC, p. 500 + ix.
- [4] EPA (2003). *EPA Assessment of Risks from Radon in Homes*, U.S. Environmental Protection Agency, Document Number 402-R-03-003.
- [5] NRC (1991). *Comparative Dosimetry of Radon in Mines and Homes*, National Academy Press, Washington, DC, p. 244 + xi.
- [6] Steck, D.J. & Field, R.W. (2006). Dosimetric challenges for residential radon epidemiology, *Journal of Toxicology and Environmental Health Part A* **69**, 655–664.
- [7] Darby, S., Hill, D., Deo, H., Auvinen, A., Barros-Dios, J.M., Baysson, H., Bochicchio, F., Falk, R., Farchi, S., Figueiras, A., Hakama, M., Heid, I., Hunter, N., Kreienbrock, L., Kreuzer, M., Lagarde, F., Makelainen, I., Muirhead, C., Oberaigner, W., Pershagen, G., Ruosteenoja, E., Rosario, A.S., Tirmarche, M., Tomasek, L., Whitley, E., Wichmann, H.E. & Doll, R. (2006). Residential radon and lung cancer—detailed results of a collaborative analysis of individual data on 7148 persons with lung cancer and 14,208 persons without lung cancer from 13 epidemiologic studies in Europe, *Scandinavian Journal of Work Environment and Health* **32**(1), 1–83.
- [8] Krewski, D., Lubin, J.H., Zielinski, J.M., Alavanja, M., Catalan, V.S., Field, R.W., Klotz, J.B., Letourneau, E.G., Lynch, C.F., Lyon, J.L., Sandler, D.P., Schoenberg, J.B., Steck, D.J., Stolwijk, J.A., Weinberg, C. & Wilcox, H.B. (2006). A combined analysis of North American case–control studies of residential radon and lung cancer, *Journal of Toxicology and Environmental Health Part A* **69**, 533–597.

Related Articles

Radon Risk

MICHAEL E. GINEVAN

Radon Risk

Background

Radon is an inert gas, produced naturally from radium in the decay series of uranium (Figure 1). Radon decays with a half-life of 3.82 days into a series of solid, short-lived radioisotopes that collectively are referred to as radon daughters, progeny, or decay products.

Radon was the first environmental respiratory carcinogen identified, being linked to increased risk of lung cancer in underground miners in the early twentieth century, even before smoking was found to cause lung cancer. Many epidemiologic studies of underground miners of uranium and other ores have established exposure to radon daughters as a cause of lung cancer [1–3]. In the miners exposed to radon in past centuries, very high lung cancer risks were observed; these were lower for more recent workers but the recent epidemiological studies still show clear evidence of existing cancer risk [3]. Cigarette smoking and radon decay products synergistically influence lung cancer risk in a manner that is supra-additive but submultiplicative,^a placing smokers who are exposed to high levels of radon at extremely high risk for lung cancer [1, 3]. Beginning in the 1970s, radon was found to be ubiquitous in the air of homes, at high levels in some, and concern quickly arose as to the associated lung cancer risk to the general population.

Increasingly elegant experimental studies have documented the occurrence of permanent damage to a cell from just one hit by an α particle [3]. This experimental finding suggests that assuming a linear nonthreshold relationship between exposure and risk at the levels found not only in mines but also in indoors is biologically appropriate. In this same type of experimental system, a bystander mutagenic effect has been demonstrated; a hit to a cell affects cells adjacent to the cell damaged by a single α particle [4]. This effect may amplify the risks of radon exposure beyond those anticipated based on the construct that passage of an α particle through a cell damages only that cell.

For historic reasons, the concentration of radon progeny in underground mines is generally expressed in working levels (WLs): 1 WL is any combination of radon progeny in 1 l of air that ultimately releases

1.3×10^5 MeV of alpha energy during decay. In the United States, concentrations of radon are generally given as picocuries per liter (pCi l^{-1}), a unit for the rate of decay. Exposure to 1 WL for 170 h equals 1 working level month (WLM) of exposure, a unit developed to describe exposure sustained by miners during the average number of hours spent underground. In SI units, concentration is described as Becquerels per cubic meter with 1 Bq m^{-3} equivalent to 1 pCi l^{-1} ; exposure is in units of joules per hour per cubic meter with 1 WLM equivalent to $3.5 \times 10^{-3} \text{ J h}^{-1} \text{ m}^{-3}$.

Radon and Lung Cancer Risk

As the biologic basis of respiratory carcinogenesis was understood and the lung dosimetry of radon and its short-lived progeny was described, it was recognized that α -particle emissions from inhaled radon progeny, not from radon itself, caused lung cancer in radon-exposed miners [3]. Two of those decay products, polonium-218 and polonium-214, emit α particles, which are high-energy and high-mass particles consisting of two protons and two neutrons that are very damaging to tissue. The energy of these particles is invariant with concentration of radon progeny so that the potential for passage of alpha particles to damage target cells is the same at high and low concentrations. When the alpha emissions take place within the lung as inhaled radon progeny decay, the DNA of cells lining the airways may be damaged and lung cancer may ultimately result. Animal studies demonstrated that radon alone through its progeny could induce cancer in the respiratory tract [1].

The relationship between radon progeny and lung cancer risk has been investigated in follow-up (cohort) studies (*see Cohort Studies*) of mining populations, including not only uranium miners but also miners of other types of ores who were exposed to radon either from the ore itself or from the radon dissolved in mine water. Those cohort studies with exposure estimates for the miners have shown that the risk for lung cancer increases proportionally with the cumulative exposure to radon progeny. These 11 studies were included in a pooled analysis reported by Lubin *et al.* [2, 3, 5] and extended by the Biological Effects of Ionizing Radiation (BEIR VI) Committee VI [3]. Although the individual cohort

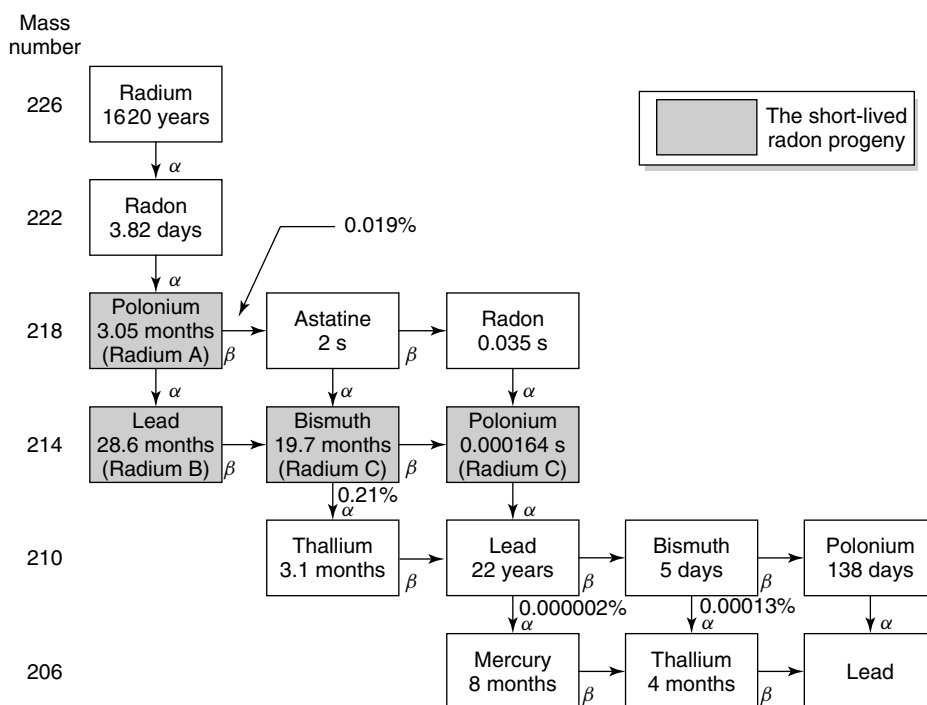


Figure 1 Uranium decay chain

studies differ somewhat in the nature of the mining populations and in their methods, and not all include detailed smoking histories from all miners, the crude estimates of excess relative risk (ERR) per WLM of radon progeny exposure were reasonably close. As anticipated from the presumed mechanism of carcinogenesis, all of the studies documented increasing risk with increasing exposure without any indication of a threshold below which risk was not manifest.

The pooled analysis showed that the risk varied with time since exposure and the rate at which cumulative exposure was achieved, and was also modified by attained age and smoking. There is evidence for an inverse dose-rate effect or protraction enhancement, so that miners who were exposed at low concentrations over long periods of time have a higher risk than miners receiving the same total exposure over shorter periods. Risk also dropped as time since exposure lengthened and independently as attained age increased. The pooled analyses indicated a synergistic interaction between smoking and radon progeny. The age at which a miner is first exposed to radon does not appear to have a consistent effect on

risk overall, but the epidemiological data are limited at younger ages.

An elevated risk for lung cancer has also been shown in miners who have never smoked cigarettes. A follow-up study of 516 white nonsmoking miners from the Colorado Plateau cohort showed 14 lung cancer deaths with only 1.1 expected, based on the never-smokers in the study of US veterans [6]. Native American miners in the Colorado Plateau cohort, mostly Navajos, were almost all nonsmokers or very light smokers; a follow-up study of 757 members of this group identified 34 deaths from lung cancer when only 10.2 were expected [standardized mortality ratio (SMR) 3.3, 95% confidence interval (CI) 2.3–4.6], confirming an earlier case–control study of lung cancer in Navajo men [7]. A pooled analysis of the 2798 never-smoking miners in the 11 miner groups found that the estimated ERR/WLM for nonsmokers (0.0103, 95% CI 0.002–0.057) was almost three times that of smokers (0.0034, 95% CI 0.001–0.015), consistent with the submultiplicative interaction between smoking and radon in the full data set [2].

Radon and Indoor Environments

As information on air quality in indoor environments accumulated during the 1970s and 1980s, it became clear that radon and its decay products that are invariably present in indoor environments and in some dwellings may reach unacceptably high concentrations, equivalent to those in mines. In fact, radon is the greatest source of radiation exposure in the United States from natural background radiation [8]. Radon primarily enters homes and other buildings as soil gas is drawn in by the pressure gradient between the structure and the ground beneath. An extensive body of literature now addresses the risks of indoor radon [3, 9, 10]. The epidemiologic studies of miners of uranium and other ores have been the principal basis for estimating the risks of indoor radon, and the epidemiologic data have been extensively analyzed to develop risk models. More recent findings from case-control studies

of indoor radon and lung cancer in the general population have been reported and their findings pooled to enhance the precision of risk estimates. These studies compare radon concentrations at homes where persons with lung cancer lived with concentrations in homes of similar people who do not have lung cancer.

In response to the need for information on risks from indoor exposure, epidemiological studies directed at lung cancer in the general population were initiated in the 1970s and 1980s. The first wave of studies was largely ecological in design and provided mixed findings because of inherent flaws of this approach [8]. By the mid- to late-1980s, many case-control studies of more sophisticated design with larger sample sizes were undertaken. The findings, although having the inherent limitation of exposure measurement error, generally support the carcinogenicity of radon exposure in homes and are consistent quantitatively with expectations of risk

Table 1 Estimated number of lung cancer deaths in the United States in 1995 attributable to indoor residential radon progeny exposure^{(a)(b)}

Smoking status	Number of lung cancer deaths	Number of lung cancer deaths attributable to radon progeny exposure			
		Exposure-age-concentration model		Exposure-age-duration model	
Males^(c)					
Total	95 400	13 000 ^(d)	12 500 ^(e)	9200 ^(d)	8800 ^(e)
Ever-smokers	90 600	12 300	11 300	8700	7900
Never-smokers	4800	700	1200	500	900
Females^(c)					
Total	62 000	9300	9300	6600	6600
Ever-smokers	55 800	8300	7600	5900	5400
Never-smokers	6200	1000	1700	700	1200
Males and females					
Total	157 400	22 300	21 800	15 800	15 400
Ever-smokers	146 400	20 600	18 900	14 600	13 300
Never-smokers	11 000	1700	2900	1200	2100

^(a) Reproduced from [11]. US Government Printing Office, 1992

^(b) National Research Council (NRC), Committee on Health Risks of Exposure to Radon, Board on Radiation Effects Research, and Commission on Life Sciences. Health Effects of Exposure to Radon (BEIR VI). 1999. Washington, D.C., National Academy Press

^(c) Assuming that 95% of all lung cancers among males occur among ever-smokers, and that 90% of all lung cancers among females occur among ever-smokers. Percentages of ever-smokers in the population were 58 and 42% for males and females, respectively

^(d) Estimates based on applying same risk model to ever-smokers and never-smokers, implying a joint multiplicative relationship for radon progeny exposure and smoking

^(e) Estimates based on applying a smoking adjustment to the risk models, multiplying the base-line ERR/WLM by 0.9 for ever-smokers and by 2 for never-smokers, the committee's preferred approach

based on downward extrapolation of models based on the miner data.

In 2005, the results of pooled analyses of data from North America [9] and from Europe [10] were reported. The North American analysis included data from seven studies with 3662 cases and 4966 controls. The risk of lung cancer was estimated to increase by 11% (95% CI 0–28%) per 100 Bqm⁻³ increment in the concentration at which exposure occurs in a home. The estimate from the pooling of 13 European studies involving 7148 cases and 14 208 controls was similar; an increment of 8% (95% CI 3–16%) per 100 Bqm⁻³ increase in home concentration was estimated.

The burden of lung cancer associated with indoor radon has been estimated by using the distribution of indoor radon concentrations from surveys and the exposure–response relationship estimated from the miner studies. The assumptions made by the Environmental Protection Agency and the BEIR IV and VI committees of the National Research Council lead to estimates that approximately 15 000–20 000 lung cancer deaths per year in the United States are caused by radon (Table 1) [11]. A substantial proportion of these premature deaths could be avoided through the use of testing programs that identify homes with unacceptably high concentrations.

End Notes

^a. That is, the combined effect exceeds that expected from adding the individual effects of radon decay products and of smoking but is less than the expected value based on multiplying the effects of radon decay products and smoking.

References

- [1] National Research Council, Committee on the Biological Effects of Ionizing Radiation (1988). *Health Risks of Radon and Other Internally Deposited Alpha-Emitters: BEIR IV*, National Academy Press, Washington, DC.
- [2] Lubin, J.H., Boice Jr, J.D., Edling, C., Hornung, R.W., Howe, G.R., Kunz, E., Kusiak, R.A., Morrison, H.I., Radford, E.P., Samet, J.M., Tirmarche, M., Woodward, A., Yao, S.X. & Pierce, D.A. (1995). Lung cancer in radon-exposed miners and estimation of risk from indoor exposure, *Journal of the National Cancer Institute* **87**(11), 817–827.
- [3] National Research Council, Committee on Health Risks of Exposure to Radon, Board on Radiation Effects Research & Commission on Life Sciences (1999). *Health Effects of Exposure to Radon (BEIR VI)*, National Academy Press, Washington, DC.
- [4] Hall, E.J. & Hei, T.K. (2003). Genomic instability and bystander effects induced by high-LET radiation, *Oncogene* **22**(45), 7034–7042.
- [5] Lubin, J.H., Boice Jr, J.D., Edling, C., Hornung, R.W., Howe, G., Kunz, E., Kusiak, R.A., Morrison, H.I., Radford, E.P., Samet, J.M., Tirmarche, M., Woodward, A. & Yao, S.X. (1995). Radon-exposed underground miners and inverse dose-rate (protraction enhancement) effects, *Health Physics* **69**(4), 494–500.
- [6] Roscoe, R.J., Steenland, K., Halperin, W.E., Beaumont, J.J. & Waxweiler, R.J. (1989). Lung cancer mortality among nonsmoking uranium miners exposed to radon daughters, *Journal of the American Medical Association* **262**(5), 629–633.
- [7] Samet, J.M., Kutvirt, D.M., Waxweiler, R.J. & Key, C.R. (1984). Uranium mining and lung cancer in Navajo men, *New England Journal of Medicine* **310**(23), 1481–1484.
- [8] Stidley, C.A. & Samet, J.M. (1993). A review of ecologic studies of lung cancer and indoor radon, *Health Physics* **93**(3), 234–251.
- [9] Krewski, D., Lubin, J.H., Zielinski, J.M., Alavanja, M., Catalan, V.S., Field, R.W., Klotz, J.B., Letourneau, E.G., Lynch, C.F., Lyon, J.I., Sandler, D.P., Schoenberg, J.B., Steck, D.J., Weinberg, C. & Wilcox, H.B. (2005). Residential radon and risk of lung cancer: a combined analysis of 7 North American case–control studies, *Epidemiology* **16**(2), 137–145.
- [10] Darby, S., Hill, D., Auvinen, A., Barros-Dios, J.M., Baysson, H., Bochicchio, F., Deo, H., Falk, R., Forastiere, F., Hakama, M., Heid, I., Kreienbrock, L., Kruezer, M., Lagarde, F., Mäkeläinen, I., Muirhead, C., Oberaigner, W., Pershagen, G., Ruano-Ravina, A., Ruosteenoja, E., Schaffrath Rosario, F., Tirmarche, M., Tomásek, L., Whitley, E., Wichmann, H.E. & Doll, R. (2005). Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case–control studies, *British Medical Journal* **330**(7485), 223.
- [11] US Environmental Protection Agency (1992). *Technical Support Document for the 1992 Citizen's Guide to Radon*, U.S. Government Printing Office, Washington, DC.

Related Articles

Environmental Health Risk

Radon

Threshold Models

JONATHAN M. SAMET

Randomized Controlled Trials

In the medical sciences, the ideal experiment is the randomized controlled clinical trial (RCT) in which the subject or patient is randomized to treatment with an experimental agent or control (either placebo or other active treatment), in a double-blind manner. The RCT is the “gold standard” in medical research to establish the efficacy of a new treatment or intervention. The benefits of such an experiment are that any observed differences between the two groups (or more groups) with respect to the outcome can be attributed to the treatment and are without bias. That is, confounding is controlled and selection bias is removed since neither the subject nor the investigator chooses the intervention [1].

In this article, we review the issues in the design, conduct, analysis, and interpretation of a randomized clinical trial.

Design Considerations

The concepts of the RCT are applicable to any clinical trial from phase I to phase III and beyond (see phase I–III trials). Most of our remarks here will apply to phase III comparative efficacy trials, but can be generalized also to earlier phase clinical trials.

Early Phase Trials

Although early phase trials are generally not randomized, the randomization framework can be useful to permit the direct comparison of toxicities and adverse events for the new treatment or intervention with toxicities and adverse events in a control group. In such cases, randomization would minimize the risk of attributing adverse events that might be due to the underlying condition from those that might be attributable to the intervention. The use of the RCT paradigm in such trials would reduce the risk of incorrectly attributing safety issues to a new drug early in the development process. Similarly, phase II studies are often uncontrolled, but in many instances randomized designs can be implemented to eliminate the selection bias associated with treatment assignment. In such cases, formal comparisons of the treatment

and control arms will be of limited utility due to low power to detect differences. However, some phase II trials can be conducted in diseases such as hypertension using RCT designs for preliminary dose finding with results to be confirmed in definitive phase III trials.

Example of Randomized Phase III Trial

The classic example of a large randomized comparative efficacy trial (phase III) is the Salk vaccine trial that was conducted in 1954. Over one million children participated in the trial. This trial was conducted to assess the effectiveness of the Salk vaccine to provide protection against paralysis or death from poliomyelitis as described in detail by Meier [2]. The ideal design is as a randomized controlled trial. Since there was considerable reluctance to allow children to receive placebo injections, an observed control study was designed with children in grade 2 to receive vaccinations and children in grades 1 and 3 to be observed for the occurrence of polio. After much debate, some states agreed to participate in a randomized component. In the RCT, 750 000 children were randomly assigned to injections either with the Salk vaccine or with a placebo salt solution. The trial was double blind, that is, neither the children nor the diagnosing physicians were aware of who had received the Salk vaccine or the placebo. Patients and/or their physicians participating in the study are often blinded to the treatment assignment to reduce the likelihood of treatment-related bias due to the knowledge of treatment assignment (*see* **Blinding**). The Salk vaccine trial was conducted in a relatively short period of time, and the endpoints were observed within this relatively short time period. The results obtained for the RCT component of the trial were unequivocal and were supported by the observed control component which, although supportive, would not have provided compelling evidence that the vaccine was effective.

Key Design Issues

At the design stage of a randomized controlled trial, the investigators need to carefully consider several key issues that include the choice of the treatment and control regimens including the methods of treatment delivery; the types of patients and severity of disease to be studied; the level of blinding; the use

of a parallel group design or some alternative (e.g., cross-over trial); the need for stratification; choice of a single or multicenter trial; the length of the treatment and/or observation period; the unit of analysis (e.g., patient or eye within patient); the outcomes (short term, long term, surrogates); the measurements of these outcomes including measurement error; and the size of a meaningful treatment effect (statistical significance *versus* clinical significance). Special issues that relate to the use of surrogate endpoints that should correlate with the long-term outcome of interest should also be considered [3, 4]. Designs of definitive randomized controlled trials of new drugs or devices also need to consider the regulatory requirements for the approval of such drugs or devices to minimize risks of failing to meet the overall objective of approval of a new drug or device for use [5, 6].

Blinding and Randomization. In the planning stages of a clinical trial, it is prudent to consider the use of a blinded RCT as the first type of design. Careful consideration should be given to the choice of the treatment regimen or intervention to be studied as well as to the selection of the control treatment or intervention. Sometimes the nature of a study precludes blinding of participants and/or health care providers (Friedman *et al.* [7] and Redmond and Colton [8]; see **Blinding**). However, even when the study participants and those administering treatment cannot be blinded, blinding of outcome assessment could still be possible and helps prevent detection bias [9, 10].

If blinding is not feasible, randomization should still be attempted, and the use of an uncontrolled (nonrandomized) design should be justified.

Choice of Study Population. The study population is defined by the specification of criteria for entry into the study. These criteria must be carefully developed in advance to allow the population under study to be adequately characterized and to facilitate the replication of the results of the study. The selection of the patient population ranges from the consideration of a homogeneous, well-defined population to a heterogeneous population that allows easier recruitment and permits broader generalization of study results. The choice of the population for study depends on the objectives of the study and the ability to detect a meaningful benefit in a well-defined

homogeneous patient population compared with the dilution of potential effects (depending upon mechanism of action) in a more heterogeneous population. For example, a treatment may work well in a relatively small proportion of the population. In such a case, if this subgroup is a relatively small group in the total study population, the study will fail to detect an overall effect.

Study Objectives and Sample Size. Sample size considerations for the RCT depend upon the trial objectives as well as the outcome measurement. The trial can be designed to compare a new treatment or regimen against a standard treatment or placebo or observation with the expectation that the new treatment will be better than the standard treatment or the placebo regimen. Or, the trial may be designed to assess whether a new treatment or intervention is no worse than the standard treatment or intervention by some specified margin. The required sample size for the trial depends upon the absolute or relative difference in the measurement of the primary outcome between the treatment and control groups that can be a binary response rate, a continuous measurement of change from baseline, or a failure or survival time (see phase III trial section for discussion of sample size considerations and [11–13] for methods).

In the design of a randomized controlled trial, the risks of committing either of the two common types of errors in hypothesis testing–based trial designs need to be controlled. A type I error is made if the study leads to a false rejection of the null hypothesis, H_0 , and a type II error is made if the study fails to establish the alternative hypothesis H_1 when it is indeed true. The probabilities of type I and type II errors, often denoted as α and β , respectively, should be minimized. The statistical power of a test is defined as the probability $1 - \beta$.

Single-Center Trials. The simplest RCT is one that is carried out at a single institution or center with patients treated uniformly according to a well-defined protocol under the direction of the same investigator. The patients (units of observation) generally constitute a fairly homogeneous group with respect to demographic and prognostic characteristics. The protocol defines the patient population under study, the criteria for entry, adequacy of design to answer the question under study, the randomization and blinding

schema, and procedures for implementing all aspects of protocol compliance, data quality and integrity, and patient outcomes.

Statistical theory assumes that there is a “population” of subjects who meet an agreed definition of eligibility, but, in practice, patients treated in a study are those who come to the attention of the investigator, meet the criteria, and agree to be randomized. Unfortunately, although the notion of randomization allows us to make inferences about the treatment for this group of patients, randomization within the study does not offer any help in generalizing the results beyond the limited population under study at that center. Therefore, it is important to measure and report characteristics of both patient and center that may provide information regarding generalizability. Furthermore, there may be some characteristic, possibly unknown, or some specific skills associated with the single center that may be uniquely related to the observed treatment effect at that center. Such effects can only be identified if the experiment is repeated at other centers.

Frequently, the number of patients available at a single center is inadequate to provide sufficient data to answer the question under study. Therefore, multiple centers and, hence, multicenter trials are needed to obtain the required number of patients in a timely manner.

Multicenter Trials. The choice of a multicenter trial, rather than multiple single-center studies, is to enable prospective planning for the combination of data from multiple centers [14, 15]. The multicenter trial requires that a common protocol be used at each center and that procedures are in place to ensure consistency in the application of the protocol and all measurements across centers, including patient inclusion/exclusion criteria, administration of randomization and stratification schema, definitions of treatment regimens/outcomes, prognostic factors, methods of patient classification, methods for outcome assessment, definition of protocol compliance, study conduct, and operational definitions. The protocol and investigator meetings, as well as procedures for the conduct and the monitoring of a multicenter trial, go a long way toward ensuring that the results of the trial will be interpretable and generalizable. The measurement of treatment effect is agreed on by all investigators, and the methods for combining the

results across individual centers are, ideally, specified as part of the protocol.

Adaptive Designs

While many randomized controlled trials are designed with a fixed sample size that is specified at the start of the study, new strategies for randomization have developed that include adaptive randomization procedures which allow new patients to be randomized on the basis of the results obtained prior to patient entry (e.g., Zelen’s play the winner rule [16]). The use of interactive voice recognition and computer-based randomization systems has enabled dynamic balancing of subjects at randomization to maintain balance among several stratification variables. These procedures are particularly useful in multicenter trials with small center sizes where balancing within center on multiple factors is not feasible.

Group Sequential Trial Designs

In sequential trials, enrollment and observation continue until a specified stopping boundary is crossed. Inference about the treatment effect in sequential designs incorporates the fact that the stopping time of the trial is random in order to avoid bias. Methods for the design and analysis of phase III trials using group sequential procedures with interim analysis are described in Proschan *et al.* [17]. The two most common approaches are (a) O’Brien–Fleming boundaries that permit early stopping for success with stringent α -level requirements and maintain almost all of the α level for the final planned look [18] and (b) Pocock boundaries that divide the α level equally among the planned looks making it easier to stop the trial early [19]. These approaches fall under the class of Lan–DeMets spending functions with different parameterizations [20].

Analysis Issues

The analysis of any randomized clinical trial follows from the design of the trial and is based on the comparison of the specified treatment effect in the new treatment group *versus* the treatment effect in the standard treatment or placebo group. The magnitude of efficacy differences among treatment groups can

be evaluated using appropriate measures such as the difference in response rates, means, hazard rates (*see Hazard and Hazard Ratio*), and median survival times along with the corresponding confidence intervals [7, 11–13].

Stratification in the randomization of patients to treatment groups is also incorporated into the analysis with stratification factors used as covariates in the appropriate models. We note that in multicenter trials, center effects are considered as fixed or random effects [14]. However, if the center size or class size for a level of stratification variable is relatively small, power can be reduced when the stratification is incorporated into the primary analysis.

When repeated measurements of response and/or covariates are taken on subjects over the time course of the trial, analysis of the trial data should incorporate the positive correlations among the measurements on the same subject (*see Repeated Measures Analyses*).

Intention to treat analyses that include all randomized patients regardless of compliance to protocol or missing data are generally preferred to obtain an unbiased test of the treatment effect based on the randomization (*see Intent-to-Treat Principle*). Lack of compliance and missing data that may be related to the treatment under study can lead to bias in the estimated treatment effect (*see Compliance with Treatment Allocation*). Harrington [1] discusses some of these choices for analysis and implications for interpretation of results.

Discussion

This article provides a brief overview of many of the statistical issues in the design and analysis of RCTs that can impact the results and interpretation of the trial. There are many excellent references that provide detailed discussions of these issues: Harrington [1] for a brief tutorial and reference books by Piantadosi [13], Friedman *et al.* [7], Pocock [19], Fleiss *et al.* [12], Chow and Liu [11] among others.

New trial designs and methods are focused on reducing the numbers of patients enrolled in randomized trials and on shortening the time to decision making. For example, Bayesian methods (*see Bayesian Statistics in Quantitative Risk Assessment*) are now being used to obtain regulatory approval for

new drugs based on randomized clinical trials. Berry [21] provides a recent review of these Bayesian approaches and recent experience in practice.

References

- [1] Harrington, D.B. (2000). The randomized clinical trial, *Journal of the American Statistical Association* **95**, 312–315.
- [2] Meier, P. (1989). The biggest public health experiment ever: the 1954 field trial of the Salk poliomyelitis vaccine, in *Statistics: A Guide to the Unknown*, 3rd Edition, J.M. Tanur, F. Mosteller, W.H. Kruskal, E.L. Lehmann, R.F. Link, R.S. Pieters & G.R. Rising, eds, Duxbury.
- [3] Prentice, R.L. (1994). Surrogate endpoints in clinical trials: definition and operational criteria, *Statistics in Medicine* **8**, 431–440.
- [4] Fleming, T.R. & DeMets, D.K. (1996). Surrogate endpoints in clinical trials: are we being misled? *Annals of Internal Medicine* **7**, 605–613.
- [5] International Conference on Harmonization (1998). E9: guidance on statistical principles for clinical trials, *Federal Register* **63**(179), 49583–49598.
- [6] European Medicines Agency, Committee for Medicinal Products for Human Use (2005). *Guideline on the Choice of the Non-Inferiority Margin*, at <http://www.emea.europa.eu/pdfs/human/ewp/215899en.pdf>.
- [7] Friedman, L.M., Furberg, C.D. & DeMets, D.L. (1998). *Fundamentals of Clinical Trials*, 3rd Edition, Springer, New York.
- [8] Redmond, C. & Colton, T. (2001). *Biostatistics in Clinical Trials*, John Wiley & Sons.
- [9] Juni, P., Altman, D.G. & Egger, M. (2001). Assessing the quality of controlled clinical trials, *British Medical Journal* **323**, 42–46.
- [10] Schulz, K.F. & Grimes, D.A. (2002). Blinding in randomized trials: hiding who got what, *The Lancet* **359**, 696–700.
- [11] Chow, S.-C. & Liu, J.-P. (2003). *Design and Analysis of Clinical Trials: Concepts and Methodologies*, 2nd Edition, John Wiley & Sons, New York.
- [12] Fleiss, J.L., Levin, B. & Paik, M.C. (2003). *Statistical Methods for Rates and Proportions*, 3rd Edition, John Wiley & Sons, New York.
- [13] Piantadosi, S. (2005). *Clinical Trials: A Methodologic Perspective*, 2nd Edition, John Wiley & Sons, New York.
- [14] Goldberg, J.D. & Koury, K.J. (1992). Design and analysis of multicenter trials, in *Statistical Methodology in the Pharmaceutical Science*, D.A. Berry, ed, Marcel Dekker, New York, pp. 201–237.
- [15] Friedman, H. & Goldberg, J.D. (1996). Meta-analysis: an introduction and point of view, *Hematology* **23**, 917–928.
- [16] Zelen, M. (1974). Play the winner rule and the controlled clinical trial, *Journal of the American Statistical Association* **64**, 131–146.

- [17] Proschan, M.A., Lan, G.K.K. & Wittes, J.T. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*, Springer, New York.
- [18] O'Brien, P.C. & Fleming, T.R. (1979). A multiple testing procedure for clinical trials, *Biometrics* **35**, 549–556.
- [19] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*, John Wiley & Sons, New York.
- [20] DeMets, D.L. & Lan, K.K.G. (1994). Interim analysis: the alpha spending function approach, *Statistics in Medicine* **13**, 1341–1352.
- [21] Berry, D. (2006). Bayesian clinical trials, *Nature Reviews: Drug Discovery* **5**, 27–36.

Related Articles

Comparative Efficacy Trials (Phase III Studies)

Cost-Effectiveness Analysis

Efficacy

Meta-Analysis in Clinical Risk Assessment

JUDITH D. GOLDBERG AND ILANA
BELITSKAYA-LEVY

Recurrent Event Data

Byar [1] presents a medical study of patients who had bladder tumors upon entry. The initial tumors were removed using surgical techniques and patients were classified according to treatment (group 1 = placebo, group 2 = thiotepa, group 3 = pyridoxine). Throughout the study, patients experienced multiple recurrences of the tumors. Each time there was a recurrence, the tumors were again removed using a surgical intervention. The times (in months) of the tumor recurrences are presented in Wei *et al.* [2]. One aspect the researchers were interested in was the effectiveness of thiotepa based on the recurrence times of the tumors.

In the aforementioned study, the patients experienced an event of interest multiple times, that being the recurrence of tumors. The resulting data is termed *recurrent event data*. Recurrent events are ordered and each event of interest is treated as being of the same nature. Such data is considered a special case of multivariate lifetime data (*see* **Dependent Insurance Risks; Lifetime Models and Risk Assessment**).

Recurrent event data occurs in a wide array of disciplines. Examples in survival analysis (*see* **Mathematics of Risk and Reliability: A Select History**) include the reoccurrence of asthma attacks, seizures, tumors, and infections. In reliability, the events could be the breakdown of brake systems in city buses, bugs that occur in software programs, and the repeated fixing of aircraft. In the social and political sciences, examples include the recidivism rate of criminals, absenteeism of employees, and the occurrence of geographical conflicts.

There are several issues to consider when analyzing recurrent event data. These include the correlation of the event-of-interest calendar times within units, the effects of interventions that are applied to bring units back to “operational condition”, the impact of accumulating event occurrences on a unit, and the effects of possibly time-dependent covariates.

Many approaches for studying recurrent event data have been proposed in the literature. These include parametric, nonparametric, and semiparametric models both in the classical and Bayesian framework. Treatment of many of these approaches can be found in the books of Hougaard [3], Kalbfleisch and Prentice [4], Rigdon and Basu [5], Therneau

and Grambsch [6], Nelson [7], and Ascher and Feingold [8].

Many models are intensity based. To incorporate the correlation amongst event-of-interest calendar times researchers use a variety of strategies in these intensity-based models (*see* **Default Correlation; Default Risk**). These include using time-dependent covariates, the use of effective age (EA) processes, and the inclusion of random effect components referred to as *frailties*. Other authors do not specify any structure for the correlation. Instead, they provide an appropriate adjusted estimate for the variance–covariance matrix. Models often differ in their definition of the at-risk process to reflect the nature of recurrent events. The use of models based on mean rate functions is given as an alternative to those that are intensity based. They impose fewer assumptions and are presented as having better interpretability.

This article describes the models of Peña and Hollander [9] (EA model), Prentice, Williams, and Peterson [10] (PWP model), Wei, Lin, and Weissfeld [2] (WLW model), and those studied by Lin, Wei, Yang, and Ying [11] (LWYY model).

The article gives the mathematical setup and describes several classes of models used to analyze recurrent event data in the section titled “General Classes of Models”. It examines the bladder tumor data of Byar [1] in the section titled “Illustrative Example”. Some conclusions and other topics to be considered when analyzing recurrent event data are presented in the section titled “Concluding Remarks”.

General Classes of Models

We consider a study with n units. Each i th unit is observed over a time interval $[0, \tau_i]$, where τ_i is a right censoring (*see* **Imprecise Reliability**) random variable. We assume that the censoring random variable is independent and noninformative (*see* Andersen *et al.* [12, p. 138 and 171]). For each i th unit, we observe events of interest at calendar times $0 \equiv S_{i,0} < S_{i,1} < \dots$. Associated with these calendar times are the gaptimes, $T_{i,k} = S_{i,k} - S_{i,k-1}$. The counting process, $N_i(s)$ is the number of observed events of interest over the time interval $[0, s]$.

For a continuous distribution function $F(s)$ with associated density function $f(s)$, define the hazard rate function (*see* **Mathematical Models of Credit**

Risk; Repair, Inspection, and Replacement Models; Reliability Data) $\lambda(s) = f(s)/(1 - F(s))$, with the convention that $0/0 = 0$. A heuristic interpretation is

$$\lambda(s) ds \approx \text{probability of an event occurring within the next instant of time} \quad (1)$$

The cumulative hazard function is defined as $\Lambda(s) = \int_0^s \lambda(w) dw = -\log(1 - F(s))$. The survivor function (see **Hazard and Hazard Ratio**) is defined as $\bar{F}(s) = 1 - F(s)$.

EA Model

A general class of models for analyzing recurrent event data is given by Peña and Hollander [9]. The class of models assumes the intensity process that has the form

$$\lambda_i(s) = Y_i(s)Z_i\lambda_0(\mathcal{E}_i(s))\rho(N_i(s-); \alpha)\psi(\beta^t X_i(s)) \quad (2)$$

where $Y_i(s) = I(\tau_i \geq s)$ is the at-risk process. It is of interest to note that the value of $Y(\cdot)$ remains one as events of interest occur.

The baseline hazard function is perturbed by the observable and predictable EA process $\{\mathcal{E}_i(s)|0 \leq s \leq \tau_i\}$. Figure 1 represents the EA process for unit 1 in a study. The EA process between $S_{1,0}$ and $S_{1,2}$ is $\mathcal{E}_i(s) = s$. At the time of the first event of interest, $S_{1,1}$ the unit is “minimally repaired”. A “perfect repair” then occurs to the unit at the time of the second event of interest, $S_{1,2}$. The EA process then is represented by a nonlinear function between $S_{1,2}$ and $S_{1,3}$ and also between $S_{1,3}$ and $S_{1,4}$. After the third event of interest, a repair is performed that is between minimal and perfect. The time of the fourth event of interest, $S_{1,4}$ is not observed and is censored at τ_1 .

The frailty (see **Risk Classification/Life**) Z_i allows the researcher to model heterogeneity among the units and acts as a random-effect component. The $\rho(\cdot; \alpha)$ function permits accounting of the weakening or strengthening impact from the accumulating event occurrences. The function $\psi(\cdot) : \mathbb{R} \mapsto \mathbb{R}^+$ links the effects of concomitant variables to the intensity rate.

The generality of this class of models was demonstrated in Peña and Hollander [9] where they showed that many existing models in reliability and survival analysis are subsumed by this class. For instance, it includes models by Andersen and Gill

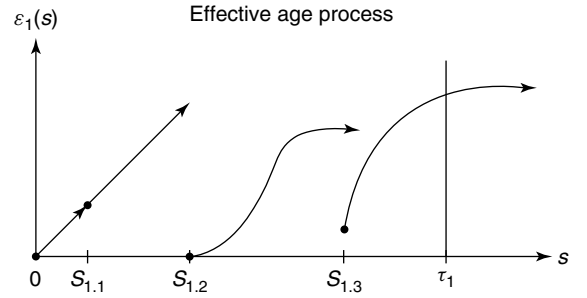


Figure 1 Graph of effective age versus calendar time for unit 1

[13], Prentice *et al.* [10], Kijima [14], Lawless [15], Aalen and Husebye [16], Dorado *et al.* [17], and Last and Szekli [18].

When $\mathcal{E}_i(s) = s$, $Z_i = 1$, and $\rho(N_i(s-); \alpha) = 1$, the EA general class of models reduces to the model of Andersen and Gill (AG) model [13]. The correlation amongst the calendar times is incorporated in the model using time-varying covariates. Therneau and Grambsch [6] discuss implementation of the AG model using SAS and Splus. They advocate a jackknife estimate of the variance-covariance matrix to adjust for the correlation in the calendar times within units.

The general class of models is studied under a fully parametric specification for the baseline hazard rate function by Stocker and Peña [19]. A treatment of the semiparametric model is given by Peña *et al.* [20]. In both papers, the authors use a reformulation of the EA model that incorporates both calendar and gaptime timescales. This reformulation yields a process that can be viewed as a generalized at-risk process. Stocker and Peña [19] demonstrate how this reformulation viewed in terms of gaptime is useful in verifying regularity conditions needed to prove asymptotic properties of parameter estimates. Peña *et al.* [20] use the reformulation to construct a methods of moments estimator for $\Lambda(\cdot|\alpha, \beta)$ in the gaptime timescale. More details can be found in the above references [19, 20].

PWP Model

Prentice *et al.* [10] model recurrent event data utilizing event-specific stratum. A unit at time s is in the k th stratum if it has experienced $k - 1$ events of interest. The use of stratum creates a definition of

the at-risk process that is different from that of the EA class of models. The PWP model has an at-risk process that changes as units experience events of interest, while the at-risk process for the EA class of models remains one. Prentice *et al.* [10] suggest using semiparametric hazard rate functions of the form

$$\lambda(s) = \lambda_{0k}(s) \exp(\beta_k^t X(s)) \quad (3)$$

and

$$\lambda(s) = \lambda_{0k}(s - S_{i,N_i(s)}) \exp(\beta_k^t X(s)) \quad (4)$$

PWP models that use hazard rates of the form given in equations (3) and (4) will be referred to respectively as *PWP calendar time* and *PWP gaptime*. Note the index k allows the hazard rate functions and regression parameters to differ from stratum to stratum. Partial likelihood functions can be formed for either choice of hazard rate function. Maximization of these likelihood functions gives estimates of the regression parameters. Large sample theory can be employed to construct inference procedures. Implementation of this model using SAS and Splus is given in Therneau and Grambsch [6].

WLW Model

Wei *et al.* [2] presented semiparametric methods to analyze multivariate lifetime data. They model each event of interest with a Cox proportional hazards model (see **Comparative Risk Assessment; Competing Risks**). This marginal modeling does not specify a form for the correlation between calendar times within units. This differs from the EA and PWP class of models that use EA processes.

For the k th event of interest ($k = 1, 2, \dots, K$) of the i th subject the intensity process is

$$\lambda_{ik}(s) = Y_{ik}(s) \lambda_{0k}(s) \exp(\beta_k^t \mathbf{X}_{ik}(s)) \quad (5)$$

where $Y_{ik}(s) = I(S_{i,k-1} \leq s < S_{i,k})$ is the at-risk process for the i th subject with respect to the k th event of interest. β_k is a q -dimensional event-specific parameter with associated covariate process $\mathbf{X}_{ik}(s)$. The unknown parameter estimates are found by maximizing the event-specific Cox partial likelihoods. Wei *et al.* [2] provide large sample theory and give an estimate of the variance-covariance matrix. Therneau and Grambsch [6] discuss implementation of these models using SAS and Splus.

Mean Rate Models

Many modeling approaches of recurrent event data have focused on the intensity-based methods of Andersen and Gill [13]. Denote $dN(s) = N(s-) - N((s + \Delta s)-)$ and $\mathcal{H}_{s-}$ as the history of the process just before time s ($\mathcal{H}_s$ is called a *filtration*). These models impose a nonhomogeneous Poisson process (see **Product Risk Management; Testing and Warranties; Reliability Growth Testing**) structure on $dN(s)$, where the intensity function is given by

$$E(dN(s)|\mathcal{H}_{s-}) = \lambda(s) ds \quad (6)$$

This assumption is seen by many authors as restrictive and interpretability of $\lambda(\cdot)$ maybe difficult for practitioners. In the AG model (see [13]), the correlation between event calendar times is modeled using time-dependent covariates. This assumption may be wrong and there is no way to check whether it is correct. Many authors have advocated the use of the mean rate function, denoted by $\mu(s|\mathbf{X}(s))$. It is defined as

$$\mu(s|\mathbf{X}(s)) = \int_0^s \exp(\beta^t \mathbf{X}(w)) d\mu_0(w) \quad (7)$$

or heuristically as

$$\begin{aligned} d\mu(s|\mathbf{X}(s)) &= E(dN(s)|\mathbf{X}(s)) \\ &= \exp(\beta^t \mathbf{X}(s)) d\mu_0(s) \end{aligned} \quad (8)$$

where $\mu_0(s)$ is an unknown continuous function. If $\mathbf{X}(s)$ consists only of external covariates (see Kalbfleisch and Prentice [4, p. 123]) then $\mu(s|\mathbf{X}(s))$ is interpreted as the mean function of recurrent events. The resulting model is termed *proportional rates model*. If the covariate process is time invariant then the model is termed the *proportional means model*. Lin *et al.* [11] study these models. They give appropriate score functions for estimation of $\mu_0(\cdot)$ and β . Asymptotics are proven using modern empirical process theory as outlined in van der Vaart and Wellner [21]. Appropriate inferential procedures for $\mu(s|\mathbf{X}(s))$ and β are given. They also provide graphical and numerical techniques for model checking.

Illustrative Example

The bladder tumor data of Byar [1] were modeled with those models described in the section titled

“General Classes of Models”. Only the 85 patients who were either from the placebo or thiotepa treatment groups were considered. A maximum of four recurrences per patient were recorded. The data is given in Wei *et al.* [2]. The covariates for each patient are a treatment indicator (placebo or thiotepa), the size in centimeters of the largest initial tumor, and the number of initial tumors. Our main interest was comparing the placebo versus thiotepa groups. The models used were AG, EA, WLW, PWP calendar time, PWP gaptime, and the mean rate models.

Uncommon and common regression coefficients were incorporated into the WLW and PWP models. Therefore, equations (3)–(5) have link functions of the form $\exp(\beta_k' X)$ and $\exp(\beta' X)$ resulting in six different models. For the EA model, $\rho(k; \alpha) = \alpha^k$ was used. For demonstration purposes, two choices were examined for the EA process: $\mathcal{E}_i(s) = s$ (minimal repair) and $\mathcal{E}_i(s) = s - S_{i, N_i(s-)}$ (perfect repair). It was assumed $Z_i \sim \text{gamma}(\eta, \eta)$ ($E(Z_i) = 1$ and $\text{Var}(Z_i) = 1/\eta$) for $i = 1, 2, \dots, 85$. This choice is common in frailty models. The results of all the analyzes are found in Tables 1–5.

Tables 1–3 contain the parameter estimates, standard errors, test statistics, and p values associated with the treatment covariate (placebo or thiotepa) for the WLW, PWP calendar time, and PWP gaptime

Table 1 Table of parameter estimates, standard errors, test statistics, and p values for the WLW model using common regression coefficients. The parameter Trt corresponds to the treatment covariate for each marginal model (1, 2, 3, and 4)

Variable	Estimate	Standard error	Test statistic	p value
Trt 1	−0.52	0.31	2.83	0.0923
Trt 2	−0.62	0.36	2.90	0.0887
Trt 3	−0.70	0.42	2.84	0.0918
Trt 4	−0.65	0.50	1.77	0.1839

Table 2 Table of parameter estimates, standard errors, test statistics, and p values for the PWP calendar time model. The parameter Trt corresponds to the treatment covariate for each stratum (1, 2, 3, and 4)

Variable	Estimate	Standard error	Test statistic	p value
Trt 1	−0.52	0.32	2.69	0.1012
Trt 2	−0.46	0.41	1.28	0.2581
Trt 3	0.12	0.67	0.03	0.8617
Trt 4	−0.04	0.79	0.00	0.9592

Table 3 Table of parameter estimates, standard errors, test statistics, and p values for the PWP gaptime model. The parameter Trt corresponds to the treatment covariate for each stratum (1, 2, 3, and 4)

Variable	Estimate	Standard error	Test statistic	p value
Trt 1	−0.52	0.32	2.69	0.1012
Trt 2	−0.26	0.41	0.41	0.5224
Trt 3	0.22	0.55	0.16	0.6873
Trt 4	−0.20	0.64	0.09	0.7613

models when uncommon regression coefficients were used. The test statistics are χ^2 distributed with one degree of freedom under the null hypothesis. For all three models, the test statistics are not statistically significant.

Table 4 contains the parameter estimates and standard errors (in parentheses) for common regression coefficients for the WLW and PWP calendar time models. Also included in Table 4 are the parameter estimates and standard errors for the AG and mean rate models. Table 5 contains the parameter estimates and standard errors (in parentheses) for the PWP gaptime and EA models. In Tables 4 and 5, the parameter vector $\beta = (\beta_1, \beta_2, \beta_3)^t$ corresponds to the covariates: treatment group, size of the initial tumors, and

Table 4 Table of parameter estimates and corresponding standard errors for the AG, mean rate, WLW, and PWP calendar time models

Parameter	AG	Mean rate	WLW	PWP calendar
α	NA	NA	NA	NA
η	NA	NA	NA	NA
β_1	−0.47 (0.20)	−0.47 (0.26)	−0.58 (0.20)	−0.33 (0.21)
β_2	−0.04 (0.07)	−0.04 (0.08)	−0.05 (0.07)	−0.01 (0.07)
β_3	0.18 (0.05)	0.18 (0.06)	0.21 (0.05)	0.12 (0.05)

Table 5 Table of parameter estimates and corresponding standard errors for the PWP gaptime and EA models. The effective age processes used in the EA models are perfect repair (EA perfect) and minimal repair (EA minimal)

Parameter	PWP gap	EA perfect	EA minimal
α	NA	0.98 (0.07)	0.79 (0.13)
η	NA	∞	0.97
β_1	-0.28 (0.21)	-0.32 (0.21)	-0.57 (0.36)
β_2	0.01 (0.07)	-0.02 (0.07)	-0.03 (0.10)
β_3	0.16 (0.05)	0.14 (0.05)	0.22 (0.10)

number of initial tumors. The results for the EA models are also found in Peña *et al.* [20]. The parameter estimates obtained using the AG and mean rate models are exactly the same. These estimates are found by solving the same score equations. The standard errors are computed differently. This results in a significantly larger standard error for the treatment indicator parameter β_1 for the mean rate model. One degree of freedom χ^2 tests indicate that thiotepa is effective at the 0.05 significance level when using the AG model.

We observe how close the parameter estimates and corresponding standard errors are for the WLW and EA minimal models. This is also seen when comparing the PWP calendar time and the EA Perfect models. This points to how the differences in models may be attributed to the EA processes.

Concluding Remarks

This article has discussed a specific form of multivariate lifetime data termed *recurrent event data*. The nature of this data requires modeling strategies to incorporate several challenging aspects. Of particular consideration is the correlation of within subject event-of-interest calendar times. A further complication is the fact that many studies have right-censored data. Four general classes of models were described that approach modeling of this data in unique ways. The bladder tumor data of Byar [5] was modeled using the different models presented in the section titled “General Classes of Models” for illustrative purposes.

Many experiments yield recurrent event data along with other data structures that were not formally discussed in this article. Often, recurrent event data is coupled with another event of interest termed a *terminal event*. The appropriate joint modeling of

the recurrent event process and the terminal event in the presence of right censoring is an important statistical challenge. The right censoring in this article was assumed to be independent and noninformative. Dependent and informative censoring are also found in many studies. These are just some of the interesting aspects that need to be addressed when studying recurrent event data.

References

- [1] Byar, D. (1980). The Veterans administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine, and topical thiotepa, in *Bladder Tumors and Other Topics in Urological Oncology*, M. Pavones-Macaluso & P. Smith, eds, Plenum, New York, pp. 363–370.
- [2] Wei, L.J., Lin, D.Y. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American Statistical Association* **84**, 1065–1073.
- [3] Hougaard, P. (2000). *Analysis of Multivariate Survival Data, Statistics for Biology and Health*, ISBN 0-387-98873-4, Springer-Verlag, New York.
- [4] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, Wiley Series in Probability and Statistics, ISBN 0-471-36357-X, 2nd Edition, Wiley-Interscience [John Wiley & Sons], Hoboken.
- [5] Rigdon, S.E. & Basu, A.P. (2000). *Statistical Methods for the Reliability of Repairable Systems*, Wiley Series in Probability and Statistics: Applied Probability and Statistics, ISBN 0-471-34941-0, John Wiley & Sons [A Wiley-Interscience Publication], New York.
- [6] Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*, Statistics for Biology and Health, ISBN 0-387-98784-3, Springer-Verlag, New York.
- [7] Nelson, W.B. (2003). *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*, ASA-SIAM Series on Statistics and Applied Probability, ISBN 0-89871-522-9, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Vol. 10.
- [8] Ascher, H. & Feingold, H. (1984). *Repairable Systems Reliability: Modeling, Inference, Misconceptions and Their Causes*, Lecture Notes in Statistics, ISBN 0-8247-7276-8, Marcel Dekker, New York, Vol. 7.
- [9] Peña, E. & Hollander, M. (2004). Models for recurrent events in reliability and survival analysis, *Mathematical Reliability: An Expository Perspective*, Kluwer Academic Publishers, pp. 105–118.
- [10] Prentice, R., Williams, B. & Peterson, A. (1981). On the regression analysis of multivariate failure time data, *Biometrika* **68**, 373–379.

- [11] Lin, D.Y., Wei, L.J., Yang, I. & Ying, Z. (2000). Semi-parametric regression for the mean and rate functions of recurrent events, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **62**, 711–730.
- [12] Andersen, P., Borgan, O., Gill, R. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer Series in Statistics, ISBN 0-387-97872-0, Springer-Verlag, New York.
- [13] Andersen, P. & Gill, R. (1982). Cox's regression model for counting processes: a large sample study, *Annals of Statistics* **10**, 1100–1120.
- [14] Kijima, M. (1989). Some results for repairable systems with general repair, *Journal of Applied Probability* **26**, 89–102.
- [15] Lawless, J. (1987). Regression methods for Poisson process data, *Journal of the American Statistical Association* **82**, 808–815.
- [16] Aalen, O. & Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes, *Statistics in Medicine* **10**, 1227–1240.
- [17] Dorado, C., Hollander, M. & Sethuraman, J. (1997). Nonparametric estimation for a general repair model, *Annals of Statistics* **25**, 1140–1160.
- [18] Last, G. & Szekli, R. (1998). Asymptotic and monotonicity properties of some repairable systems, *Advances in Applied Probability* **30**, 1089–1110.
- [19] Stocker, R. & Peña, E. (2007). A general class of parametric models for recurrent event data, *Technometrics* **49**, 210–220.
- [20] Peña, E., Slate, E. & González, J. (2007). Semiparametric inference for a general class of models for recurrent events, *Journal of Statistical Planning and Inference* **137**, 1727–1747.
- [21] van der Vaart, A.W. & Wellner, J.A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer Series in Statistics, ISBN 0-387-94640-3, Springer-Verlag, New York.

RUSSELL S. STOCKER

Reinsurance

The Role of Reinsurance for Direct Insurers

Most individuals and organizations prefer to transfer some of the risks they stand to another entity, paying a certain agreed price. Insurance companies are no exception.

Under a contract of insurance, the insurer promises to indemnify the policyholder on the occurrence of an uncertain specified event. The policyholder agrees to pay the premium. The very same thing happens in reinsurance contracts.

Reinsurance is a form of insurance, but it has many important distinctive features. These differences result from the fact that it is *insurance for insurers*. Reinsurance contracts are celebrated between a direct *insurer* and a *reinsurer*, traditionally, “on the basis of utmost good faith”, with the purpose of transferring to the latter part of the risks assumed by the former, in its business. They are a vital instrument to the management of risk and capital. Reinsurers have contractual obligations only to direct insurers, not to policyholders.

Before reinsurance widespread, insurance companies used to constrain their acceptances only to amounts they could bear. When a risk was too large for an insurer to take alone (*see Large Insurance Losses Distributions; Ruin Probabilities: Computational Aspects*), coinsurance was practiced and each one of the direct underwriters would take a share. The first specialized company, the Kölnische Rück, was founded in the nineteenth century, after a catastrophic fire devastated Hamburg, in 1842.

Every insurer recurrently faces the same dilemma: to accept business that can induce a severe result on its claims costs, or to reject it and take the consequences of losing place to other insurers – and the smaller the company, the greater the constraints on the risks it can accept. The problematic risks are those carrying either the possible occurrence of very large individual losses or the possible accumulation of losses from one single event – normally because individual risks are not mutually independent. An insurance portfolio has to include numerous similar and independent risks to be considered balanced, so that the central limit theorem applies. Unfortunately, that is often not the case, and some extra loadings are

needed for claims fluctuations, but the market limits the size of the premiums.

The above are the main causes why reinsurance is required. Although nonlife business predominates (many life policies have small insurance risk), its benefits apply to both life and nonlife insurance and can be summarized as follows (*cf.* [1, 2]):

- To give protection against large individual losses and accumulations of losses that may endanger the direct insurer’s survival. By covering large sums insured and highly exposed risks, relative to income and reserves, the direct insurer intends to reduce the probability of ruin.
- To improve the balance of the portfolio and to stabilize underwriting results. By assuming a part of the risk of random fluctuations of small and medium size claims, reinsurance limits the instability of the annual aggregate claims.
- To aid insurers meet their solvency requirements and to provide them with additional underwriting capacity to accept individual risks and types of business otherwise unbearable (and to spread their overheads). Without reinsurance the potential market would be very limited to the majority of the underwriters and some risks would possibly not be insurable.
- To allow direct insurers to free themselves from risks that they do not want to stand alone or which they would even prefer to avoid. The flexibility in respect of the acceptable risks for the amount of capital available is increased.
- To guarantee liquidity and income, by means of financial reinsurance (also known as *finite risk* or *capital-motivated reinsurance*). It is transacted mainly to accomplish financial purposes and combines risk transfer with elements of risk finance. Financial reinsurance helps direct insurers to deal with difficult capital-planning problems and in the financing of acquisitions. More efficient use of capital and better returns on investments are obtained.
- To advise on the planning of reinsurance programmes and on the handling of very large losses – the reinsurer usually reserves the right to participate in the negotiations, contributing to an efficient adjustment of the claims.
- To offer many kinds of other services, fundamental in particular, to new companies in developing countries that lack experience and have

not enough information to rate risks and propose reasonable prices. As reinsurers collect data all over the world, they are able to help in the assessment of special and unusual risks, to give consultation in loss prevention and capital investment, to provide loss adjustment support, to perform actuarial work, to assist in the development and pricing of new products, to find cooperation associates, and even to recruit and train personnel.

Outside the field of reinsurance protection are the issues of inadequate pricing, uncontrolled costs, and bad investment policy. Reinsurance cover does not prevent direct insurers from suffering the consequences of their failures; in some cases they may continue to be very limited in the risks they can accept.

Looking from the reinsurers' perspective, they have to supply the cover demanded by their clientele and at the same time they must structure and preserve their own portfolio, risks being in line with the existing capital. This is accomplished by dispersing activities geographically and throughout the classes of business. A reinsurer may also reinsure risks it has accepted, by means of a retrocession contract. The ceding reinsurer is then the retrocedent and the other one is the retrocessionaire.

Reinsurance Forms

All reinsurances can be classified into two types: proportional and nonproportional. In the last years alternative risk transfer (ART) techniques have appeared, e.g., life and no life finite (*see Nonlife Loss Reserving*) risk reinsurance and property/casualty multiyear, multirisk, reinsurance, which are extensions of the conventional types.

Proportional reinsurance, quota share, and surplus, were the first to appear, developing from coinsurance. In this type of reinsurance, cessions are linked to the sums insured and the direct insurer and the reinsurer share premiums and losses at a contractually defined ratio. The reinsurer accepts a fixed share of the liabilities assumed by the cedent under the original contract, and receives the same proportion of the original premium, minus a commission. This commission, used to compensate the insurer for expenses met for the benefit of both, is usually defined as a percentage of the ceded premium.

Under a *quota share*, a fixed proportion of every risk accepted by the primary insurer is ceded to the

reinsurer. This proportion, *the quota*, establishes how premiums and losses are distributed between them. The pros of this kind of reinsurance are its simplicity, the low cost administration, and a certain identity between insurer and reinsurer. The main disadvantage is that quota share does not help to balance the portfolio: by reinsuring the same proportion of all risks, the relative variability and the incurred loss ratio on the retained portfolio are not reduced; it can neither limit adequately the threat of peak risks nor provide enough protection against the accumulation of losses. Since all the underwritten risks are covered, the reinsurer gets a share of the premium for policies that really do not need to be reinsured. In spite of this, quota share is often an appropriate part of the reinsurance programme, particularly when the company is entering a new market.

Surplus reinsurance, possibly the most common form of reinsurance, is also proportional, but the reinsurer does not participate in all the risks. The direct insurer retains every risk up to a certain amount, the *retention*, and the reinsurer is obliged to accept only the surplus, that is to say, the amounts accepted by the insurer above that retention. When the sum insured is below the retention, the insurer retains the entire risk. For each reinsured risk, the ratio between the retained and the ceded amounts determines how the premiums and losses are distributed to them. The ceding company can choose different retention limits directly related to the different degrees of exposition of the reinsured risks, and there is an upper limit to the reinsurer's obligation for each of them. This limit is usually defined as a certain multiple of the direct insurer's retention, known as *lines*. The cover is split into several surpluses, when it is more easily placed that way. When the sum insured exceeds the surplus, the direct insurer must either take the rest or plan additional suitable reinsurance cover. Surplus reinsurance has the merit of balancing the cedent's portfolio. By reducing the range of possible retained losses and the relative variability of costs, it also limits the highest retained exposures and allows adjusting the amount of risk. Like quota share, surplus reinsurance gives no effective protection against the accumulation of losses.

Nonproportional reinsurance, excess of loss and stop loss, won popularity in the last decades. Cessions are no longer linked to the sums insured, but to the losses, and there is not a predetermined ratio for dividing premiums and losses between insurer

and reinsurer. The share of losses paid by each one depends on the amount of the loss incurred. There is an amount up to which the direct insurer pays all losses – the deductible, net retention, excess point, or priority – and the reinsurer pays all losses above it, also up to a defined cover limit. The contract forces the reinsurer to pay only when the reinsured portfolio or risk incurs a loss that exceeds the deductible. The difficulty is to set premiums that are fair to both parties. In rating the price, the reinsurer tries to get an adequate part of the primary premium, considering the loss experience of past years (experience rating) or, if events are rare, the losses to be expected from that sort of risk (exposure rating, normally based on other available information and expert opinion).

Excess of loss (XL) can be contracted under a risk basis (per risk XL) or under an occurrence basis (per event XL). Per risk XL protects against possible large losses produced by *one policy* and the reinsurer pays any loss in excess of the deductible. Per event XL protects against an accumulation of individual losses due to *a single event* and the reinsurer pays when the deductible is exceeded by the aggregate loss from any one occurrence.

When the deductible is low enough so that it is expected that some claims will activate the contract every year, the excess of loss is named working excess of loss (WXL) cover. Catastrophe excess of loss (CatXL) is defined per event and it is used in property reinsurance to protect direct insurers against accumulation of losses due to catastrophic occurrences (earthquakes, floods, etc). A clash cover, sometimes referred to as *unknown accumulation cover*, is an excess of loss that protects against losses due to unknown accumulations (e.g., the exposure to a larger loss than anticipated in such an occurrence).

XL reinsurance is very efficient to stabilize the results of the primary insurer, by reducing its exposures on individual risks, and can also be used to deal with the problem of accumulations and catastrophe risks. On the other side, some reinsurers do not accept contracts with relatively low retentions, which may produce numerous claims, and XL cover has often to be organized in layers, thus increasing the costs. There are variations of XL covers especially devoted to the largest claims, for example, ECOMOR reinsurance (*excédent du coût moyen relatif*).

Stop loss covers are intended to protect against fluctuations in annual loss experience, in a certain

class of business: the reinsurer pays if the *aggregate losses* for a year, net of other reinsurance covers, exceed the agreed deductible and is not relevant whether it is exceeded by one single large loss or an accumulation of small- and medium-sized losses. The deductible, now expressed as a function of the aggregate net losses, is settled either as a monetary limit (stop loss) or in terms of a proportion of premium income (XL ratio). Only stop loss reinsurance has the virtue to offer simultaneous protection against increases in the severity and the frequency of losses. Administration costs are lower, but premium rating poses important practical problems. The reinsurers always limit their responsibility and the loading they charge reduces the reserves of the direct insurers. They are often reluctant to underwrite these kinds of contracts because of the large amount of risk transferred – all the losses, great and small, determine their liability – in contrast to their limited means of influencing the exposure. There is a danger of the cedent manipulating the needed information, particularly when the business is more internationalized.

The *aggregate XL* is very similar to stop loss, except that only the losses exceeding a certain amount are considered when calculating the company's annual claims burden.

The term *umbrella cover* is used when several classes of business are protected by combining the contracts of the different classes into one reinsurance contract.

Placing Reinsurance

Both types of reinsurance, proportional and nonproportional, may be either facultative or obligatory, the two basic methods of placing reinsurance.

The *facultative method* was the first to be used, when reinsurance was mainly the mutual placing of individual risks between primary insurers. The word facultative means that the deal is optional: neither is the primary insurer obliged to offer the business nor is the reinsurer obliged to accept it. Like in direct insurance, a reinsurer offered a risk can examine it and decide whether or not to accept it. The facultative method has lost significance owing to the high administration costs and the inherent inadequacy when the direct insurer has to take, without delay, a risk in excess of its retention. It is used mostly in two situations: when the insurer has exhausted its

retention and the capacity provided by the obligatory reinsurance, but additional reinsurance is still needed, and when it underwrote risks that are not included in the existing reinsurance covers and need to be individually reinsured.

In *obligatory* (or *treaty*) *reinsurance* the direct insurer is obliged to cede and the reinsurer is obliged to accept an agreed share of the risks defined in the reinsurance treaty. A treaty is a contract between a direct insurer and one or more reinsurers, where the insurer agrees to cede certain of the written risks and the reinsurer agrees to reinsure them. The terms of the treaty stipulate the risks covered and the limits of the reinsurer's liability. In obligatory reinsurance, entire and precisely defined portfolios are ceded. The reinsurer neither can refuse to provide insurance protection for an individual risk falling within the scope of the treaty, nor can the direct insurer decide not to cede a risk to the reinsurer. Thus, once a treaty has been concluded, a direct insurer (or a reinsurer, under a retrocession treaty) can obtain reinsurance automatically for risks accepted and covered by that treaty. The automatic character of treaties, normally set annually, makes obligatory reinsurance less expensive than facultative.

Facultative/obligatory (or *open cover*) arrangements are similar to treaties, with the important difference that the ceding company decides which of the covered risks are in fact reinsured. It is used mostly when the sum insured exceeds the surplus and the direct insurer must either take the rest or plan additional suitable reinsurance cover. Facultative/obligatory reinsurance gives the ability to reinsure immediately the remaining surplus falling within the scope of the arrangement and therefore to accept without delay those large risks that the restrictions in the reinsurance treaties would not allow. *Obligatory/facultative* contracts function in the opposite way.

The Design of Reinsurance Programmes

When designing a reinsurance programme, somehow, there is an attempt to decide optimally on the type of reinsurance and on how much to reinsure, but what is meant by "optimal" depends on many factors and on the significance given to each of them: the current and future business models and resultant loss exposures, the financial strength and risk aversion, the market

conditions and opportunities. Altogether, they make of each company a single case, requiring a distinct reinsurance programme.

Gerathewohl *et al.* states that "... the 'ideal' type of reinsurance applicable in all cases does not exist" [3, p. 124], and the previous paragraphs pointed out that there is no type of reinsurance which is able to protect the primary insurers against *all* the causes; why then is reinsurance required?

Following the pioneer works of de Finetti and Borch, there are many theoretical results [4] in favor of this or that type of reinsurance, depending on the optimality criteria and the premium principle that have been chosen.

The main reason why the market practices show only a weak adherence to these theoretical results is the fact that nearly all the models assume that the premium principle is independent of the particular type of reinsurance under consideration. For instance, in Borch's famous result, proving that stop loss is the optimal form of reinsurance – in the sense that, for a fixed net reinsurance premium, it gives the smallest variance of the net retention – it is assumed that the loading coefficient on the net premium is not different from that in a conventional quota treaty. This is not a reasonable assumption and Borch himself said later [5, pp. 293–295]:

I do not consider this a particularly interesting result ... Do we really expect a reinsurer to offer a stop loss contract and a conventional quota treaty with the same loading on the net premium? If the reinsurer is worried about the variance in the portfolio he accepts, he will prefer to sell the quota contract, and we should expect him to demand a higher compensation for the stop loss contract.

Yet, the exceptions to this somewhat unrealistic approach are very unusual (see a few in [6]).

An example of the current practice is provided by Schmutz [7] in the form of a "pragmatic approach" for the designing of reinsurance programmes. The proposed method supplies a set of empirical rules that are exclusively applicable to property reinsurance, a branch where typically all forms and types of contracts are used. For instance, it is common to start by contracting a facultative proportional arrangement, to reduce the largest risks, followed by a surplus treaty, to homogenize the retained portfolio, and then a quota share on the retention of the surplus treaty and a WXL, to protect against major losses. As natural

hazards may occur, in some cases it will be wise to protect the per risk retention with a CatXL treaty.

As Gerathewohl *et al.* [3] explain, although the design of a reinsurance programme is essentially a management decision, there are various standards for determining the retentions, for instance, the premiums and shareholders' funds criteria and the claims ratios and fluctuations criteria. In practice, the variance is the most used measure for these fluctuations. Naturally, the final decision has always to balance the level of the direct insurer's retention with the desired level of business profitability.

References

- [1] Carter, R.L. (1979). *Reinsurance*, Kluwer Publishing, Brentford.
- [2] Swiss-Re (2002). *An Introduction to Reinsurance*, 7th edition, Zurich.
- [3] Gerathewohl, K., Bauer, W.O., Glotzmann, H.P., Hosp, E., Klein, J., Kluge, H. & Schimming, W. (1980). *Reinsurance Principles and Practice*, Verlag Versicherungswirtschaft e. V., Karlsruhe, Vol. 1.
- [4] Centeno, M.L. (2004). Retention and reinsurance programmes, *Encyclopedia of Actuarial Science* 3, 1443–1452.
- [5] Borch, K. (1969). The optimum reinsurance treaty, *ASTIN Bulletin* 5, 293–297.
- [6] Centeno, M.L. (1986). Some mathematical aspects of combining proportional and non-proportional reinsurance, in *Insurance and Risk Theory*, M. Goovaerts, F. e Vylder & J. Haezendonck, eds, D. Reidel Publishing Company, Holland.
- [7] Schmutz, M. (1999). *Designing Property Reinsurance Programmes – the Pragmatic Approach*, Swiss Re Publishing, Zurich.

Further Reading

Patrik, G.S. (2001). Reinsurance, in *Foundations of Casualty Actuarial Science*, 4th Edition, Casualty Actuarial Society, Virginia, Chapter 7, pp. 343–484.

Related Articles

Enterprise Risk Management (ERM)

Risk Measures and Economic Capital for (Re)-insurers

Ruin Theory

ONOFRE SIMÕES AND MARIA DE LOURDES
CENTENO

Relative Risk

Risk is often quantified by calculating risk ratios. The presentation of risk influences its perception. For example, one might be worried to hear that one’s lifestyle doubled the risk of a serious disease compared to some other way of life. However, the statement that the risk had increased from one in a million to two in a million might be less worrying. In the first case, it is a *relative* risk that is presented, in the second an *absolute* risk.

The relative risk measures how the probability of an event, typically of getting a disease, varies between two categories, namely, a group that is exposed to a risk factor and one that is not. The specific definitions are most easily understood in terms of a 2×2 table of exposure by disease as shown in Table 1. Therein, N denotes the population size and a, b, c , and d the absolute frequencies of the respective combinations of the levels of the risk factor (exposed or not exposed) and the disease factor (present or absent). The (absolute) risk of the disease within a subpopulation is defined as the proportion of the subgroup that have the disease, i.e., $r(\text{disease present} | \text{exposed group}) = a / (a + b)$ and $r(\text{disease present} | \text{nonexposed group}) = c / (c + d)$. Then the relative risk ratio (RRR), of disease comparing the exposed with the nonexposed subpopulation is given by the ratio $RRR = \frac{a/(a+b)}{c/(c+d)}$.

For example, a relative risk of coronary heart disease, comparing obese people with nonobese people, of two would be interpreted as a doubling of the risk of heart disease when putting on weight. Similarly, a relative risk of contracting measles comparing vaccinated children with nonvaccinated children of 0.5 would be interpreted as a protective effect.

While the relative risk is a popular choice for expressing the effect of exposure on disease, there are a number of other indices that can be used for

this purpose. They divide into measures that contrast event probabilities, for example, *absolute risk reduction* (see **Absolute Risk Reduction**), attributable fraction (see **Attributable Fraction and Probability of Causation**), and indices that contrast other measures of chance, for example, odds ratios (see **Odds and Odds Ratio**).

It is important to use an appropriate study design when estimating the relative risk in the population from a sample. Cross-sectional studies typically fix the sample size N and in *cohort studies* (see **Cohort Studies**), the sizes of the cohorts of exposed and non-exposed subjects (the row totals in Table 1) are under the control of the investigator. Both these restrictions allow the risk of disease within the exposure groups and hence the RRR to be estimated. In contrast, in a *case-control study* (see **Case-Control Studies**), the number of subjects with the disease and the number of subjects without the disease (the column totals in Table 1) are chosen by the investigator, making it impossible to estimate disease risks from the sample data. However, when the disease is rare, it is possible to use the odds ratio as an approximation (see **Odds and Odds Ratio**).

Related Articles

Epidemiology as Legal Evidence

History of Epidemiologic Studies

Logistic Regression

SABINE LANDAU

Table 1 Two-way classification of exposure by disease

		Disease		Total
		Present	Absent	
Risk factor	Exposed	a	b	$a + b$
	Not exposed	c	d	$c + d$
	Total	$a + c$	$b + d$	$N = a + b + c + d$

Reliability Data

The need for quantitative reliability assessment (*see* **Imprecise Reliability; Mathematics of Risk and Reliability; A Select History; Near-Miss Management; A Participative Approach to Improving System Reliability**) in industrial risk management has been dictated by major hazards legislation (*see* **Economic Criteria for Setting Environmental Standards**), concern for protection of the environment, and the safety of personnel working in or near to industrial installations. At a meeting of experts from all sectors of industry organized jointly by The Engineering and Physical Sciences Research Council and the Institution of Mechanical Engineers in 1996, the lack of good quality reliability and maintenance data was a common complaint. Ten years later, the situation is little different due primarily to uncertainties of reliability data requirements in different areas of reliability analysis, the different types of data and methods for their collection and analysis. However, a variety of reliability data sources are now featured on the Internet.

Basically, reliability data (*see* **Reliability Demonstration; Lifetime Models and Risk Assessment; Reliability Integrated Engineering Using Physics of Failure**) are required during the three main phases of an engineering system's lifetime:

1. in the design phase, as input to the models developed to predict the reliability, availability, and maintainability of safety and production systems;
2. prior to manufacture and operation, to support the preparation of reliability specifications, quality control procedures, test programs, and inspection and maintenance planning;
3. during operation, to monitor the efficacy of reliability performance and identify areas where hazards and production losses can be reduced by application of reliability based techniques such as reliability centered maintenance and risk-based inspection.

Generic statistics of equipment failure characteristics from in-house and published sources generally form the database for the first two areas. In-service experience provides more detailed, system-specific failure information suitable for ongoing reliability monitoring and feedback into the generic database.

The principal types of data required are known respectively as generic reliability data (*see* **Probabilistic Risk Assessment**) and in-service reliability data. In both cases, it is important to ensure that reliability data are based on clear definitions, and collected and analyzed to well-defined procedures. Vague descriptions of equipment, engineering, and functional attributes, system boundaries, and failure descriptors can lead to uncertainties in the results. The objective of this article is therefore to cover the basics of reliability data collection and analysis. Other publications cover the subject in more detail and include more detailed methodologies for generic and in-service reliability data analysis, its collection, and quality control [1–4].

Generic Reliability Data

The reliability assessment of an engineering system requires a variety of different types of generic data. A database of relevant reliability information is required related to the study objective – say, to predict whether the system will meet some previously defined reliability target. This could be the probability of failure on demand of a safety system or the expected steady-state availability of a manufacturing system. In all cases, the study aims to identify the relative importance of the different pieces of equipment in the system, the critical failure modes, and the distribution of hazards and/or downtime with respect to the different subsystems and equipment.

Typical data required are as follows:

- overall equipment failure rates
- principal failure modes and proportions
- common cause failure rates
- human error rates
- spurious trip rates
- fail-to-start probabilities
- active repair and waiting times.

Although most of these data classes are identical for both safety and availability studies, the emphasis is different. In safety studies, the principal objective is to identify the potentially dangerous equipment and component failure modes, whereas in availability studies, the main focus is on the equipment and failure modes that cause forced outage or production loss.

Generic reliability data can be obtained from a variety of sources. These include in-house information collected from operating plant, data tabulated in reliability data handbooks, and the variety of reliability data now available *via* the Internet. There is a general reluctance on the part of most organizations to publish reliability data since it is often viewed as commercially sensitive. In 1993, therefore, the IMechE Mechanical Reliability Committee's Data Acquisition Working Party carried out a review of existing generic reliability data sources. The results were published by the IMechE and reviewed in 1993 [5]. The 20 generic data sources are shown here in Table 1. Sources of electronic component and equipment data are also published from time to time by the Rome Air Development Center, for example [6].

As one progresses from data in the public domain to in-house sources, the extent of the information available increases and uncertainties diminish. The relevant information should now be assembled on a project database before more detail analysis is applied to determine the best estimates of equipment failure rates in the critical failure modes.

Determining best estimates from the very diverse and frequently small data available can be a daunting task. Generic reliability data available in the public domain has significant limitations and is frequently confined to a single figure of failure rate for a specific type of equipment. Besides the effect of different physical and environmental stresses, uncertainties arise from the imprecision of the item description and the operational, maintenance, and industrial conditions associated with the systems from which the generic failure rate was derived. Where there are a large number of failure rates available for a specific type of equipment, the range is likely to be very large. For example, Figure 1 illustrates the lognormal distribution of circuit breaker failure rates listed in one generic database which shows a factor of approximately 100 between highest and lowest estimates. This is quite typical for a wide range of equipment in manufacture, power production, transport, and military systems.

To derive a best-estimate failure rate (*see Repair, Inspection, and Replacement Models*) from such diverse information requires a structured approach and the application of engineering judgment to

initially define a reference point within the distribution of generic failure rates. The reference-point failure rate is generally assumed to be the median of the generic failure rate distribution and to reflect "average industrial conditions", since 50% of the observed failure rates will be above and 50% below the reference point. This is termed the *base failure rate* (λ_b). For circuit breakers, λ_b indicated by the data in Figure 1, is 4 failures/ 10^6 h.

The second step is to scale this base failure rate to reflect the different conditions of the system being assessed and the equipment failure mode(s) that could trigger system failure. A common model employed for electromechanical equipment is given by

$$\lambda_{XA} = \lambda_b \cdot \alpha_A \Pi k_i \quad (1)$$

where λ_{XA} is the estimated failure rate for equipment X in failure mode A, λ_b is the base generic failure rate, α_A is the proportion of all failures in failure mode A, and Πk_i is the product of the different stress factors.

A subjective methodology for estimating stress factors based on the differences between the "average industrial conditions" associated with the reference-point failure rate (λ_b) and the system being assessed is the DOE (Design, Operation, Environment) model. Scores for design, operation, and environment stress attributes are allocated on the scale -1 to $+1$.

Stress	Low	Average	High
Score (z)	-1	0	$+1$

Fractions within and outside the range are allowed, and k factors are calculated from the expression

$$k = e^z \quad (2)$$

where z is the estimated score for the design, operation, and environment attributes of the system. Assuming a circuit breaker is installed in the electrical distribution system of a petrochemical plant located on a UK estuary, the k factors would perhaps be

Design z score say, $+0.25$ – (air-operated, high voltage circuit breaker)

Operation z score say, 0 – (normal operation, regularly tested and maintained)

Table 1 Principal Reliability Data Sources

Data source	Title	Publisher and date
1. CCPS	<i>Guidelines for process equipment reliability</i>	American Institute of Chemical Engineers, 1990
2. Davenport and Warwick	A further survey of pressure vessels in the UK 1983–1988	AEA Technology – Safety and Reliability Directorate, 1991
3. DEFSTAN 0041, Part 3	<i>MOD practices and procedures for reliability and maintainability, Part 3, Reliability prediction</i>	Ministry of Defence, 1983
4. R. F. de la Mare	Pipeline reliability; report 80-0572	Det Norske Veritas/Bradford University, 1980
5. Dexter and Perkins	Component failure rate data with potential applicability to a nuclear fuel reprocessing plant, report DP-1633	EI Du Pont de Nemours and Company, USA, 1982
6. EIREDA	<i>European industry reliability data handbook</i> , Vol. 1, <i>Electrical power plants</i>	EUROSTAT, Paris, 1991
7. ENI Data Book	<i>ENI reliability data bank – component reliability handbook</i>	Ente Nazionale Indrocarburi (ENI), Milan, 1982
8. Green and Bourne	<i>Reliability technology</i>	Wiley Interscience, London, 1972
9. IAEA TECDOC 478	<i>Component reliability data for use in probabilistic safety assessment</i>	International Atomic Energy Agency, Vienna, 1988
10. IEEE Std 500-1984	<i>IEEE guide to the collection and presentation of electrical, electronic sensing component and mechanical equipment reliability data for nuclear power generating stations</i>	Institution of Electrical and Electronic Engineers, New York, 1983
11. F. P. Lees	<i>Loss prevention in the process industries</i>	Butterworth, London 1980
12. MIL-HDBK 217E	<i>Military handbook – reliability prediction of electronic equipment</i> , Issue E	US Department of Defense, 1986
13. NPRD-3	<i>Non-electronic component reliability data handbook – NPRD-3</i>	Reliability Analysis Centre, RADC, New York, 1985
14. OREDA 84	<i>Offshore reliability data (OREDA) handbook</i>	OREDA, Hovik, Norway, 1984
15. OREDA 92	<i>Offshore reliability data</i> , 2nd edition	DnV Technica, Norway, 1992
16. RKS/SKI 85-25	<i>Reliability data book for components in Swedish nuclear power plants</i>	RKS – Nuclear Safety Board of the Swedish Utilities and SKI – Swedish Nuclear Power Inspectorate, 1987
17. H. A. Rothbart	<i>Mechanical design and systems handbook</i>	McGraw-Hill, 1964
18. D. J. Smith	<i>Reliability and maintainability in perspective</i>	Macmillan, London, 1985
19. Smith and Warwick	A survey of defects in pressure vessels in the UK (1962–1978) and its relevance to primary circuits, report SRD R203	AEA Technology – Safety and Reliability Directorate, 1981
20. WASH 1400	<i>Reactor safety study. An assessment of accident risks in US commercial nuclear power plants</i> , Appendix III, Failure data	US Atomic Energy Commission, 1974

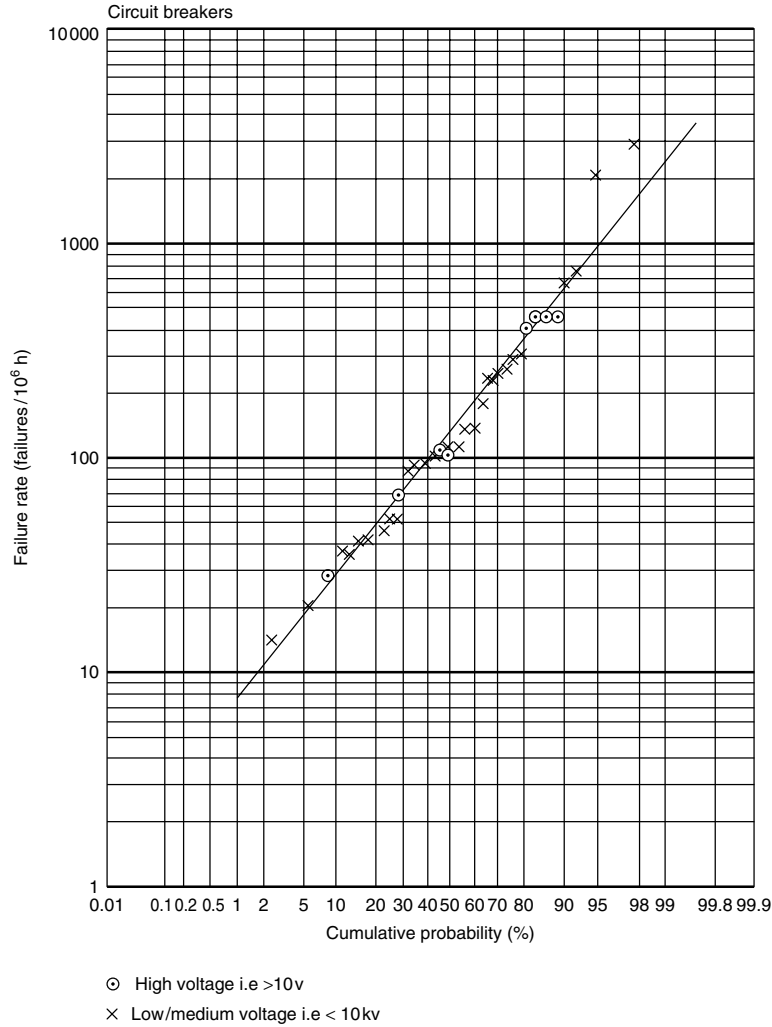


Figure 1 Lognormal distribution of circuit breaker generic failure rates

Environment z score $+0.5$ – (exposed marine location).

On the basis of these z scores, estimated design, operation, and environment k factors are as follows:

$$k_1 = e^{0.25} = 1.3$$

$$k_2 = e^0 = 1.0$$

$$k_3 = e^{0.5} = 1.6$$

Therefore, Πk_i is approximately 2. If it is assumed that the “fail-to-danger” mode is circuit breaker fail-closed with say 80% of failures in this mode, then a

best-estimate “dangerous” failure rate is

$$\begin{aligned} \lambda_d &= \lambda_b \cdot \alpha_D \Pi k_i \\ &= 4 \times 0.8 \times 2 = 6.4 \text{ failures}/10^6 \text{ h} \end{aligned} \quad (3)$$

For a system failure mode of “fail to operate on demand” and the assumption of a test interval (T) of 3 months (say 2000 h), the fractional dead time (mean unavailability) due to circuit breaker failure is

$$FDT = \frac{\lambda_d T}{2} = \frac{6.4 \times 2000}{2 \times 10^6} = 6.4 \times 10^{-3} \quad (4)$$

This is the probability of failure of a circuit breaker to disconnect the electricity supply on demand for input to a system fault tree. A similar approach would be adopted for the other pieces of equipment in the system with more detailed investigation focusing on failure data for the critical failure modes.

In-service Reliability Data

In-service reliability data analysis is required for monitoring equipment reliability performance during operation and to derive estimates of equipment failure rates for the in-house generic database. During analysis, it is important to differentiate between component and equipment failures. Components are items that can only fail once. Equipment, on the other hand, are repairable systems, and “failure” does not necessarily cause end of life; the piece of equipment can generally be repaired or restored to its previous operational state by replacement of the failed component. Different methods of data analysis and different definitions of the rate function are required for component and equipment reliability data analysis. For components, the rate function is termed the *hazard rate function* $h(t)$ (see **Mathematical Models of Credit Risk; Lifetime Models and Risk Assessment; Markov Modeling in Reliability**) and for equipment the *failure rate function* $z(t)$.

Component Data Analysis

Components, by definition, are not repairable; they have a finite life and lifetime characteristics can generally be represented by one of the standard distributions. The important point to note is that it is the characteristic of a population of identical components that fail randomly at different ages. Some sample populations exhibit decreasing hazard rate – generally because of quality control problems, others constant hazard rate – due to random shocks such as power surges, or increasing hazard rate – due to aging. When a component fails in a repairable system, it will be replaced by an identical new component with the assumption that the repairable system is restored to “as-good-as-new” (AGAN) condition. This assumption of “AGAN” is very important in component reliability analysis and includes the assumption that all observed events are from

components of the same type, and therefore that component lifetimes are independent and identically distributed (i.i.d.).

For detailed analysis of component failure characteristics, reliability engineers generally turn to Weibull analysis because of the distribution’s flexibility – it can model decreasing, constant, and increasing hazard rates. Graphical Weibull analysis (see **Reliability Growth Testing**) is preferred, particularly for small sample populations because of the snapshot it gives of scatter and the general structure of the data. Special probability plotting paper has been developed for the Weibull distribution. There are also a number of software packages available for determining the distribution of component failures.

Equipment Data Analysis

The analysis of reliability data from repairable systems can never be as precise as for items that are not repairable. Components are frequently bench tested under conditions where uncertainties concerning operating environment, component quality, age, etc., are closely controlled. Even when component data are extracted from operating system records, it can be assumed that failed components are replaced by new ones with the equipment considered only as the test bed.

When analyzing failure data from equipment in operating systems, however, the position is not so clear cut. Equipment, by definition, are repairable, and it is seldom possible to determine the age and condition of a piece of equipment at the start of surveillance. (Equipment that are not repairable are sometimes referred to as *supercomponents*.) Because of the diversity of components that make up equipment, the assumption of i.i.d. times to failure is no longer valid. Certain basic assumptions therefore need to be made, some of which are as follows:

- repairs are perfect, restoring the equipment to “as new” condition;
- failures occur at random, but at an underlying constant failure rate, λ . This implies that failures are the result of a homogeneous Poisson process (HPP) giving interfailure times t that are exponentially distributed, and thus failures in a specified time interval have a Poisson distribution.

Under the HPP model, the expected number of failures over any time period is λt and the probability

of observing an arbitrary number of failures x is given by the Poisson distribution where

$$P(X = x) = \frac{e^{-m} m^x}{x!} \quad (5)$$

and $m = \lambda t$. When the failure rate is not constant, the model is a nonhomogeneous Poisson process (NHPP).

The concept of increasing or decreasing failure rates for repairable equipment is completely separate from the issue of increasing or reducing hazard rates associated with components. Equipment can be simple or complex, but in either case, comprise a family of component parts. These components have different lifetime characteristics and unless there is a dominant failure mode, the mixing of component lifetimes and replacements when a component fails generally ensures that equipment failure rates quickly settle down to a constant value. However, and particularly with mechanical equipment, some components “bed-in” resulting in a gradually reducing failure rate as equipment age. On the other hand, other failure modes such as component wear may eventually lead to a gradually increasing failure rate. Monitoring changes in the reliability performance of systems and equipment to detect the onset of increasing or decreasing failure rates is a fundamental requirement for the availability-critical and safety-critical systems of an industrial plant.

Equipment Failure Characteristics

Component lifetime analysis is concerned with aging characteristics (*see Structural Reliability*). A component operates for a period of time and, for a population of components, these characteristics can be represented by the Weibull distribution in terms of the characteristic life (η) and shape factor (β) from which other parameters such as hazard rate $h(t)$ and mean time to failure (MTTF) can be determined. With this information, predictions can be made as to their likely reliability performance when installed in equipment.

The concept of “lifetime” is different for equipment, which operate within significantly longer time frames than their component parts. There will come a time for all industrial equipment when replacement is obviously essential because of the combined effects of the industrial and operating environments. However, for most of the time, a useful life period can be defined where the assumption of AGAN after component failure and replacement is valid. With this

assumption, the reliability characteristic of equipment can be categorized by a (constant) failure rate or its reciprocal – the equipment mean time between failures (MTBF) (*see Accelerated Life Testing*).

It is the author’s personal view that time between failures should always relate to operating time rather than operating plus repair time since repair strategies can vary significantly between different installations even within the same organization. Repair time data is, of course, important since it impacts on equipment, system, and plant availability, but its analysis needs to be considered separately from the analysis of failure time data. For failure data that meet these criteria, the MTBF is the total operating time of a population of equipment divided by the total number of failures.

Example

Three failures were recorded from a population of 10 pumps in a period of 1 year. Assuming an average repair time per failure of 8 h and AGAN after repair, then

$$\begin{aligned} MTBF &= \frac{(8760 \text{ h} \times 10 \text{ pumps}) - (3 \times 8 \text{ h})}{3 \text{ failures}} \\ &= 29192 \text{ h} \end{aligned} \quad (6)$$

Failure rate is the reciprocal of the MTBF so $\lambda = 1/(29192) = 3.4 \times 10^{-5} \text{ h}^{-1} = 0.3 \text{ failures/year}$.

Recalling the Poisson distribution as in equation (5) or in its more familiar form, the reliability $R(t) = e^{-\lambda t}$, when $x = 0$ since $(0.3)^0$ and $0!$ both equal unity.

Hence, the reliability (probability of no failure) of a pump in a period of 1 year is

$$R(t) = P(X = 0) = e^{-\lambda t} = e^{-0.3} = 0.74 \quad (7)$$

The probability of no failures in the population of 10 pumps during the year is

$$R(t) = e^{-(0.3 \times 10)} = 0.05 \quad (8)$$

And, taking this a step further, one can estimate the probabilities for $X = 1$, $X = 2$, etc.

For example,

$$P(X = 2) = \frac{e^{-m} m^x}{x!} = \frac{e^{-3.0} .3^2}{2 \times 1} = \frac{0.05 \times 9}{2} = 0.45 \quad (9)$$

This is the probability of exactly two failures in 1 year of operation.

Obviously the probability of more than one failure = $1 - R(t) = 1 - 0.05 = 0.95$

Reliability Data Collection

For the occasional *ad hoc* investigation triggered, say, by concern that some reliability problem has emerged concerning a piece of equipment, or a population of nominally identical items, investigations are generally short and clearly focused. In these situations, manual extraction of failure information from maintenance records and input to one of the standard spreadsheet programs is usually adequate. This methodology can also be acceptable for bench testing of components and equipment. However, when there is a routine reliability program with ongoing monitoring of critical systems and equipment, the use of database management programs such as ACCESS can be worthwhile. Standard database programs facilitate the organization of the basic inventory and event data into formats suitable for further processing by the use of spreadsheet programs or specialized management information and reliability analysis software. A typical information flow diagram for a reliability data collection and analysis system is shown in Figure 2.

Most of the information obtained from plant records are processed and stored in inventory and

event data files in standard formats. Inventory data comprises the engineering and process information required to identify the major items of equipment with respect to their design, operation, and maintenance features. Their location in the plant will be defined by a hierarchical code at plant, system, and equipment level.

The three main types of data needed in a reliability data system are as follows:

1. inventory data – defines the design and functional characteristics of the item;
2. failure event data – textual descriptions of each failure event (with additional codes to relate to the item's inventory) failure codes and operational and environmental conditions;
3. operating state data – dates and times of failure, operating state prior to failure, active repair and waiting times, etc.

Operating state data is seldom recorded, except for major equipment such as rotating machinery, hence, in most cases, estimates of operating time and time on standby are included with the failure event data.

It is worth making the point that many reliability data projects fail because of lack of attention to data quality. Before starting data collection, it is necessary to review the source of the data so that realistic objectives can be established within the available budget. A quality assurance plan should then be developed, which defines the objectives of the exercise and details the procedures to be followed in data collection and analysis. Finally, a quality audit should be carried out on completion, to assess the degree of success and to feed back processed data and recommendations to the host management. Feedback of information is essential for data validation and to ensure cooperation in future studies. In large schemes, it is worthwhile setting up a steering committee that meets at regular intervals to monitor progress on the study.

References

- [1] Andrews, J.D. & Moss, T.R. (2002). *Reliability and Risk Assessment*, 2nd Edition, John Wiley & Sons.
- [2] Moss, T.R. (2005). *The Reliability Data Handbook*, John Wiley & Sons.
- [3] Newton, D.W. (1994). *Part of Reliability of Mechanical Systems*, 2nd Edition, PEP.

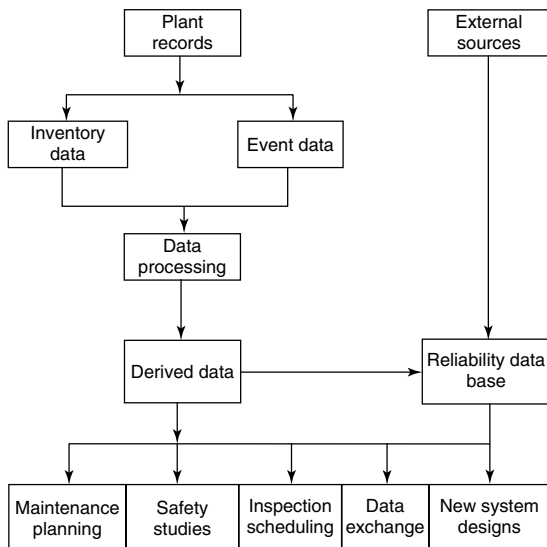


Figure 2 Typical information flow diagram

- [4] Ansell, J.I. & Phillips, M.J. (1994). *Practical Methods for Reliability Data Analysis*, Oxford Science Publications.
- [5] Moss, T.R. & Strutt, J.E. (1993). Data sources for reliability design analysis, *IMechE Proceedings*, ImechE Technical Note E00293, Institute of Mechanical Engineers.
- [6] Bloch, H.P. & Geitner, F.K. (1999). Use equipment failure statistics properly, *Hydrocarbon Processing*, Rome Air Development Center.

Further Reading

Kirsteiger, C., Teichmann, T. & Richwine, R.R. (1987). *Availability Predictions for Combustion Turbines, Combined Cycle and Gasification Combined Cycle Power Plants*, ASME paper 87-JPGR-Pwr-64.

Proccacia, H., Arsenis, S. & Aufort, P. (1998). *European Industries Reliability Data Bank: EIReDA*, 3rd Edition, Crete University Press, Heraklion.

US Department of Defense (1980). *Maintainability Analysis*, MIL-HDBK 472.

Some Published and Estimated Failure Rates for Use in Fault Tree Analysis, Du Pont, 1981.

Moss, T.R. (1994). *T-Book: Reliability Data of Components in Nordic Nuclear Power Plants*, 4th Edition, ATV, Stockholm.

Consulting Engier (Reliability) (1999). *Handbook of Quality of Reliability Data: ESReDA*, Det Norske Veritas.

THOMAS R. MOSS

Reliability Demonstration

The purpose of a reliability demonstration test plan is to formally verify that a product meets a specified reliability requirement with a certain level of confidence. These plans are used in industry to provide assurance to consumers and external regulating bodies, as well as for internal quality control purposes. Consumers include both individual customers, as well as large companies that buy components from suppliers.

This article focuses on demonstration plans using binary (success/failure) data, also called *attribute testing* (see **Axiomatic Models of Perceived Risk; Repeated Measures Analyses**). While this is the most common scenario, there do exist other plans that use auxiliary information in addition to the success/failure data. One example is information on time at which the failures occurred (often called *time-to-failure data*). We discuss this case briefly later in the article.

The section titled “Demonstration Test Plans” provides an overview of zero-failure test plans and general k -out-of- n plans. The power behavior (probability of successful demonstration) is discussed in the section titled “Power of Demonstration Plans”. Two general approaches for increasing the efficiency of demonstration testing, Bayesian techniques and extreme testing, are reviewed in the section titled “Improving the Efficiency of Test Plans”. Additional information on demonstration testing and references to the literature can be found in [1, Chapter 10, pp. 466–486] and [2, Chapter 10.6, pp. 247–249]. Furthermore, several popular statistical software packages have modules for constructing demonstration test plans.

Demonstration Test Plans

Reliability demonstration plans (see **Comparative Risk Assessment**) have two specific requirements: a reliability target that has to be verified and a confidence level that has to be achieved. Reliability targets can be specified in terms of the reliability of a product at time t_0 , $R(t_0) = R_0$, or the design life corresponding to a reliability of R_0 , $L(R_0) = L_0$. Recall that $L(R)$ is the $(1 - R)$ th quantile. There is a one-to-one relationship between design life (see

Reliability Integrated Engineering Using Physics of Failure) and reliability, so we restrict attention to the reliability target R_0 throughout.

We begin with the zero-failure test plan, which is commonly used. Under this plan, a simple random sample of n product units are put on test until time t_0 . The test is said to successfully demonstrate the target reliability of R_0 if there are no failures during the test; otherwise, it is not successful. The sample size n is chosen to satisfy the confidence level requirement. Specifically, we must have

$$R_0^n \leq \alpha \quad \text{or} \quad n \geq \frac{\log(\alpha)}{\log(R_0)} \quad (1)$$

Letting $\lceil x \rceil$ be the smallest integer greater than or equal to x , we get the required sample size as

$$n = \left\lceil \frac{\log(\alpha)}{\log(R_0)} \right\rceil \quad (2)$$

The expression R_0^n in equation (1) corresponds to the probability of observing no failures out of n units when the true reliability is R_0 . This is sometimes called the “success-run” formula in the engineering literature [3].

The sample size is chosen so that the probability of observing zero failures is no greater than α when the true reliability is R_0 or smaller. Thus, if we do not observe any failures during the demonstration test, we can conclude with confidence (of at least $(1 - \alpha)$) that the required reliability must be at least as high as R_0 .

The actual probability of a successful demonstration depends on the true but unknown reliability R . We refer to this as power and discuss it in more detail in the next section. In practice, the value of R needs to be much higher than the target value R_0 in order to have a reasonable probability of successful demonstration.

The zero-failure plan is a special case of a k -out-of- n demonstration test plan. In this general case, one tests a simple random sample of n product units. The test is deemed successful if the number of failures is less than or equal to k . If we observe more than k failures, the test is unsuccessful. Here, the number of allowable failures k is fixed *a priori*. The sample size of the test plan is then chosen so that

$$\sum_{i=0}^k \binom{n}{i} (1 - R_0)^i R_0^{n-i} \leq \alpha \quad (3)$$

The expression on the left in equation (3) is the probability of getting k or fewer failures from testing n units with failure probability $(1 - R_0)$. This is simply the cumulative binomial probability.

Example 1 Suppose we need to demonstrate reliability of $R_0 = 0.9$ with 95% confidence. From equation (2), a zero-failure plan would require a sample of $n = 29$ units. On the other hand, a 1-out-of- n failure plan would require a sample of $n = 46$ units, since $n = 46$ is the smallest integer value for which equation (3) holds with $k = 1$.

A number of factors are involved in choosing the number of allowable failures k , some of which will be discussed later in the article. One concern is that the sample size n is not too large as to be unaffordable. As seen in Example 1, the sample size n increases with k . For this reason, the zero-failure plan ($k = 0$) is a popular choice.

Nevertheless, plans with $k > 0$ are also often used. These are sometimes preferred over zero-failure plans because they have an increased power or likelihood of being successful compared to zero-failure plans. The trade-offs between sample size and power are discussed in the next section.

Reliability demonstration plans are mathematically equivalent to certain hypothesis tests in statistics. Specifically, a reliability demonstration plan with $(1 - \alpha)$ confidence is equivalent to an α -level hypothesis test of $H_0 : R \leq R_0$ versus $H_1 : R > R_0$. A successful reliability demonstration plan corresponds to rejecting the null hypothesis. Further discussion of this equivalence can be found in [4, pp. 404–405].

Power of Demonstration Plans

In hypothesis testing, recall that the “power” of a test is defined as the probability of rejecting H_0 . We can analogously define the power for a k -out-of- n reliability demonstration plan as the probability that the demonstration test is successful, i.e., the probability of observing k or fewer failures.

The actual value of the power depends on the true reliability R and is given by

$$P(R) = \sum_{i=0}^k \binom{n}{i} (1 - R)^i R^{n-i} \quad (4)$$

Since R is unknown in practice, the power of the demonstration plan is also unknown. However, one can compute the power for various values of R and examine the power curve before conducting the test. Demonstration tests can be costly, and an unsuccessful test may result in loss of customer’s confidence, so one has to be reasonably certain up front that the test will be successful. The power $P(R)$ of a demonstration plan also depends on the choice of k (and hence also n) in a k -out-of- n plan. As mentioned, the power increases with k , but this, in turn, requires a larger sample size n .

Example 2 Consider again the set up in Example 1 with $R_0 = 0.9$ and confidence level of 95%. Suppose that the true reliability for this product is $R = 0.99$. The zero-failure plan with sample size of $n = 29$ has a power of 0.75. On the other hand, the $k = 1$ -failure plan requires a larger sample size of $n = 46$, but also has a larger power of 0.92. These comparisons can be seen in the first plot in Figure 1.

In practice, test planners must balance the need for a large power with the cost of a large sample size. Typically, they calculate the power in equation (4) for different values of the true reliability R and different plans before deciding on the best plan to use. In many cases, a specific minimum power requirement is used to determine the values of k and n . For instance, in Example 2, if the test planner required that the power be at least 0.95 when the true unknown reliability is 0.99, then neither the zero-failure plan nor the $k = 1$ -failure plan would be adequate. Rather, the plan with the smallest sample size n that meets this requirement is the $k = 2$ plan, which requires a sample size of $n = 61$ (as displayed in the first plot in Figure 1). If this sample size is not affordable, the test planner would either need to lower the power requirement, decrease the required confidence $(1 - \alpha)$, or use other testing schemes. (See, for example, the discussion on Bayesian techniques and extreme testing later in this article.)

Note that while the terms *power*, *level*, and *confidence* are common in statistics, engineers typically use the terms *manufacturer’s risk* or *producer’s risk*, and *consumer’s risk* (see, for example [5, pp. 244–245]). Producer’s risk is the probability or risk of rejecting the product (failing the demonstration test) when the product reliability meets or exceeds the target. Consumer’s risk, on the other hand, is the probability or

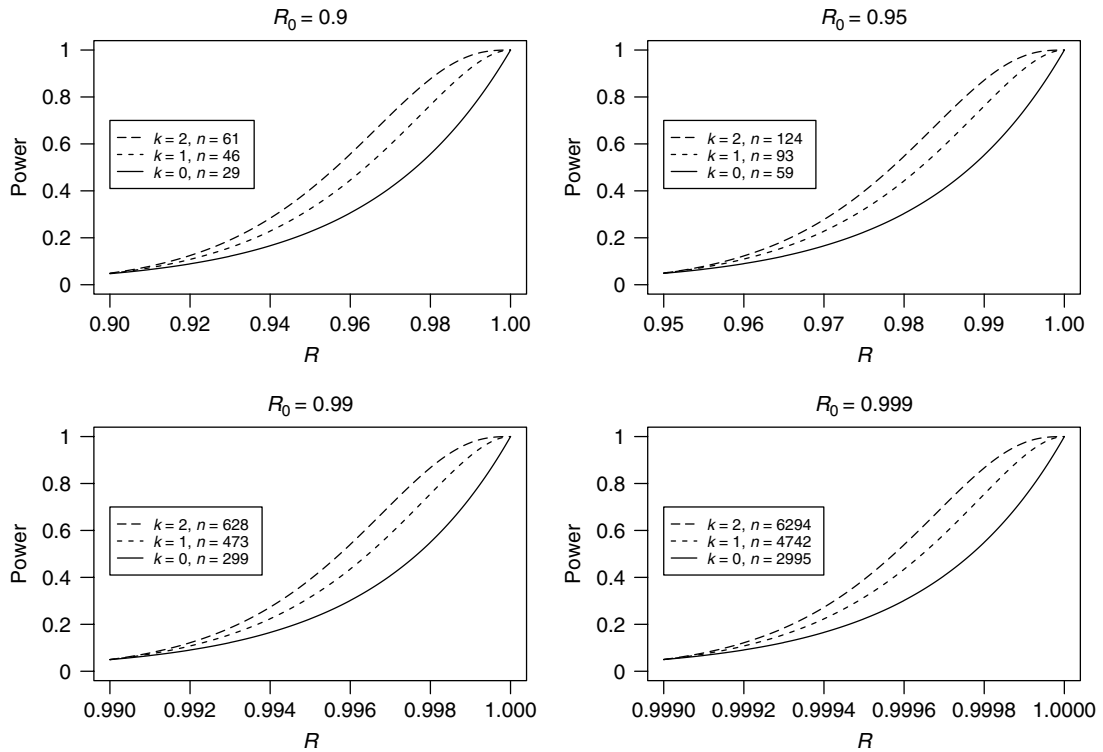


Figure 1 Power of demonstration plans for different values of k and R_0

risk of accepting the product when the true reliability does not meet the target value.

Figure 1 provides a comparison of the power of $k = 0$, $k = 1$, and $k = 2$ test plans for different values of R_0 as a function of the true unknown reliability R with level $\alpha = 0.05$. The four plots display the power of these test plans for R_0 equal to 0.9, 0.95, 0.99, and 0.999, respectively. In each plot, the x axis begins at R_0 and ends at 1. Each curve begins at $\alpha = 0.05$ on the y axis since the power, by definition, is equal to α when R is equal to R_0 . Similarly, each curve ends at 1 on the y axis since the power of any demonstration plan is necessarily equal to 1 when R is equal to 1. The curves for each value of k are remarkably similar across the four graphs despite the different sample sizes (and different scales for the x axis). As noted earlier, larger values of k lead to larger power, but also require larger sample sizes. Some software packages, such as Minitab (Version 14.0) can be used to compute sample size and power calculations for reliability demonstration plans.

Improving the Efficiency of Test Plans

In many cases, the number of units required for the demonstration plans can be prohibitively large. In this section, we discuss two alternative approaches that use additional information to reduce sample sizes and test costs. See also [1, Chapter 10] for sequential tests.

Extreme Testing

Increasing test time or overtesting (also called *bogey testing*) is a common approach for reducing sample sizes (see, for example [2, Chapter 10.6] and [3]). Suppose the goal is to demonstrate reliability of $R(t_0) = R_0$ at time t_0 . Instead of testing units only until time t_0 , one can test them for a longer period of time, say $t_1 > t_0$. If we have a parametric model that relates the reliability at time t_0 to reliability at time t_1 , we can trade-off sample size for increased test time (see **Accelerated Life Testing**). This approach is beneficial if test time is relatively cheaper than test units.

Example 3 Consider again Example 1 with $R_0 = 0.9$ and $\alpha = 0.05$, where the traditional zero-failure plan requires $n = 29$ units. Suppose R_0 corresponds to reliability at time $t_0 = 1$ million cycles of operation. In this situation, suppose we can test the units until time t_1 with $t_1 > t_0 = 1$ million cycles. For the sample-size calculation, we need to know the form of the underlying reliability or failure-time distribution. If it is exponential, the sample-size formula is

$$n = \left\lceil \frac{s \log(\alpha)}{\log(R_0)} \right\rceil \quad (5)$$

where $s = t_0/t_1$ [2, p. 248]. This yields $n = 15$ for $t_1 = 2$ million cycles and $n = 10$ for $t_1 = 3$ million cycles. For the Weibull distribution with known shape parameter β , the sample-size formula becomes [2, p. 248]

$$n = \left\lceil \frac{s^\beta \log(\alpha)}{\log(R_0)} \right\rceil \quad (6)$$

where again $s = t_0/t_1$.

It is natural to wonder why one should not continue increasing the test time to reduce the sample size all the way to $n = 1$. One reason is that the assumption about the failure-time distribution is less likely to hold in the extreme tail, so one must be cautious in increasing the test time much beyond the target condition. Additionally, testing for a very long period can introduce failure modes that are not present at the target condition.

There are also expressions for sample sizes for general k -out-of- n plans under overtesting, but they are more involved. Some software packages, such as Minitab (Version 14.0), have capabilities for computing the sample sizes for selected log-location-scale distributions, assuming that the scale parameter is known.

The authors of [6] generalize approaches such as overtesting, which seek to improve demonstration test efficiency through increasing failure probability. They use the term *extreme testing* or (“X-testing” for short) to describe such methods. Another well-known example of X-testing is the use of accelerated stress conditions to induce early failures. While this is common with time-to-failure data, it can also be useful with binary data. However, one needs information about the acceleration model (see **Competing Risks**), which requires testing at more than one stress level. The authors of [6] discuss this and other,

newer approaches to X-testing. These include testing “weaker units”, the use of series-system structures for testing components, etc. The fundamental idea in X-testing is that it is often more efficient to demonstrate reliability under “extreme” conditions that correspond to a lower reliability target. The X-transform $\tau(\cdot)$ maps R , the reliability under standard conditions, to $\tau(R)$, the reliability under the extreme conditions. For suitable X-transforms, this leads to demonstrating a lower reliability target $\tau(R_0)$. X-testing assumes that the X-transform $\tau(\cdot)$ is known. A number of situations in which this is feasible are discussed in [6]. For example, in the overtesting case with the exponential distribution we have

$$\tau(R) = e^{s \log(R)} \quad \text{or} \quad \log(\tau(R)) = s \log(R) \quad (7)$$

where $s = t_0/t_1$. This requires information about the failure-time distribution which can be obtained from previous studies.

An important consideration in using X-testing is the effect on the power of the demonstration plan. In the case of zero-failure plans, the power of an X-test can be equal to, smaller than, or even larger than the power of the corresponding traditional plan (despite having a smaller sample size). The power of the X-test is completely determined by its X-transform, which, in turn, depends on the methodology used to induce failures. The authors of [6] discuss different methods of inducing failure and show that many of them lead to location or scale X-transforms.

Bayesian Techniques

The Bayesian approach provides a formal framework for introducing prior knowledge about the product reliability R and using it to reduce test costs or sample sizes. This knowledge is represented through a prior probability distribution on R , and may come from past testing or even expert opinion.

To describe the formulation, suppose the prior distribution of R is uniform on $(0, 1)$. (This is a conservative case; in practice, the prior information will be more informative.) Conditional on the value of R , the number of failures among n units tested is (as before) binomial with probability $(1 - R)$. Let us consider first the simple case of zero-failure plans. If we observed zero failures, the posterior probability density of R , given this event, turns out to be $(n + 1)R^n$. Thus, the posterior probability that R is less than the target value R_0 is $\int_0^{R_0} (n + 1)R^n dR = R_0^{(n+1)}$. Under

the Bayesian formulation, we choose the sample size n so that this posterior probability is less than or equal to α . In this particular case, this leads to $R_0^{(n+1)} \leq \alpha$, so

$$n = \left\lceil \frac{\log(\alpha)}{\log(R_0)} \right\rceil - 1 \quad (8)$$

which is one less than the classical expression in equation (2). This reduction in sample size may seem counterintuitive since the uniform prior could be viewed as conservative. Priors with more information (with more probability mass at higher reliabilities) will result in substantially smaller sample sizes.

The Bayesian versions of k -out-of- n plans are obtained similarly. As before, k is fixed *a priori*, and the sample size n is then determined. Given a prior distribution, we can compute the posterior probability density of R , conditional on observing k or fewer failures. Let $p(R; n, k)$ denote this posterior probability density function of R . Then, we choose n so that

$$\int_0^{R_0} p(R; n, k) dR \leq \alpha \quad (9)$$

For more details, see [1, Chapter 10, pp. 466–486].

Priors in the beta family of probability distributions are especially convenient since the resulting posteriors will also be in the β family. However, recent advances in Bayesian computing and the use of (*see Nonlife Loss Reserving; Bayesian Statistics in Quantitative Risk Assessment*) methods Markov Chain Monte Carlo (MCMC) allow us to also handle more general prior distributions, so the appropriateness of the prior for the problem at hand should be the main concern.

Detractors of the Bayesian approach may characterize the selection of a prior distribution as “making up data,” while proponents of the approach would point out that not all information comes in the form of data and that many users of statistical methods interpret results in a Bayesian manner anyway. In any case, it is certainly true that prior information can greatly reduce the required sample size for a reliability demonstration test.

Concluding Remarks

The current emphasis in industry is on building quality and reliability into the product up front. On the

other hand, reliability demonstration planning shares some of the philosophy and terminology with sampling inspection for quality assurance, which has fallen out of favor. Nevertheless, reliability demonstration remains an important part of current industrial practice for verifying customer requirements or government regulations. Reliability demonstration test plans are also used internally in many companies, although these demonstration tests may be used informally, without requirements on confidence levels.

Because demonstration test plans can be costly in terms of test units or test time, there is a need to identify and develop alternative methods that can reduce test costs. We have discussed two classes of approaches here, but both use only the binary (success/failure) data. Even with these k -out-of- n plans, if $k > 0$, we may have information on the actual time of failure, which can be exploited. For example, if we observe one failure, then knowing whether the failure occurred early in the test or just before the end of the test could provide useful, additional information.

Other methods of increasing efficiency include using the actual failure-time data to estimate the reliability and then computing appropriate confidence bounds to demonstrate that the target is met. In applications with very high reliability, one may need to use accelerated testing to induce failures and observe the failure times. An alternative approach that is becoming more common is the collection and analysis of degradation data. All of these methods, like the X-testing approach discussed in this article, require additional assumptions such as knowledge of the failure-time distribution, the acceleration model, the X-transform, and so on. Many customers and regulatory agencies are often reluctant to accept these assumptions, leading to the continued reliance on traditional demonstration tests.

Acknowledgment

Nair's research was supported in part by NSF grant DMS 0505535.

References

- [1] Martz, H.F. & Waller, R.A. (1982). *Bayesian Reliability Analysis*, John Wiley & sons, New York, Chapter 10, pp. 466–486.

- [2] Meeker, W.Q. & Escobar, L.A. (1998). *Statistical Methods for Reliability Data*, John Wiley & sons, New York, Chapter 10.6, pp. 247–249.
- [3] Lipson, C. & Sheth, N.J. (1973). *Statistical Design and Analysis of Engineering Experiments*, McGraw-Hill, New York.
- [4] Mann, N.R., Schafer, R.E. & Singpurwalla, N.D. (1974). *Methods for Statistical Analysis of Reliability and Life Data*, John Wiley & Sons, New York, Chapter 85, pp. 404–409.
- [5] Small, B.B. (1985). *AT&T Statistical Quality Control Handbook*, 11th printing, Delmar Printing Company, pp. 244–245.
- [6] Mease, D. & Nair, V.N. (2006). Extreme (X-)testing with binary data and applications to reliability demonstration plans, *Technometrics* **48**, 399–410.

DAVID MEASE, VIJAYAN N. NAIR AND
JEFFREY A. ROBINSON

Reliability Growth Testing

The first prototypes of a new product generally contain design, manufacturing, and/or engineering flaws that prevent specified reliability requirements to be met. Reliability growth is the process by which the reliability problems are identified and corrected by introducing modifications in the product design and/or manufacturing process. Accordingly, the reliability of the product grows as the product goes through successive stages of the development program.

Permanent actions addressed to prevent reliability problems both of tested prototypes and of all future copies of the product constitute a reliability growth. Reworks, screening of units with lower reliability, and temporary fixes do not constitute reliability growth.

Reliability growth programs should be carried out for all new products and all existing products subject to major design changes, to reduce the risk in launching new technologies, and to avoid mass production of products that do not meet reliability requirement and customer satisfaction. Clearly, early implementation of a reliability growth program can minimize the impact of design and/or process changes on production scheduling and total cost of the product, because redesigning the product late in the development phase or retrofitting products already in the field is very expensive.

Implementing correctly reliability growth programs can produce increased quality and reliability of product and process, reduced costs and risk of product development, improved customer satisfaction, improved market position, and reduced warranty claims and costs. A collateral result is the increased knowledge and experience gained by the reliability team on the identification and solution of reliability problems.

The degree of possible improvement in product reliability depends on the underlying technology of the product, the available resources, and the knowledge of the reliability team. The underlying technology is an obvious limiting factor, as well as economic and human resources that are always limited, thus conditioning the management strategies toward performing tests and introducing reliability improvements. For example, limited resources may prevent the team from introducing very expensive

modifications. The knowledge of the reliability team affects both the ability to identify the cause of observed failures and the degree of effectiveness of implemented actions, because a corrective action does not always entirely eliminate a failure mode from occurring again, even if the cause has been identified.

The limitation of resources and development time compels the reliability growth process to be as efficient as possible. Clearly, efficiency is a relative concept and the attained improvement has to be compared with the reliability level at the beginning of testing and to the reliability goal.

In general, a reliability growth program may have more than one reliability goal. For example, one goal may be associated with all the failures occurring during the warranty period and the second goal would be to achieve a very low probability of occurrence of safety related failures.

Of course, the reliability growth has to be monitored to compare the actual reliability level to the target value and to estimate whether the development program is progressing as scheduled. If the achieved improvement is less than the planned one and the reliability goal is not going to be met, then new strategies should be considered, by devoting a larger amount of resources to the program.

Although reliability growth generally refers to improvements in products reliability, a very important field where reliability growth programs find large application is the control of industrial accidents (*see Human Reliability Assessment; Experience Feedback*) [1].

Reliability growth programs generally consist of a number of test stages. Each stage can be viewed as a distinct period of time when the product is subjected to testing and subsequent corrective actions (or fixes). During each stage, one of three basic types of test and fix programs may be carried out, the main distinction among them being when fixes are introduced into the products. Fixes can be implemented each time a failure occurs and the failure cause is identified, or can be delayed and implemented in block at one time, typically at the end of each stage (*delayed fixes*). The three types of testing program are as follows: (a) test-analyze-and-fix (TAAF) (*see Hazard and Hazard Ratio*) program, when fixes are implemented just after the failure occurrence, (b) Test-Find-Test (TFT), when all fixes are delayed, and (c) TAAF program, with some delayed fixes. The TAAF program is the one

with a more or less smooth, continuous growth curve, because each fix introduced during testing generally produces a small increase in product reliability. On the contrary, the TFT produces significant jumps in the product reliability, resulting in a step function [2].

In general, reliability growth programs are carried out on a number of identical prototypes, manufactured at the same time. However, in some cases, especially when the products are very expensive and their manufacturing time is very long (i.e., locomotives, airplanes, industrial plants, etc.), the development program involves preseries products that are put on test progressively, as soon as they are manufactured. In this case, design and/or process defects, discovered during the development program, are corrected and product design and/or manufacturing process improve as the program goes on. Thus, the initial reliability of tested products is not the same: a product put on test at the beginning of the program has more defects than one product put on test when the program is going to end (see, for example [3]).

The reliability growth of a product can be expressed in terms of quantities such as the decrease in the rate of occurrence of failures $\rho(t)$ (see **Repairable Systems Reliability; Near-Miss Management: A Participative Approach to Improving System Reliability**), the decrease in the number of failures, and the increase in the mission success probability. These quantities are generally mathematically related and can be derived from each other. The changes in these quantities, as a function of the testing time, are generally referred to as *reliability growth trends* and are described through statistical models called *reliability growth models* (see **Repairable Systems Reliability; Reliability Data**). Depending on the failure mechanisms, the strategy adopted to improve the product reliability, and the kind of observed data, different models can be utilized. In the next section, a number of reliability growth models are briefly illustrated.

Reliability Growth Models

The general term, reliability growth models, is usually used to denote those models that are specifically applicable to the analysis of time-to-failure data. In this case, the reliability growth is generally modeled through stochastic point processes under the assumption that failures are highly localized events that occur

randomly in the continuum. Reliability growth model for time-to-failure data can be roughly classified into *continuous* and *discontinuous* models, and are briefly discussed in the following.

However, a broad class of models is devoted to the analysis of one-shot systems (such as guns, rockets, and missile systems) whose relevant data consist of sequences of dichotomous success–failure outcomes from successive testing stages. A complete survey of such models, generally referred to as *discrete* reliability growth models, can be found in [4].

Continuous Models

Continuous models are mathematical tools able to describe the *overall* trend in product reliability, resulting from the repeated applications of corrective actions. They are commonly applied in TAAF programs where the underlying failure process can be thought of as a decreasing stepwise function with test time (whose jumps depend on the effectiveness of corrective actions) that is fitted by a continuous curve representing the overall trend (see Figure 1).

The most commonly used of such models is the power-law process (PLP) (see **Repairable Systems Reliability**) [1], often known as the *Duane model* [5], whose intensity function $\lambda(t)$

$$\lambda(t) = \frac{\beta}{\alpha} \left(\frac{t}{\alpha} \right)^{\beta-1}, \quad \alpha, \beta > 0 \quad (1)$$

numerically equal to the rate $\rho(t)$, is a decreasing function with the test time t when $\beta < 1$. The expected number of failures up to t , say $M(t) = \int_0^t \rho(x) dx = (t/\alpha)^\beta$, is linear with t in a log–log scale.

The growth parameter β measures the management and engineering efforts in eliminating the failure modes during the development program. Small β values (e.g., $\beta \leq 0.5$) arise when an aggressive program is conducted by a knowledgeable team.

If the reliability growth program ends at the time T and no further improvement is introduced afterward, then the reliability level of tested prototypes at time T will characterize the reliability of the product as it goes into production. Besides, the intensity function of the product as it goes into production is generally assumed to remain constant during the so-called useful life period (see **Environmental Exposure Monitoring; Hazard and Hazard Ratio**)

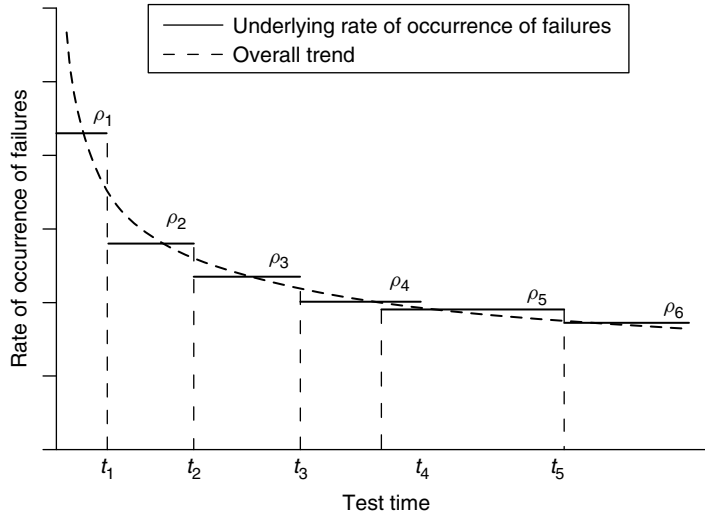


Figure 1 The underlying failure process in a TAAF program and its overall trend, where the product reliability in each testing stage i is measured by the rate of occurrence of failures

and equal to the intensity $\lambda_T \equiv \lambda(T)$ at the end of the development program [6]. The *current reliability* $R(\tau)$ (see **Repairable Systems Reliability; Mathematics of Risk and Reliability: A Select History**), i.e., the probability that the next failure occurs after τ during the useful life period, is

$$R(\tau) = \exp(-\lambda_T \cdot \tau), \quad \tau > 0 \quad (2)$$

which gives the success probability of a new product for a mission of fixed duration τ , too.

The PLP enjoys large popularity, partly because of the simplicity of inference procedures. Indeed, given the first n successive failures times observed during the test time T , say $t_1 < t_2 < \dots < t_n \leq T$, one can easily obtain graphical estimates of parameters by exploiting the linear behavior of $M(t)$ with t in a log-log scale [7]. Again, the maximum-likelihood (ML) estimators of parameters β and α are in closed form:

$$\hat{\beta} = \frac{n}{\sum_{i=1}^n \ln(T/t_i)} \quad \text{and} \quad \hat{\alpha} = \frac{T}{n^{1/\hat{\beta}}} \quad (3)$$

and exact confidence intervals for β can be easily obtained because $2n\beta/\hat{\beta}$ is a χ^2 random variable (r.v.) both in failure-truncated sampling (where

$T \equiv t_n$) and in time-truncated sampling (see **Detection Limits**), with $2(n-1)$ and $2n$ degrees of freedom, respectively [1]. Likewise, in failure-truncated sampling, one can obtain exact confidence intervals for $\lambda_n \equiv \lambda(t_n)$ and $R(\tau)$ by exploiting the result

$$\frac{\lambda_n}{\hat{\lambda}_n} \sim \frac{Z \cdot S}{4n^2} \quad (4)$$

where Z and S are independent χ^2 r.v.'s with $2(n-1)$ and $2n$ degrees of freedom, respectively [8]. Table 1 of [9] allows one to obtain exact confidence intervals for λ_n and $R(\tau)$. For time-truncated samples, approximate confidence intervals for λ_T and $R(\tau)$ can be obtained by using Table 2 of [9].

For the case of failure-truncated sampling, one can predict the *current lifetime* τ (see **Persistent Organic Pollutants**) by exploiting the result that $V = \hat{\lambda}_n \cdot \tau$ is a pivotal quantity, whose percentiles are given in Table 1 of [10] for selected n values.

For illustrative purposes, let now consider the $n = 14$ failure times (in a failure-truncated sampling) of a complex type of aircraft generator subject to TAAF

Table 1 Failure times in hours of aircraft generator^(a)

10	55	166	205	341	488	567
731	1308	2050	2453	3115	4017	4596

^(a) Reproduced from [11]. © American Statistical Association, 1998

program [11] (see Table 1). The product reliability is measured by the success probability for a mission of $\tau = 100$ h, and the reliability requirements are met if the lower 90% confidence limit for $R(100)$ is greater than 0.75. By using equations (3), the ML estimates are $\hat{\beta} = 0.613$ and $\hat{\alpha} = 61.9$ h, and the 90% confidence interval for β is (0.337, 0.851), thus providing empirical evidence that the reliability growth program has been effective. The ML estimates of λ_n and $R(100)$ are $\hat{\lambda}_{14} = 0.00187 \text{ h}^{-1}$ and $\hat{R}(100) = 0.830$, respectively. By using Table 1 of [9], the upper 90% confidence limit for λ_n is $\lambda_U = 0.00187/0.7087 = 0.00263 \text{ h}^{-1}$, from which the lower 90% confidence limit for $R(100)$ results in $R_L(100) = \exp(-\lambda_U \cdot 100) = 0.768$. Then, at the end of the development program the product meets the reliability requirements.

Another model frequently used in TAAF programs is the log-linear process (LLP) [12] whose intensity function:

$$\lambda(t) = \exp(\alpha + \beta t) \quad -\infty < \alpha, \beta < \infty \quad (5)$$

is more realistically equal to the positive value of $\exp(\alpha)$ at $t = 0$. In the reliability growth context, when $\beta < 0$, the LLP yields a defective distribution for the time to first failure (the mean of the first failure time is infinity).

Both the PLP intensity and the LLP intensity tend to zero as t approaches infinity, and hence such models cannot be used, for example, to analyze data collected during a TAAF program followed by a demonstration test dedicated to verify the attainment of reliability goals. Indeed, during the demonstration phase the rate of occurrence of failures is expected to be roughly constant and greater than zero. In this case, a more realistic model is the three-parameter LLP, whose failure intensity $\lambda(t) = \exp(\alpha + \beta t) + c$ ($-\infty < \alpha < \infty$, $\beta < 0$, and $c > 0$) approaches more realistically to the positive value of c as the test time increases.

When the initial reliability of the prototypes depends on the age of the manufacturing process, the *process-related reliability growth* model of [13] can be used. This process describes the design process through a PLP (with $\beta < 1$) and assumes that the scale parameter α is not constant, but depends on the age X of the manufacturing process when the prototype is manufactured.

Discontinuous Models

Discontinuous models are able to describe both TAAF programs and the cases where some or all of the fixes are delayed.

In case of TAAF program, Sen [11] and Pulcini [14] proposed two parametric models that assume that the failure intensity λ_i is constant during testing stage i ($i = 1, 2, \dots$), so that the resulting plot of the failure intensity as a function of test time is a nonincreasing series of steps. Thus, the interfailure times $t_i - t_{i-1}$ are independent exponential r.v.'s with mean $1/\lambda_i$. In particular, in [11] $\lambda_i = (\eta/\theta)i^{1-\theta}$ ($\eta > 0$ and $\theta > 1$), where the ratio η/θ is the starting failure intensity of the product. In [14], $\lambda_i = \mu\delta^{i-1}$ ($\mu > 0$ and $0 < \delta < 1$), where μ is the starting failure intensity, the *growth parameter* δ is equal to the ratio λ_{i+1}/λ_i , and hence $1 - \delta$ is the fraction of failures eliminated by each corrective action.

In case of delayed fixes, the reliability growth model depends also on the assumptions regarding failure mechanisms and repair policy during each development stage, because more than one failure can occur during each stage.

The nonparametric model of [15] assumes that the prototype is perfectly repaired at failure during each stage of a TFT program. Thus, the failure process is a series of renewal processes, where the probability distribution governing the interfailure times changes stage by stage. The model in [16] assumes that the interfailure times during each stage are exponentially distributed, so that the failure process is a series of homogeneous Poisson processes (see **Repairable Systems Reliability**) whose failure intensity decreases stage by stage.

Both these models can be used for nonrepairable or simple products, but are not adequate for the analysis of complex repairable systems subject to TFT program, for which the discontinuous model proposed by Ebrahimi [17] and Calabria *et al.* [18] is more suitable. This model assumes that the prototypes can experience deterioration with operative time during each stage, are substituted with new improved copies at the beginning of each stage, and are minimally repaired at failure during each stage. The failure process in each stage is described by a PLP with $\beta > 1$, where the parameter β measures the deterioration rate of the system within each stage and is assumed to be constant over stages (fixes do not alter the failure mechanism). The scale

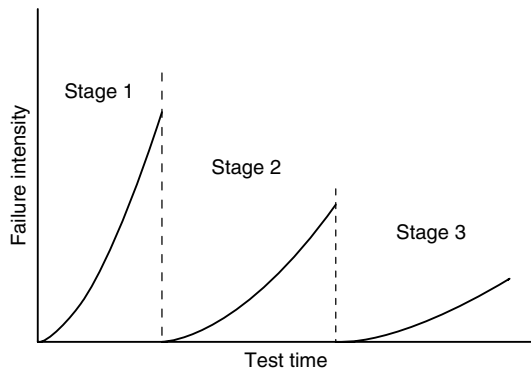


Figure 2 Hypothetical behavior of the failure intensity of a deteriorating system undergoing a Test-Find-Test program

parameter α increases stage by stage, as a result of reliability growth. Figure 2 depicts a hypothetical behavior of the failure intensity during such a TFT program. ML estimates can be found in [17, 19], whereas [18] proposes a Bayesian approach that is mainly addressed to the solution of a decision-making problem (to accept the current design of the equipment for mass production, or to continue the development program).

References

- [1] Crow, L.H. (1974). Reliability analysis for complex, repairable systems, in *Reliability and Biometry*, F. Proschan & R.J. Serfling, eds, SIAM, Philadelphia, pp. 379–410.
- [2] Benton, A.W. & Crow, L.H. (1989). Integrated reliability growth testing, *Proceeding Annual Reliability and Maintainability Symposium*, Atlanta, pp. 160–166.
- [3] Giorgio, M., Guida, M. & Pulcini, G. (2006). Reliability-growth analysis of locomotive electrical equipment, *Journal of Quality Technology* **38**, 14–30.
- [4] Fries, A. & Sen, A. (1996). A survey of discrete reliability-growth models, *IEEE Transaction on Reliability* **45**, 582–604.
- [5] Duane, J.T. (1964). Learning curve approach to reliability, *IEEE Transactions on Aerospace* **2**, 563–566.
- [6] Crow, L.H. (1975). Tracking reliability growth, in *Proceedings 20th Conference on Design of Experiments*, Report 75-2, ARO, pp. 741–754.
- [7] Murthy, D.N.P., Xie, M. & Jiang, R. (2004). *Weibull Models*, John Wiley & Sons, New York.
- [8] Lee, L. & Lee, S.K. (1978). Some results on inference for the Weibull process, *Technometrics* **20**, 41–45.
- [9] Crow, L.H. (1982). Confidence interval procedures for the Weibull process with application to reliability growth, *Technometrics* **24**, 67–72.
- [10] Calabria, R. & Pulcini, G. (1996). Maximum likelihood and Bayes prediction of current system lifetime, *Communications in Statistics: Theory and Methods* **25**, 2297–2309.
- [11] Sen, A. (1998). Estimation of current reliability in a Duane-based reliability growth model, *Technometrics* **40**, 334–344.
- [12] Cox, D.R. & Lewis, P.A.W. (1978). *The Statistical Analysis of Series of Events*, Chapman & Hall, London.
- [13] Heimann, D.I. & Clark, W.D. (1992). Process-related reliability-growth modeling. How and why, *Proceeding Annual Reliability and Maintainability Symposium*, Las Vegas, pp. 316–321.
- [14] Pulcini, G. (2001). An exponential reliability-growth model in multicopy testing program, *IEEE Transaction on Reliability* **50**, 365–373.
- [15] Robinson, D.G. & Dietrich, D. (1987). A new nonparametric growth model, *IEEE Transaction on Reliability* **36**, 411–418.
- [16] Robinson, D. & Dietrich, D. (1989). A nonparametric-Bayes reliability-growth model, *IEEE Transactions on Reliability* **38**, 591–598.
- [17] Ebrahimi, N. (1996). How to model reliability-growth when times of design modifications are known, *IEEE Transactions on Reliability* **45**, 54–58.
- [18] Calabria, R., Guida, M. & Pulcini, G. (1996). A reliability-growth model in a Bayes-decision framework, *IEEE Transactions on Reliability* **45**, 505–510.
- [19] Pulcini, G. (2002). Correction to: how to model reliability-growth when times of design modifications are known, *IEEE Transactions on Reliability* **51**, 252–253.

Related Articles

Repair, Inspection, and Replacement Models

GIANPAOLO PULCINI

Reliability Integrated Engineering Using Physics of Failure

Reliability

Reliability is the ability of a product to perform without failure and within specified performance limits, for a specified time, in its life-cycle application conditions. The development of a reliable product is not a matter of chance or good fortune; rather, it is a rational consequence of conscious, systematic, and rigorous efforts conducted through the entire life cycle of the product. Acceptable product reliability can only be assured through robust product designs, capable processes that are known to be within tolerances, and qualified components and materials from vendors whose processes are also capable and within tolerances. Quantitative understanding and modeling of the relevant failure mechanisms can guide design, manufacturing, and the planning of test specifications and tolerances.

Physics of failure (PoF) is a methodology for reliability assessment, which is fundamental to design for reliability (DfR) (see **Reliability Optimization**), reliability prediction, test planning, and prognostics. The objective is to design and test a product using parts and materials that have been sufficiently characterized in terms of performance over time when subjected to the manufacturing and application (environmental and operational) profile conditions. To carry out this objective, it is necessary to identify the potential mechanisms of failure, such as corrosion, fatigue, hot carriers, and electromigration, to name a few. The assessment of failure mechanisms forms the foundation for PoF (see Table 1 for definitions).

A failure mechanism (see **Failure Modes and Effects Analysis; Common Cause Failure Modeling**) is a chemical, mechanical, thermodynamic, or other process that is the root cause of the failure occurrence. Whether a specific failure mechanism will arise in a product and cause failure will depend on the environmental load conditions (e.g., temperature, humidity, contaminants, and radiation), and the operating load conditions (e.g., voltage, current, and generated temperature). The failure mechanisms will also depend on the product materials and the product

composition, including how the materials are composed and connected.

Failure sites are the locations where failure mechanisms precipitate. In electronic products, failure sites include semiconductor chips, component interconnects, circuit board metalization, interconnects (solder joints) between the component and the printed circuit, and external connections, such as connector and wiring.

The PoF methodology uses failure mechanism models (sometimes called *PoF models*) to assess the onset of potential failure mechanisms at specific failure sites in the product. A failure mechanism model will depend on the product composition and the stress (especially at the failure site). Stress is defined as the *intensity of a load* (e.g., electrical, thermal, mechanical, and chemical) at a failure site. The result of the PoF model is a probabilistic time to failure for a percentage of likely products to fail due to the failure mechanism.

The Physics-of-Failure Methodology

PoF methodology requires a set of information as input on which it performs a series of analysis and as a result provides a set of outputs that can be customized depending on the application. This section describes the steps, as shown in Figure 1, in this process.

Inputs

The PoF methodology requires inputs of the hardware configurations and the life-cycle profile (LCP) of the product under analysis. The details of the inputs may vary depending on the stage of product development.

Hardware Configurations. A product consists of a number of components and subassemblies working together to deliver the overall function of the product. Each subassembly may consist of lower level assemblies that are also interconnected. The architecture of a product describes the physical and functional relations between the subassemblies. The hardware configuration of the product being evaluated by PoF method describes the design of the components and subassemblies and the product architecture. It may also include the effects of the manufacturing processes on the final product in the form of tolerances on the dimensions and material properties.

This situation may require the elastic modulus of the connector elements, loading elements, and their housings to determine the contact force and its degradation pattern. Properties for common materials used in electronic products can be found in [1, 2].

Products are not normally produced by a single manufacturing process, but require a sequence of different processes to achieve all the required attributes of the final product. The manufacturing process applies stresses on materials, may produce residual stress, and may even modify some of material properties. For example, lead-free reflow profile can change the thermophysical properties of a printed circuit board. The variations in geometry and material properties caused by different manufacturing processes need to be inputs to the PoF methodology.

Life-Cycle Profile (LCP). Loads applied to the product during its life cycle from environment and operation drive the processes that lead to product degradation and failure. The life cycle of a product includes manufacturing and assembly, testing, rework, storage, transportation and handling, operation (e.g., modes of operation and on-off cycles), repair, and maintenance. The life-cycle loads include thermal (steady-state temperature, temperature ranges, temperature cycles, and temperature gradients), mechanical (pressure levels, pressure gradients, vibrations, shock loads, and acoustic levels), chemical (aggressive or inert environments, ozone, pollution humidity levels, contamination, and fuel spills), physical (radiation, electromagnetic interference, and altitude), and/or operational loading conditions (power, power surge, heat dissipation, current, and voltage spikes). These loads, in various combinations, can influence the reliability of the product. The extent and rate of product degradation depend upon the nature, magnitude, and duration of exposure to such loads.

An LCP is a time history of events and conditions associated with an item of product from its release from manufacturing to its removal from service. The life cycle should include the various phases that an item will encounter in its life, such as handling, shipping, and storage prior to use; mission profiles while in use;^a phases between missions, such as standby or storage, transfer to and from repair sites and alternate locations; and geographical locations of expected deployment.

LCP influences the stress levels on the product and failure mechanisms that will be active on a product. For example, the frequency of use may determine the number of temperature cycles to which a product is exposed on a daily basis. During transportation, the product may be subjected to substantial vibration loading. In storage, elevated humidity levels may be present. Under operation, an electrical bias may exist between isolated conductors. The level of temperature, vibration, humidity, and voltage potential may be inputs to determine thermomechanical stress, mechanical strain, moisture content, and dendritic growth, respectively.

It is also necessary to identify the critical loads to a product under certain conditions. A product will experience numerous loads, but it is not always necessary to process all the loads in the analysis steps. Some of the loads will play major roles in activating and accelerating the failure of the product, while other loads can be ignored. For example, radiation can often be ignored for ground-based electronic products, since the radiation level is too low to produce any effect to the function of the product or cause any damage. The same load may also be considered to have different criticality for different products under different conditions. For example, a mechanical shock caused by dropping a cell phone onto the hard ground, which is critical to the electronic parts in the cell phone, will have no effect on a tire and can be ignored. The loads that can or cannot be ignored depend on the critical failure mechanisms that are identified in the analysis step.

Analysis

PoF analysis steps determine the failure mechanisms that are active on a product and evaluate the effects of those failure mechanisms on the possible failure sites of the product. The analysis steps are conducted in sequence and the result of one step is used as a part of input to the next step.

Failure Modes, Mechanisms, and Effects Analysis (FMMEA). Failure modes, mechanisms, and effects analysis (FMMEA) is an essential step in the PoF analysis that uses hardware configuration and the LCP to identify critical failure mechanisms. FMMEA utilizes the basic steps in developing a traditional FMEA in combination with knowledge of PoF. It then uses LCP to identify active stresses and select

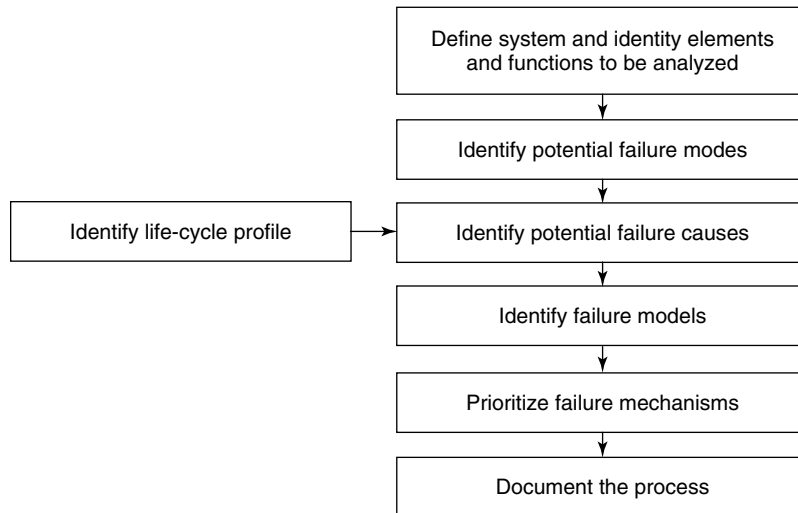


Figure 2 FMMEA methodology

the potential failure mechanisms. Knowledge of load type, level, and frequency combined with failure site are used to prioritize failure mechanisms according to their severity and likelihood of occurrence. Figure 2 is a schematic diagram of FMMEA.

FMMEA is based on understanding the relationships between product requirements and the physical characteristics of the product (and their variations in the production process), the interactions of product materials with loads (stresses at application conditions), and their influence on the product's susceptibility to failure. FMMEA combines life-cycle environmental and operating conditions and the duration of the intended application with knowledge of the active stresses and potential failure mechanisms. Potential failure mechanisms are determined on the basis of appropriate available mechanisms corresponding to the material system, stresses, failure modes, and causes.

FMMEA prioritizes the failure mechanisms on the basis of their occurrence and severity to provide guidelines for determining the major operational stresses and environmental and operational parameters that must be accounted for in the design or be controlled. LCP is used for evaluating failure susceptibility. If certain environmental and operating conditions are nonexistent or generate a very low-level stress, the failure mechanisms that are exclusively dependent on those environmental and operating conditions are assigned low occurrence. Severity ratings

are obtained from the failure modes associated with the mechanism. The high-priority failure mechanisms identified through combination of occurrence and severity are the critical mechanisms. Each critical failure mechanism has one or more associated sites, modes, and causes in an FMMEA result.

The basic categories of failures are overstress (i.e., based on stress strength interference) and wear out (i.e., based on damage accumulation); they are often identified through a mode of going beyond performance tolerance (e.g., excessive propagation delays). Overstress and wear-out failures generally result from irreversible material damage; however, some overstress failures can be caused by reversible material damage (e.g., elastic deformation) [3]. Failures can also be broadly categorized by the nature of the loads that trigger or accelerate the mechanisms. Examples of both failure types for electronic assemblies are given in Table 2.

Failure models (*see Probabilistic Risk Assessment*) are used as tools to assess failure propensity. There are two different categories of models: analytical models and PoF models. PoF methodology needs inputs of PoF-based models rather than analytical models. Many models, such as the Arrhenius model, the Coffin–Mason model and the Steinberg model, exist for predicting the behavior of components and products. Different models have different associated assumptions, which limit their applications to specific ranges of conditions.

Table 2 Examples of overstress and wear-out failure mechanisms for electronic products^(a)

Failure mechanism			
Overstress (environmental extreme)		Wear out (life)	
Mechanical	Yield, fracture, and delamination	Mechanical	Fatigue, creep
Thermal	Softening due to glass transition	Thermal	Diffusion voiding
Electrical	Electrical overstress, electrostatic discharge, and dielectric breakdown	Electrical	Electromigration, conductive filament
Radiation	Single-event upsets	Radiation	Embrittlement charge tapping
Chemical	Etching	Chemical	Corrosion, dendrites, and intermetallic growth

^(a) Reproduced from [4]. © IEEE, 2003

In PoF models, the stresses and the various stress parameters and their relationships to materials, geometry, and product life are considered. Each potential failure mechanism is represented by one or more of the prevalent models. For electronic products, there are many PoF models describing the behaviors of components like printed circuit boards, interconnections, and metalizations under various conditions, such as temperature cycling, vibration, humidity, and corrosion. A model should provide repeatable results, reflect the variables and interactions that are causing failures, and predict the behavior of the product over the entire domain of its operational environment. This type of model allows development of accelerated testing and may help reduce the number of test runs.

Many PoF models are available in the literature, such as strain-range-based model for solder joint fatigue [5], which describes temperature-cycle-induced solder interconnect fatigue; Black's model and its variations [6], which describes the electromigration in semiconductor devices; the Fowler–Nordheim model [7], which describes the time-dependent dielectric breakdown due to tunneling in gate oxide devices; and Pecht and Rudra model [8] that described conductive filament formation in printed circuit board. Models applicable to electronic products are available in [9–11]. If no models are available or if the models are found to be not applicable to specified failure sites and loads, then new models can be developed by using controlled experiments that identify the design and environmental

factors governing failure and the mathematical relationship linking those factors to the time to failure.

Stress Analysis. Stress analysis is used to convert the loads on the system to stresses at the potential failure sites. The product's life-cycle loads and configuration information are inputs to the stress analysis process. Stress analysis is conducted for the stresses that drive the critical failure mechanisms identified by FMMEA. For electronic products, thermal analysis, thermomechanical analysis, and vibration analysis are often performed to determine the inputs for the relevant failure models.

Thermal analysis is used to determine the temperature distribution within a product. For a printed wiring board (PWB) in an electronic product, the thermal analysis will output the temperature on and within the board, and at electronic part cases and junctions. Thermal analysis involves the solving of heat transfer equations for three fundamental modes of heat transfer: conduction, convection, and radiation. Thermal analysis may be transient or steady state. In most cases, steady-state thermal analysis results provide sufficient inputs to identified failure models. For example, a temperature cycle is usually defined to occur between a powered-on operating state and a powered-off state. Here, a steady-state thermal analysis may be used to establish the powered-on state. In some instances, the transient behavior may be important. For example, if a significant delay in the change of temperature occurs between connected structures,

the connection point may be subjected to a more significant stress than estimated by a steady-state evaluation. In terms of the component temperature, thermal resistance networks are often used to evaluate case and junction temperatures. Once a thermal analysis problem can be expressed, there are usually several ways of solving it, including finite difference, finite element, resistance network, and bulk analysis [1].

Vibration analysis is often used to determine the response of the PWB due to the random oscillating motion of the structure that contains the PWB. While calculating the natural frequency of a PWB, the boundary conditions are assigned on the basis of understanding of the mounting conditions of the board in hardware configurations to conditions such as free, simple, or clamped. Numerically, the natural frequency and first-order response of a rectangular PWB with complete edge supports can be determined using first-order approximations. However, finite element modeling is usually required to evaluate a more diverse and practical range of supports.

The results of stress analysis are often verified and calibrated using experimental results. For example, the natural frequency of a PWB can be experimentally determined using the strain gauge or accelerometer on a PWB, attaching the PWB to a dynamic shaker, and measuring the response of the PWB to a known input.

Reliability Assessment. With the failure sites, stress inputs, and failure models, the reliability of a product is estimated and reported in terms of times to failure of its individual functional units. To define failure, damage metric is established. The damage metric value is augmented as the effects of failure mechanisms degrade a failure site. For fixed duration load events, the augmentation is based on the severity and duration of the event. For example, drop of a handheld device results in damage representative of the height of the drop. For repetitive events, a damage rate is established and the damage metric is updated on the basis of number of repetitive events. For thermal cycling events, the damage rate is defined for each cycle and the damage accumulation is proportional to the number of thermal cycles the product endures. When the damage metric exceeds a threshold value, the failure site is identified as failed and the time to failure for the site is determined

(see **Design of Reliability Tests; Reliability Growth Testing; Mathematics of Risk and Reliability; A Select History; Human Reliability Assessment**).

In general, time-to-failure data is obtained as a distribution for each failure site and failure mechanism. This distribution on time to failure is achieved by using the inputs to the failure models as distributions. In reality, all dimensional and material properties are distributed about a nominal value as a result of the normal variations in manufacturing. The same is true for the environmental loads. PoF-based reliability assessment allows for utilization of these natural variations in the reliability assessment. With the time-to-failure distribution on each site known, reliability can be evaluated in different metrics such as hazard rate, warranty return rate, or mean time to failure.

Sensitivity analysis is used to evaluate the sensitivity of the product life to the application, and define the safe operating region for the desired LCP. Sensitivity analysis involves, for each of the dominant failure mechanisms, identification of the functional relationships of time to failure to geometry, material property, and environmental parameters of interest. Sensitivity analysis can be used to identify the optimal values of accelerated stress loads, based on the sensitivity of the dominant failure mechanisms to the loads.

Outputs

The PoF analysis results can be obtained as output in different ways and can be used for various applications. The results can be as basic as a ranked list of potential failures or as sophisticated as design of life-consumption monitoring (LCM).

Ranked Potential Failures. The PoF methodology output can be obtained as a ranked list of potential failures, with associated potential failure modes, sites, and mechanisms. The ranking is obtained by evaluating the severity and occurrence of various failure mechanisms. The failure mechanisms with a higher ranking are considered riskier than others. Higher risk failure mechanisms are given higher priority when product design is updated, screening conditions are determined, accelerated testing plans are designed, and prognostics and remaining life prediction are conducted.

Design Trade-Offs. PoF analysis can be performed on multiple product design and manufacturing variants to compare their reliability and other attributes to perform trade-off analysis between metrics of interest, such as cost and reliability. The sensitivity analysis in PoF assessment also provides a way to obtain reliability as a function of product attributes. Hence, implementing PoF into the design makes the process of trade-off analysis efficient and effective. Typically, the PoF approach outputs the time to failure and the ranked list of failure mechanisms associated with failure sites. The ranked list allows designers to concentrate on locations and mechanisms that are most likely to cause product failure. The PoF methodology may be used to guide design improvement by identifying the drivers for the dominant failure mechanisms during each life-cycle phase. Design trade-offs can then be evaluated by determining the sensitivity of the dominant degradation mechanisms to the drivers of the mechanisms.

If the reliability of a specific design made by a specific manufacturing process cannot meet the reliability requirements within the cost and lead time constraints, guidance from the PoF analysis can be used to improve the reliability. The options for the trade-offs include using a different material with better thermomechanical properties, improving a current manufacturing process with lower variability, or implementing a new process to reduce or avoid damage to materials during the process, and/or revising a

product's architecture, such as adding or redesigning a heat sink. Since factors like these are inputs to the PoF analysis, the effects of these changes on product reliability can be computed and compared with the original and alternative designs.

Virtual Qualification. Virtual qualification is a methodology for assessing and improving the durability of electronic equipment through the use of validated failure models/simulation tools. Virtual qualification (also called *simulation-assisted reliability assessment*) assesses whether a part or system can meet defined life-cycle requirements based on its materials, geometry, and operating characteristics. The technique involves the application of simulation software to model physical hardware to determine the probability that the system will meet the desired life goals [12–14].

By implementing PoF, virtual qualification can be applied at different design stages and hence move reliability assessment into the design phase [15, 16]. This accelerated product qualification process emphasizes the physical process that produces failures in electronic systems. It allows the design team to consider qualification at the initial stages of design, technology and functional definition, and supplier selection. A flowchart of virtual qualification is shown in Figure 3. This system takes advantage of advances in computer-aided engineering software permitting components and systems to be qualified

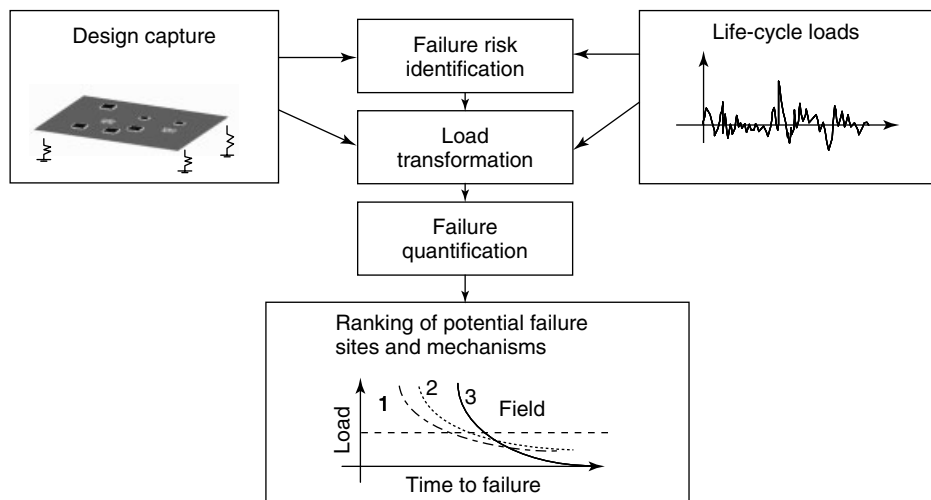


Figure 3 Flowchart of virtual qualification

on the basis of analysis of the susceptibility of their designs to failure due to a number of fundamental physical and chemical mechanisms. The reliability assessment tool assesses the candidate and existing package or module designs for reliability in many different environments, using a database of fully validated PoF models. It calculates times to failure for the mechanisms that cause failures and evaluates the effects of different manufacturing processes on reliability by calculating the time to failure as a function of typical manufacturing tolerances and defects. It facilitates the selection of cost-effective test parameters for validating reliability assessment and design and also aids in the selection of components by permitting their virtual qualification [17].

Screening Conditions. Screening conditions of electronic products are often selected from government/military specifications as MIL-STD-883 [18], but screening conditions are arbitrary and not based on understanding of the defects in products to its failure propensity. PoF approach to screening is based on the fact that the stresses and stress magnitudes, which affect the dominant failure mechanisms, are controlled by the levels of defects in a product. The screening conditions are thus determined by the calculation of threshold levels of defects in a product. Below these threshold levels of defects, a product is likely to be reliable. The PoF outputs identify the magnitude of defects that can be allowed to pass the screens. PoF is then used to determine the screening loads and magnitudes to isolate the products that have defects above the threshold level.

Since PoF identifies the correlation between the defect magnitudes and operational life, the failures during in-process testing and screening can be evaluated against the prediction for modes and mechanisms. Thus, PoF method of screening condition development allows for validating the effectiveness of screens and allows its improvement.

Accelerated Test Conditions. Accelerated testing is intended to shorten the normal time to failure of a product due to a specific failure mechanism, generally through application of higher stress levels on a product. The application of increased stresses is usually for the purpose of (a) ruggedizing the design and manufacturing process of the package through systematic step-stress testing and increasing the stress margins by corrective action (reliability enhancement

testing), (b) conducting compressed/accelerated life tests in the laboratory to verify in-service reliability (qualification testing), and (c) eliminating weak or defective specimens from the main population (quality conformance testing). In PoF-based design, the accelerated test formulation is based on an understanding of the failure mechanisms and associated causes. Accelerated testing can be a value-added activity that allows verification of design goals over a short period, thereby achieving faster design modification. However, the value of accelerated tests can be realized only if there is an associated predictive capacity regarding the effect of accelerated tests on failure modes and mechanisms.

The accelerated test conditions can be seen as a corollary to virtual qualification, where the output is provided in the form of a test plan. This test plan will include test stress type(s), stress levels, duration of the test, definitions of failure, expected times of failures, and the site, mode, and mechanism. In addition, if the accelerated tests are meant to be qualification or acceptance tests, then acceptance criteria for such decisions can also be identified in the plan. From a test standpoint, acceleration factors can be established between the anticipated LCP and the set of qualifying tests. Redundant and nonvalue added tests can be eliminated.

Prognostics and Life-Consumption Monitoring Process. Prognostics is a method of evaluating the extent of a product's reliability in terms of product degradation in its life-cycle environment. Determining the impending failure based on actual life-cycle application conditions allows procedures to be developed to mitigate, manage, or maintain the product. To reduce the design margin and to predict life more accurately, an increasing emphasis has been placed on prognosis of remaining useful life of components through physics-based damage models or models based on microstructure and constitutive properties. To assess damage and predict the remaining useful life and to use the entire capability of a given material, it is critical to convey the PoF of the material and the material constitutive properties to the designers involved in selecting materials and sizing components. The results of PoF analysis can be used to help determine what to monitor, where to monitor it, and how to react to the results of monitoring.

LCM is one of the prognostic approaches that assess the life consumption and the remaining life of a

product in its life cycle. The life-consumption process involves continuous or periodic measurement, sensing, recording, and interpretation of physical parameters associated with a system's life-cycle environment to quantify the amount of system degradation [19]. The PoF analysis output for a prognostics and life-cycle monitoring process will include a life-cycle environment and operation monitoring scheme and the failure mechanisms and models that can be used to determine the remaining life of the products.

PoF analysis also provides a basis for performing environmental stress management. Stress management solutions can be evaluated in terms of their effectiveness in reducing harmful loads and increasing the useful life of a product. For example, the response of a circuit card to various levels of vibration could be used to determine the amount of damping required to adequately protect the circuit card. Stress management can also take the form of derating components to reduce their electrical stress levels and increase their times to failure.

Applications of Physics of Failure in Product Life Cycle

PoF methodology and design tools allow reliability engineering to become "proactive" rather than "reactive", and have numerous applications in product design, manufacturing, and support. The PoF methodology and tools have reduced the number of time-consuming and expensive hardware design and test iterations, making it feasible to assess and improve reliability in the conceptual design stage by bringing together the reliability team members in the supply chain before expensive manufacturing commitments are made. Some successful implementations of PoF are discussed in this section.

Virtual Qualification

Virtual qualification of electronic hardware has been used for many years. One of the first documented cases of virtual qualification involved the examination of circuit cards in military radios.

In this case, three circuit cards were modeled in virtual qualification software [20, 21]. The life-cycle load consisted of temperature cycles due to power-on and power-off cycles and vibration load due to mounting in a mobile platform. From the hardware

and the loading conditions, solder interconnects were identified as failure sites and vibration and temperature cycling fatigue were identified as failure mechanisms. Evaluation of the anticipated LCP resulted in a prediction that the product would not meet its defined life-cycle requirement. To verify this result, an accelerated test composed of both temperature cycling and vibration was developed using the virtual qualification model. Test of 10 production units closely match simulation results. On the basis of this finding, the design was modified and reassessed. The modified design was estimated to provide required life and demonstrated to match simulation in test. As a result of the change, the program savings was estimated to be in excess of \$27 million.

In another study, the virtual qualification was used to justify the use of a commercial circuit card assembly (CCA) in place of a ruggedized version. Again, solder interconnects were the identified failure sites and fatigue was the identified failure mechanism. Simulating the life expectancy for the anticipated LCP, the commercial CCA was found to have sufficient life margin for use in the application. Testing verified this simulation. This justification resulted in a savings of \$12 000 per assembly and a program saving of nearly \$1 million for only the first 10 systems [1, 22].

Accelerated Testing

Understanding failure mechanisms is essential for conducting successful accelerated life tests, as advocated in PoF-based reliability prediction methodologies. One of the critical issues is the need for a rational method to quantitatively relate the results of accelerated wear-out tests to in-service reliability, using a scientific acceleration transform. PoF methodology involves quantitative modeling of the expected root-cause failure mechanisms. The output is an estimate of the acceleration factor, which, when multiplied by the accelerated wear-out test data, results in life estimation under life-cycle conditions. The range of accuracy of life estimates in the field can now be related to the scatter in the accelerated test data and to the accuracy in estimating the acceleration factor (which varies with the accuracy of input information and the assumptions built in the failure model) [23]. A successful PoF-based accelerated stress testing program meets product requirements, lowers life-cycle costs, and reduces the time to market for a product.

Guidelines to implement PoF in accelerated wear-out testing have been presented with a case study on an avionics electronic module in an avionics box located in the electronic equipment (EE) bay of a commercial aircraft by Upadhyayula and Dasgupta [23]. This case study focused on the reliability of the surface-mount interconnects, particularly the possibility of enhancing test-time compression by exploring the effects of combined vibration and temperature cycling loads on interconnect fatigue.

The sources of reliability risks under combined life-cycle loads were determined by identifying the potential weak links and dominant failure mechanisms associated with the product. PoF activities include understanding the hardware configuration, determining the product life-cycle loads in the EE bay of the aircraft, and performing an initial PoF assessment of the potential failure modes of the interconnects in the circuit card assembly module. The output ranks the potential weak links expected under life-cycle load combinations.

After all the relevant material properties, part lists, and part drawings were identified and documented, combined thermal and mechanical loads were applied to the assembly, which targets the interested failure mechanism, fatigue of interconnects. The interaction effects between thermal and vibration loads were captured, as shown in Figure 4. Thermomechanical stress analysis was conducted for the temperature history, using viscoplastic finite element method (FEM), to estimate the cyclic thermomechanical stress state in the solder joint. Solder joint life predictions were obtained using the generalized Coffin–Manson damage model; the total damage accumulated was computed using Palmgren–Miner’s hypothesis at each increment, rather than for the entire load history.

The PoF approach to accelerated stress tests includes designing the test loads and the test vehicle, as well as issues related to test setup. Different combinations of loads were applied on different test vehicles. The actual magnitudes of the accelerated stress test loads were determined through a combination of exploratory tests and PoF analysis. Test vehicle characterization was carried out to investigate and calibrate the specimen response to applied thermal and vibration loads. Failure data from over-stress tests was used to compute, using PoF-based timescaling techniques, the accelerated test load levels required to excite wear-out failure mechanisms

in accelerated life tests of the specimen. The specimen response is calibrated over the entire accelerated load range. The characterization, therefore, takes into account the interaction effects between the applied thermal loads and the observed vibration response during the test.

The PoF approach is applied to the accelerated test environment to predict the failure trends by quantifying the rate of damage accumulation as a function of the level of applied stress. Results from this exercise confirm the PoF conjecture that thermal stresses, under combined vibration and thermal environments, can actually inhibit damage caused by vibration fatigue in surface-mount interconnects.

The last step is to verify the reliability assessment. It was confirmed by experimental observations that certain surface-mount interconnects are more vulnerable to repetitive shock vibration (RSV) at room temperature than to a combined application of thermal cycling and vibration loading. The PoF approach captures the complex synergy between vibration and temperature by quantifying the rate of damage accumulation owing to RSV loading as a function of the applied temperature level and the mean stresses experienced due to thermal cycling.

Prognostics and Life-Consumption Monitoring

Prognostics and health management (PHM) is a method that permits reliability assessment of a product or system under its actual application conditions. LCM is a prognostic method to assess the remaining life of a product in its actual life-cycle environment by continuously or periodically measuring the product’s performance parameters and environmental conditions.

One case study of LCM is the remaining life estimation of a test board assembly [19], which was placed under the hood of a 1997 Toyota 4Runner and subjected to normal driving conditions. First, failure mode and mechanism analysis was conducted on the circuit card assembly. Potential failure modes and mechanisms of the circuit board assemblies under certain conditions were identified. Then virtual reliability assessment was conducted by computing the time to failure for the identified failure modes and mechanisms. The temperature and vibration data were used as inputs for stress and damaged models. After temperature and vibration analysis, stress and damage accumulation analysis was conducted to

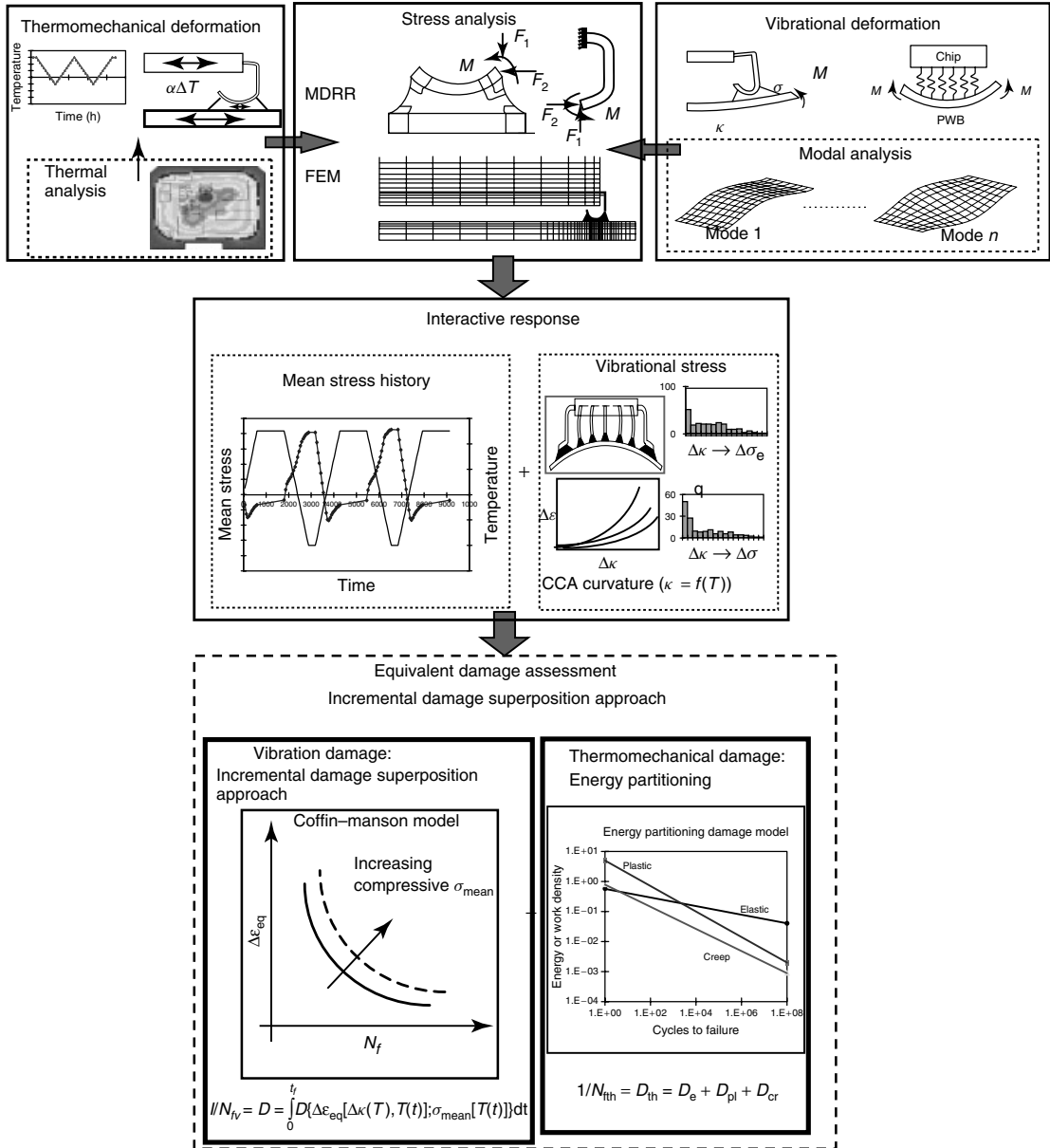


Figure 4 Assessment of potential failures for combined accelerated test loads [Reproduced from [23]. © John Wiley & Sons, Ltd, 1998.]

estimate the damage accumulation. Palmgren–Miner linear damage accumulation theory was used for the computation, as shown in the following equation.

$$\sum_{i=1}^m \frac{n_i}{N_i} = C \quad (1)$$

where C is a constant. This theory states that where there are m different stress magnitudes in a spectrum, S_i ($1 \leq i \leq m$), each stress contributing $n_i(S_i)$ cycles, then if $N_i(S_i)$ is the number of cycles to failure at a constant stress S_i , failure occurs when the above equation is satisfied.

Then, the remaining life of the circuit card assembly was estimated by subtracting the life consumption of a day from the remaining life of the previous day in an iterative manner, as shown in equation (2)

$$RL_N = RL_{N-1} - DR_N \cdot TL_{N-1} \quad (2)$$

where RL_N is the remaining life at the end of the day N , TL_N is the estimated total life at the end of day N , and DR_N is the damage ratio accumulated at the end of day N . The calculated remaining life and the experimental life are plotted in Figure 5. The remaining life estimation was verified by the experimental monitoring of resistance during the test process. The remaining life was correlated to the resistance increase in the solder joints, and failure was defined as occurrence of 15 resistance increase spikes within a certain time range. It can be seen from the plot that the estimated remaining life value and actual remaining life value are close to each other accidentally.

Acceptance of Physics-of-Failure Methodology

There are many reasons for assessing the reliability of a product. We perform reliability assessment to establish reliability requirements for preliminary design

specifications, planning documents, and requests for proposals; to evaluate the feasibility of designs and manufacturing processes to determine whether the reliability requirements can be met and to make trade-offs, to identify and rank potential failures for corrective actions; to reduce the time needed to gather data for risk assessment; to plan for qualification and test tailoring for root-cause analysis; to assist in warranty, life-cycle cost and product support planning (e.g., maintenance and spares); to address certification and regulatory concerns including safety; and to increase customer confidence.

PoF methodology determines the root cause of failures. It provides a methodology for building in reliability. It assesses the impact of hardware configuration and life-cycle stresses on root-cause failure mechanisms. PoF represents good engineering practices. The PoF approach incorporates reliability into the design process by establishing a scientific basis for evaluating new materials, structures, and technologies. Information to plan tests and screens and to determine electrical and thermomechanical stress margins are identified by the approach.

Estimation of failure mechanisms as the basis of reliability assessment and qualification test design for a system has been accepted by the EIA/JEDEC and SEMATECH, an association of semiconductor manufacturing companies like Intel, IBM, AMD, Infineon,

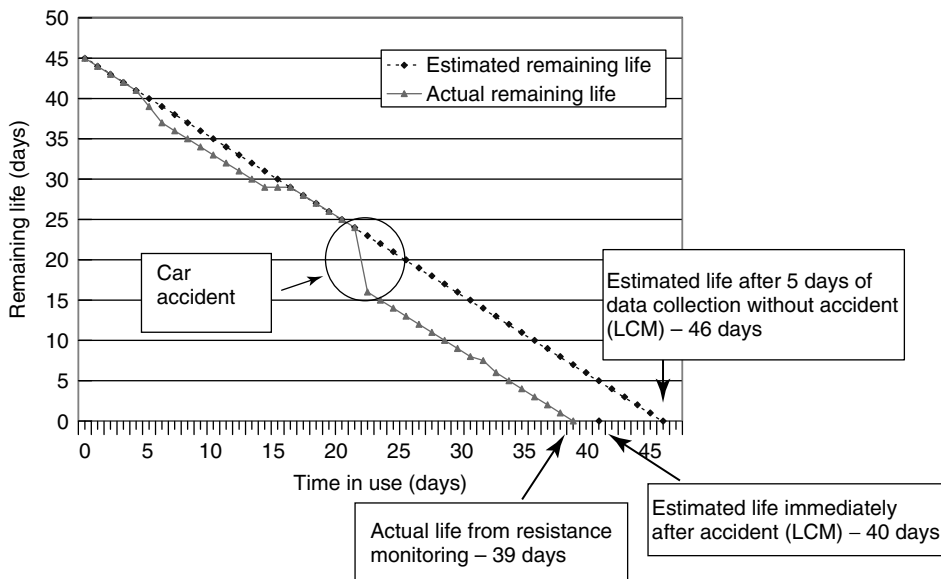


Figure 5 Plot of estimated remaining life as a function of time [Reproduced from [19]. © IEEE, 2003.]

Phillips, and Texas Instrument.^b Intel's product qualification method [24], which defines environmental, lifetime, and manufacturing use conditions based on target market segments (e.g., desktop, server, and notebook), determines probable stresses with reliability implications, estimate stress levels for modeling, and the modes of testing (e.g., standalone and board mounted), defines accelerated stress conditions necessary to identify failure mechanisms and determines the final stress conditions for testing, follow the PoF principles.

The IEEE 1413 standard [25] identifies the framework for the reliability prediction process for electronic systems (products) and equipment. An IEEE 1413-compliant reliability prediction report must include reasons why the reliability predictions were performed, the intended use of the reliability prediction results, cautions as to how the reliability prediction results must not be used, and where precautions are necessary.

In order to assist the user in the selection and use of a particular reliability prediction methodology complying with IEEE 1413, a list of criteria is provided in IEEE 1413.1 [26]. The criteria consist of questions that concern the inputs, assumptions, and uncertainties associated with each methodology, enabling the risk associated with the methodology to be identified. The PoF-based methodology for reliability prediction fared best in comparison to traditional reliability prediction methods in the assessments performed by industry experts in IEEE 1413.1.

Major organizations in manufacturing superpowers in electronics industry such as China and Korea are using PoF methods in product development and realization. Government of Korea has established centers of excellence to develop and implement PoF-based reliability methods in collaboration with its industrial giants. There is demand for training and implementation assistance for realization of PoF tools in emerging economies such as Turkey, India, and Vietnam.

Summary

To ensure product reliability, an organization must follow certain best practices during the product development process. These practices impact reliability through the selection of materials, product

design, manufacturing, assembly, shipping and handling, operation, maintenance, and repair.

PoF is the preferred methodology to assess product reliability, because it provides the information to rank potential failures, make design trade-offs, plan screens and tests, and conduct prognostics. Currently, the PoF methodology is widely accepted in industry, and professional societies, including the IEEE, EIA/JEDEC, and SEMATECH, have been developing standards to incorporate PoF in the product development and evaluation process.

End Notes

^aIn some cases, the environmental factors experienced by constituents of the product begin before manufacturing e.g., storage of parts (material) far in advance of their use in manufacturing.

^bJEDEC standards, JESD659-A, "Failure-mechanism-driven reliability monitoring", JEP143A, "Solid-state reliability assessment and qualification methodologies", JEDEC, JEP150, "Stress-test-driven qualification of and failure mechanisms associated with assembled solid-state surface-mount components", JESD74, "Early life failure rate calculation procedure for electronic components", JESD94, "Application-specific qualification using knowledge-based test methodology", JESD91A, "Method for developing acceleration models for electronic component failure mechanisms", and SEMATECH standards, #00053955A-XFR, "Semiconductor device reliability failure models", #00053958A-XFR, "Knowledge-based reliability qualification testing of silicon devices", #04034510A-TR, "Comparing the effectiveness of stress-based reliability qualification stress conditions", #99083810A-XFR, "Use condition-based reliability evaluation of new semiconductor technologies".

References

- [1] Pecht, M., Agarwal, R., McCluskey, P., Dishongh, T., Javadpour, S. & Mahajan, R. (1999). *Electronic Packaging Materials and Their Properties*, CRC Press, Boca Raton.
- [2] Pecht, M., Nguyen, L. & Hakim, E. (1995). *Plastic Encapsulated Microelectronics: Materials, Processes, Quality, Reliability, and Applications*, John Wiley & Sons, New York.
- [3] Osterman, M. & Stadterman, T. (1999). *High Performance Printed Circuit Boards*, McGraw-Hill, New York.

- [4] Snook, I., Marshall, J. & Newman, R. (2003). Physics of failure as an integrated part of design for reliability, *Proceedings of the IEEE Annual Reliability and Maintainability Symposium*, Tampa, January 27–30, pp. 46–54.
- [5] Osterman, M., Dasgupta, A. & Han, B. (2006). A strain range based model for life assessment of Pb-free SAC solder interconnects, *Proceedings of the 56th Electronic Component and Technology Conference*, San Diego, May 30–June 2, pp. 884–890.
- [6] Clement, J.J. (2001). Electromigration modeling for integrated circuit interconnect reliability analysis, *IEEE Transactions on Device and Materials Reliability* **1**(1), 33–42.
- [7] Lee, J.C., Chen, I.C. & Hu, C. (1988). Modeling and characterization of gate oxide reliability, *IEEE Transactions on Electron Device* **35**(22), 2268–2278.
- [8] Rogers, K. & Pecht, M. (2006). A variant of conductive filament formation failures in PWBs with 3 and 4 mil spacings, *Circuit World* **32**(3), 11–18.
- [9] Pecht, M., Radojic, R. & Rao, G. (1999). *Guidebook for Managing Silicon Chip Reliability*, CRC Press, Boca Raton.
- [10] Lall, P., Pecht, M. & Hakim, E. (1997). *The Influence of Temperature on Microelectronic Device Reliability*, CRC Press, Boca Raton.
- [11] Li, J. & Dasgupta, A. (1994). Failure mechanism models for material aging due to inter-diffusion, *IEEE Transactions on Reliability* **43**(1), 2–10.
- [12] Cunningham, J., Valentin, R., Hillman, C., Dasgupta, A. & Osterman, M. (2001). A demonstration of virtual qualification for the design of electronic hardware, *Proceedings of the Institute of Environmental Sciences and Technology Meeting*, Phoenix, April 24.
- [13] Cushing, M., Mortin, D., Stadterman, T. & Malhotra, A. (1993). Comparison of electronics-reliability assessment approaches, *IEEE Transactions on Reliability* **42**(4), 542–546.
- [14] Larson, T. & Newel, J. (1997). Test philosophies for the new millennium, *Journal of the Institute of Environmental Sciences* **40**(3), 22–27.
- [15] Caruso, H. & Dasgupta, A. (1998). A fundamental overview of analytical accelerated testing models, *Journal of the Institute of Environmental Sciences* **41**(1), 16–30.
- [16] Hu, J., Barker, D., Dasgupta, A. & Arora, A. (1993). The role of failure mechanism identification in accelerated testing, *Journal of the Institute of Environmental Sciences* **36**(4), 39–45.
- [17] McCluskey, P., Pecht, M. & Azarm, S. (1997). Reducing time-to-market using virtual qualification, *Proceedings of the Institute of Environmental Sciences Conference*, Mount Prospect, pp. 148–152.
- [18] Military Standard (2006). Microcircuits, test method standard, MIL-STD 883G, U.S. Department of Defense.
- [19] Ramakrishnan, A. & Pecht, M. (2003). A life consumption monitoring methodology for electronic systems, *IEEE Transactions on Components and Packaging Technologies* **26**(3), 625–634.
- [20] Pecht, M. & Dasgupta, A. (1995). Physics-of-failure: an approach to reliable product development, *Journal of the Institute of Environmental Sciences* **38**(5), 30–34.
- [21] Osterman, M. & Stadterman, T. (1999). Failure assessment software for circuit card assemblies, *Proceedings of the IEEE Annual Reliability and Maintainability Symposium*, Washington, DC, January 18–21, pp. 269–276.
- [22] Stadterman, T., Hum, B., Deckert, M. & Mortin, D. (1995). Use of physics of failure for reliability assessment in support of commercial circuit card use in military systems, *Proceedings of the Institute of Environmental Sciences Symposium*, Chicago, pp. 57–62.
- [23] Upadhyayula, K. & Dasgupta, A. (1998). Physics-of-failure guideline for accelerated qualification of electronic systems, *Quality and Reliability Engineering International* **14**(6), 433–447.
- [24] Mencinger, N.P. (2000). A mechanism-based methodology for processor package reliability assessments, *Intel Technology Journal* **Q3**, 1–8.
- [25] IEEE Standard 1413–1998 (1998). IEEE standard methodology for reliability prediction and assessment for electronic systems and equipment, IEEE.
- [26] IEEE Standard 1413.1–2002 (2003). IEEE guide for selecting and using reliability predictions based on IEEE 1413, IEEE.

WEIQIANG WANG, DIGANTA DAS, MICHAEL
OSTERMAN AND MICHAEL PECHT

Reliability of Consumer Goods with “Fast Turn Around”

Trends

In the past decades, the context of new product development in the consumer goods industry has changed. A study on the trends impacting reliability of technical systems has identified four major trends [1]:

- *Increasingly complex products*

Technology and functionalities become available at a higher pace, and at lower prices. Owing to these lower prices, the technologies and functionalities are introduced in products for the mass market faster. These lower prices for functions also lead to combining more functionality into one single product, e.g., mobile phones today are not just a phone, but they are also a camera, an MP3 player, an agenda, etc. As a result the products become increasingly complex.

- *Strong pressure on time-to-market and fast adoption cycles*

Competition is fierce, and to be able to have a reasonable market share and profit, companies must be the first (or at least one of the first few) to hit the market with new products. Being late often results in competition from cheaper products with similar functionality or equally priced products with increased functionality. At the same time consumers adopt the products faster. For example., it took over 15 years for the VCR to reach commodity price levels and become a standard item in most households, whereas this only took around three years for DVD recorders [2]. Consumer goods with “fast turn around” refer to this type of highly innovative products with even shortened economical life cycle (*see Longevity Risk and Life Annuities*).

- *Increasingly global economy*

The increasingly global economy manifests itself twofold. On the one hand, through the penetration of Internet consumers are much more aware of new products, functions, and features and expect those to be available at their local stores immediately.

Companies can no longer introduce products region by region, thus allowing for local adaptations to meet the varying needs of consumers in the regions. On the other hand, companies also need to utilize the capabilities of subcontractors all over the world to utilize local capabilities in order to achieve optimal product costs. Value chains are disintegrating and development activities are outsourced globally.

- *Decreasing tolerance of consumers for quality and reliability problems*

Although most people do understand the service a product offers, many do not understand (and therefore do not accept) the complexity of the underlying systems. People see, for example, a mobile phone as the functional equivalent of an old wired phone without realizing that complex systems are required to allow the use of these products. At the same time there is a trend of extended warranty periods (*see Imprecise Reliability*) and higher coverage, which allows consumers to return products for a wider range of reasons [3]. For example, in the United States, many retailers have a “no questions asked” return policy and in the European Union consumer rights are increasing and warranty has to be granted in certain cases even outside the product specification [4].

Trying to cope with these four trends in product development is an increasing challenge, especially for highly reliable and innovative products. Many businesses in the consumer electronics industry in the western world have chosen a strategy to bring fundamentally new products to the market with a much higher frequency than before (fast turn around of products). For these products there this will result in two major types of uncertainty [5]:

- Market uncertainty (*see Dynamic Financial Analysis*): the difficulty of potential customers to articulate their needs and thus uncertainty about the market opportunities.
- Technology uncertainty (*see Reliability Optimization*): the difficulty in translating technological advancements into product features and benefits for the customer.

Modern product creation processes require methods and tools to structurally manage these uncertainties during product creation. The following research question naturally arises: To what extent can the commonly used quality and reliability methods and

tools be applied to manage market and technology uncertainty in modern product creation processes?

In the following section, a number of commonly used quality and reliability methods and tools are reviewed on their ability to deal with especially these uncertainties in new product development projects.

Applicability of Quality and Reliability Methods and Tools

Traditional quality and reliability management (*see Reliability Growth Testing*) approaches, starting with the philosophies of Deming [6], Juran [7], and Crosby [8], and building upon those by many authors since then (e.g., Besterfield [9], Chase [10], and Evans and Lindsay [11]), do start from the customer perspective: achieving customer satisfaction through identifying the customer needs and highly effective process improvements. However, these approaches often treat customer needs as stable, and they have rather lengthy processes for data collection and improvement to ensure high quality and reliability. For consumer goods with fast turn around, the consumer needs are highly uncertain, so they cannot be treated as stable and further insight needs are needed during the course of the project. As a result, methods and tools like design for Six Sigma (*see Operational Risk Development*) [12] and robust design [13], which are commonly used, do not work for these products. The same holds for methods that are commonly used in the area of manufacturing to guarantee high product quality at zero hour, such as design of experiments (DOE) and statistical process control (SPC) [14], as well as several test strategies, including accelerated tests [15]. Other methods such as quality function deployment (QFD) [16] have difficulties in dealing with the increasing technology uncertainty [17], as well as market uncertainty [18, 19]. Risk management is one of the more general tools aimed at addressing a variety of risks during new product development projects [20–23]. However, during risk assessments and failure mode and effect analysis (FMEA) (*see Human Reliability Assessment; Probabilistic Risk Assessment*) in projects with higher uncertainties, many potential problems are consciously or unconsciously ignored under strong time-to-market pressure [19]. In addition, quantitative risk assessment methods require

the deployment of the best available information to understand the effects of the identified potential risks. However, owing to the increasing market and technology uncertainties during the creation of consumer goods with fast turn around, the required information is often unavailable or incomplete. Therefore, these tools will no longer be effective for consumer goods with fast turn around. Moreover, the traditional processes for data collection about customer feedback take too much time [24, 25], and do not fit within the tight schedules of the projects in this sector.

Therefore, it can be concluded that, owing to the increasing market as well as technology uncertainties, it is impossible to make accurate prediction and to use maturity measurement techniques during product development. It is also impossible to collect sufficient appropriate data for analysis and improvement prior to market introduction under the strong pressure on time to market. As a result, the commonly known approaches are not applicable to deal with the identified uncertainties in the context of consumer goods with fast turn around.

Reliability in an Innovative Context

Reliability is defined as “the probability that a system will perform its intended function for a specific period of time under a given set of conditions [26]”. As discussed in the previous section, the commonly used quality and reliability methods and tools are incapable of dealing with uncertainties embedded in the development. Given the definition of reliability, this has a number of implications:

- To be competitive, companies introduce new technologies and features into their products early, even though these may not have been proved as mature, leading to higher technical uncertainties on the feasibility of the product specification.
- Globalization and fast adoption leads to a wider range of consumers buying and using the products in a wider variety of environments. As products get more functions, there are also more applications the products can be used in, and often more connectivity options. It is much more difficult to predict consumer expectations with all these parameters, leading to higher uncertainties on the specification and its coverage of expectations.

Product reliability is usually measured by the number of products returned by dissatisfied consumers. Currently many companies in the consumer goods industry are facing an increasing number of product returns for nontechnical reasons, as recent studies have shown [26]. These studies show that around half of the returned products are technically fully functioning according to the specification: the most important reasons for consumers to return a product are that it does not meet their expectation, or that they are not able to get it working [27]. In service centers or repair workshops the products are analyzed to establish the root cause of the problem, but as there is no technical failure, this results in a growing number of “failure not found” classifications [26].

Most companies execute tests during product development and manufacturing to find problems in the product reliability prior to market introduction. These tests are under pressure for a number of reasons. First of all, reliability tests do take time, and are often on the critical path of the project schedule. Owing to the ever decreasing time-to-market, project teams may skip tests to meet the deadline [28]. Secondly, tests do not address all potential problems with the product as project teams tend to overlook the uncertainties in the project. As a result, consumers complain on products in the market for reasons that had not been identified during the product development project. Thirdly, as the variation in consumers, products, applications, and environments are increasing, it is impossible to test all combinations of these variations within the limited time available. As a result, products are released to the market with an unknown reliability level.

In summary, under the four trends discussed earlier, companies have to deal with products entering the market with an unknown reliability level, and it can be expected that a number of products will be returned as they are not fulfilling the consumer’s expectation, although technically these products are according to the specifications.

Managing Reliability in an Innovative Context

As the commonly known approaches to ensure high reliability prior to market introduction are no longer applicable for consumer goods with fast turn around,

other approaches will have to be applied. Managing reliability in new product development projects covers the following aspects: target setting and prediction, maturity measurement during development, and improvement (prevention).

Owing to the uncertainties, it is impossible to use accurate prediction and maturity measurement techniques during new product development. Therefore, the appropriate approach is to create fast learning cycles for both technical and nontechnical risks. This means that risk-driven test strategies need to be developed, which include tests with consumers to reduce the market uncertainties, as well as functional tests to reduce the technological uncertainties. These test strategies include a series of tests, starting as early as possible (using paper concepts, simulations, or early prototypes). Owing to the limited time available in projects (due to the time-to-market pressure), it will be impossible to get large amounts of data from these tests. The tests should therefore focus on increasing the quality of the data. This can be done by collecting more information from a limited number of tests. For example, when tests are performed with consumers, the main factors influencing the reliability should be included in the tests, such as consumer types, types of environment, and types of tasks consumers will execute with the product. High contrast consumer testing (HCCT) has proved to be a powerful method to set up a limited set of tests that generates rich information [29] that can be used by product development teams to improve the design such that the problems encountered by the participants in the test can be prevented.

Example 1^a: A multinational consumer electronics manufacturer would like to introduce a new ironer to the market. The new ironer consisted of a digital user interface comparing with traditional ironer and the development time was expected to be shortened. In the early development process, the development team realized that, given the uncertain consumer use of the user interface, it was necessary to collect as early as possible consumer use information by testing a number of early prototypes with real consumers. HCCT was applied there and the development team was able to very quickly identify a number of product flaws, which was not possible to be identified using traditional quality and reliability risk assessment methods, and necessary corrective actions were implemented accordingly.

Conclusion

Companies developing consumer goods with fast turn around are faced with increased levels of uncertainty in both market and technology. Commonly available methods and tools for product quality and reliability are not applicable in this context. As a result, there is an explosion in the number of consumers returning products that are fully working according to the specifications. To prevent these complaints, companies will need to change their test strategies. These strategies should be consumer oriented and focus on a limited number of tests providing data of a higher quality, and that can be used to improve the design prior to market introduction.

End Notes

^aName and details of the manufacturer cannot be disclosed here owing to reasons of confidentiality. Full details, however, are known to the authors.

References

- [1] Brombacher, A.C. (2005). Reliability in strongly innovative products; a threat or a challenge? *Reliability Engineering and System Safety* **88**, 125.
- [2] Minderhoud, S. & Fraser, P. (2005). Shifting paradigms of development in fast and dynamic markets, *Reliability Engineering and System Safety* **88**, 127–135.
- [3] Berden, T.P.J., Brombacher, A.C. & Sander, P.C. (1999). The building bricks of product quality: an overview of some basic concepts and principles, *International Journal of Production Economics* **67**, 3–15.
- [4] The European Parliament and the Council of European Union (1999). Directive 1999/44/EC of the European Parliament and the Council of 25 May 1999 on certain aspects of the sale of consumer goods and associated guarantees. *European Parliament Official Journal* **L171**, 0012–0016.
- [5] Mullins, H.W. & Sutherland, D.J. (1998). New product development in rapidly changing markets: an exploratory study, *Journal of Product Innovation Management* **15**, 224–236.
- [6] Deming, W.E. (1986). *Out of the Crisis*, MIT Centre for Advanced Engineering, Cambridge.
- [7] Juran, J.M. & Gyra, F.M. (1980). *Quality Planning and Analysis*, 2nd Edition, McGraw-Hill, New York.
- [8] Crosby, P.B. (1979). *Quality is Free*, McGraw-Hill, New York.
- [9] Besterfield, D.H. (1998). *Quality Control*, 5th Edition, Prentice-Hall, Upper Saddle River.
- [10] Chase R.L. (ed) (1988). *Total Quality Control, an IFS Executive Briefing*, Springer-Verlag, New York.
- [11] Evans, J.R. & Lindsay, W.M. (1999). *The Management and Control of Quality*, South-Western College Publishing, Cincinnati.
- [12] Creveling C.M., Slutky J.L. & Antis Jr, D. (2003). *Design for Six Sigma in Technology and Product Development*, Upper Saddle River, USA.
- [13] Phadke, M.S. (1989). *Quality Engineering Using Robust Design*, Prentice Hall: AT&T Bell Laboratories, Englewood Cliffs.
- [14] Bhote, K.R. & Bhote, A.K. (1991). *World Class Quality, Using Design of Experiments to Make it Happen*, 2nd Edition, Amacom American Management Association, New York.
- [15] Lu, Y., Loh, H.T., Brombacher, A.C. & Den Ouden, E. (2000). Accelerated stress testing in a time-driven product development process, *International Journal of Production Economics* **67**, 17–26.
- [16] Hauser, J. & Clausing, D. (1988). The house of quality, *Harvard Business Review* **66**(3), 63–73.
- [17] Griffin, A. (1992). Evaluating QFD's use in US firms as a process for developing products, *Journal of Product Innovation Management* **9**, 171–187.
- [18] Mill, H. (1994). Enhanced quality function deployment, *World Class Design to Manufacture* **1**(3), 23–26.
- [19] Lu, Y. (2002). *Analysing reliability problems in concurrent fast product development processes*, PhD thesis, Eindhoven University of Technology and National University of Singapore.
- [20] Cooper, R.G. (2001). *Winning at New Products, Accelerating the Process from Idea to Launch*, 3rd Edition, Perseus Publishing, Cambridge.
- [21] Fine, C.H. (1998). *Clock Speed: Winning Industry Control in the Age of Temporary Advantage*, Perseus Books, New York.
- [22] Smith, P.G. (1999). Managing risks as product development schedules shrink, *Research Technology Management* **42**, 25–32.
- [23] Wheelwright, S.C. & Clark, K.B. (1992). *Revolutionising New Product Development: Quantum Leaps in Speed, Efficiency and Quality*, The Free Press, New York.
- [24] Petkova, V. (2003). *An analysis of field feedback in consumer electronics industry*, PhD thesis, Eindhoven University of Technology.
- [25] de Visser, I.M., Lu, Y. & Nagappan, G. (2006). Understanding failure severity in new product development processes of consumer electronics products., *3rd International Conference on Management of Innovation and Technology*, 21–23 June 2006, Singapore, pp. 571–575.
- [26] Lewis, E.E. (1996). *Introduction to Reliability Engineering*, 3rd Edition, John Wiley & Sons.
- [27] Den Ouden, E., Lu, Y., Sonnemans, P.J.M. & Brombacher, A.C. (2006). Quality and reliability problems from a consumer's perspective: an increasing problem overlooked by businesses? *Quality and Reliability Engineering International* **22**(7), 821–838.

- [28] Minderhoud, S. (1999). Quality and reliability in product creation: extending the traditional approach, *Quality and Reliability International* **15**(6), 417–425.
- [29] Boersma, J., Loke, G., Lu, Y., Brombacher, A.C. & Loh, H.T. (2003). Reducing product rejection via a high contrast consumer test, *European Safety and Reliability Conference*, The Netherlands, 15–18 June 2003, pp. 191–194.
- Overton, D. (2006). *No Fault Found returns cost the mobile industry \$4.5 Billion per year*, WDSGlobal white paper, at <http://www.wdsglobal.co.za/news/whitepapers/20060717/20060717.asp>, posted on 18 July.

ELKE DEN OUDEN, YUAN LU AND AARNOUT
BROMBACHER

Further Reading

Brombacher, A.C., Sander, P.C., Sonnemans, P.J.M. & Rouvroye, J.L. (2005). Managing product reliability in business processes ‘under pressure’, *Reliability Engineering and System Safety* **88**, 137–146.

Reliability of Large Systems

Many technical systems belong to the class of complex systems as a result of the large number of components they are built of and their complicated operating processes. As a rule, these are series systems composed of a large number of components (*see Reliability Optimization*). Sometimes the series systems have either components or subsystems reserved and then they become parallel-series or series-parallel reliability structures (*see Multi-state Systems; Game Theoretic Methods*). We find large series systems, for instance, in transportation of water, gas, oil, and various chemical substances through pipes. Large systems of these kinds are also used in electrical energy distribution. A city bus transportation system composed of a number of communication lines each serviced by one bus may be a model series system, if we treat it as not failed, when all its lines are able to transport passengers. If the communication lines have several buses at their disposal, we may consider it as either a parallel-series system or an “ m out of n ” system. The simplest example of a parallel system or an “ m out of n ” system may be an electrical cable composed of a number of wires, which are its basic components, whereas the transmitting electrical network may be either a parallel-series system or an “ m out of n ” series system. Large systems of these types are also used in telecommunication, in rope transportation, and in transport using belt conveyors and elevators. Rope transportation systems like port elevators and ship-rope elevators used in shipyards during ship docking are model examples of series-parallel and parallel-series systems.

In the case of large systems, the determination of the exact reliability functions of the systems leads us to complicated formulae that are often useless for reliability practitioners. One of the important techniques in this situation is the asymptotic approach to system reliability evaluation (*see Bayesian Statistics in Quantitative Risk Assessment*). In this approach, instead of the preliminary complex formula for the system reliability function, after assuming that the number of system components tends to infinity and finding the limit reliability of the system, we obtain its simplified form.

The mathematical methods used in the asymptotic approach to the system reliability analysis of large systems are based on limit theorems on order statistics distributions, considered in very wide literature, for instance, in [1–4]. These theorems have generated the investigation concerned with limit reliability functions of the systems composed of two-state components. The main and fundamental results on this subject that determine the three-element classes of limit reliability functions for homogeneous series systems and for homogeneous parallel systems have been established by Gnidenko [5]. These results are also presented, sometimes with different proofs, for instance, in subsequent works [6, 7]. The generalizations of these results for homogeneous “ m out of n ” systems have been formulated and proved by Smirnow [8], where the seven-element class of possible limit reliability functions for these systems has been fixed. As it has been done for homogeneous series and parallel systems, classes of limit reliability functions have been fixed by Chernoff and Teicher [9] for homogeneous series-parallel and parallel-series systems. Their results were concerned with so-called quadratic systems only. They have fixed limit reliability functions for the homogeneous series-parallel systems with the number of series subsystems equal to the number of components in these subsystems, and for the homogeneous parallel-series systems with the number of parallel subsystems equal to the number of components in these subsystems. The generalizations of these results for nonquadratic and nonhomogeneous series-parallel and parallel-series systems are formulated and proved in [7].

All the results so far described have been obtained under the linear normalization of the system lifetimes. This article contains exemplary results described above for homogeneous series and parallel systems and comments on newest generalizations recently presented in [10].

Reliability of Two-State Systems

We assume that

$$E_i, i = 1, 2, \dots, n, \quad n \in N \quad (1)$$

are two-state components of the system having reliability functions

$$R_i(t) = P(T_i > t), \quad t \in (-\infty, \infty) \quad (2)$$

where

$$T_i, \quad i = 1, 2, \dots, n$$

are independent random variables representing the lifetimes of components E_i with distribution functions (see **Multistate Systems; Imprecise Reliability**)

$$F_i(t) = P(T_i \leq t), \quad t \in (-\infty, \infty) \quad (3)$$

The simplest two-state reliability structures are series and parallel systems. We define these systems first.

We call a two-state system series if it is not failed, if and only if all its components are not failed. It means that the series system lifetime T is given by

$$T = \min_{1 \leq i \leq n} \{T_i\} \quad (4)$$

and therefore its reliability function is given by

$$\bar{R}_n(t) = \prod_{i=1}^n R_i(t), \quad t \in (-\infty, \infty) \quad (5)$$

The scheme of a series system is given in Figure 1.

We call a two-state system parallel if it is failed, if and only if all its components are failed. It means that the parallel system lifetime T is given by

$$T = \max_{1 \leq i \leq n} \{T_i\} \quad (6)$$

and therefore its reliability function is given by

$$R_n(t) = 1 - \prod_{i=1}^n F_i(t), \quad t \in (-\infty, \infty) \quad (7)$$

The scheme of a parallel system is given in Figure 2.

We call a two-state system homogeneous if its component lifetimes have an identical distribution function $F(t)$, i.e., if its components have the same reliability function $R(t) = 1 - F(t)$, $t \in (-\infty, \infty)$.

The above definition and equations (5)–(7) result in the simplified formulae for the reliability functions of the homogeneous systems as follows:

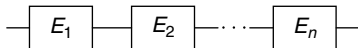


Figure 1 The scheme of a series system

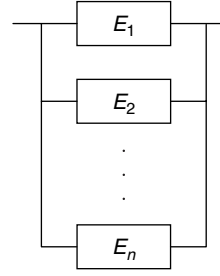


Figure 2 The scheme of a parallel system

- for a series system

$$\bar{R}_n(t) = [R(t)]^n, \quad t \in (-\infty, \infty) \quad (8)$$

- for a parallel system

$$R_n(t) = 1 - [F(t)]^n, \quad t \in (-\infty, \infty) \quad (9)$$

Asymptotic Approach to System Reliability

The asymptotic approach (see **Repair, Inspection, and Replacement Models; Extreme Values in Reliability**) to the reliability of two-state systems depends on the investigation of limit distributions of a standardized random variable

$$\frac{(T - b_n)}{a_n} \quad (10)$$

where T is the lifetime of a system and $a_n > 0$ and $b_n \in (-\infty, \infty)$ are suitably chosen numbers called *normalizing constants*.

Since

$$\begin{aligned} P((T - b_n)/a_n > t) &= P(T > a_n t + b_n) \\ &= R_n(a_n t + b_n) \end{aligned} \quad (11)$$

where $R_n(t)$ is a reliability function of a system composed of n components, the following definition becomes natural.

Definition 1 We call a reliability function $\mathfrak{R}(t)$ the limit reliability function of a system having a reliability function $R_n(t)$ if there exist normalizing constants $a_n > 0$, $b_n \in (-\infty, \infty)$ such that

$$\lim_{n \rightarrow \infty} R_n(a_n t + b_n) = \mathfrak{R}(t) \text{ for } t \in C_{\mathfrak{R}} \quad (12)$$

where $C_{\mathfrak{R}}$ is the set of continuity points of $\mathfrak{R}(t)$.

Thus, if the asymptotic reliability function $\mathfrak{R}(t)$ of a system is known, then for sufficiently large n , the approximate formula

$$\mathbf{R}_n(t) \cong \mathfrak{R}((t - b_n)/a_n), \quad t \in (-\infty, \infty) \quad (13)$$

may be used instead of the system exact reliability function $\mathbf{R}_n(t)$.

Reliability of Large Two-State Series Systems

The investigations of limit reliability functions of homogeneous two-state series systems are based on the following auxiliary theorem [5–7].

Lemma 1 *If*

- (i) $\bar{\mathfrak{R}}(t) = \exp[-\bar{V}(t)]$ is a nondegenerate reliability function,
- (ii) $\bar{\mathbf{R}}_n(t)$ is the reliability function of a homogeneous two-state series system defined by equation (8),
- (iii) $a_n > 0, b_n \in (-\infty, \infty)$,

then

$$\lim_{n \rightarrow \infty} \bar{\mathbf{R}}_n(a_n t + b_n) = \bar{\mathfrak{R}}(t) \text{ for } t \in C_{\bar{\mathfrak{R}}} \quad (14)$$

if and only if

$$\lim_{n \rightarrow \infty} nF(a_n t + b_n) = \bar{V}(t) \text{ for } t \in C_{\bar{V}} \quad (15)$$

Lemma 1 is an essential tool in finding limit reliability functions of two-state series systems. It is also the basis for fixing the class of all possible limit reliability functions of these systems. This class is determined by the following theorem [5–7].

Theorem 1 *The only nondegenerate limit reliability functions of the homogeneous two-state series system are*

$$\begin{aligned} \bar{\mathfrak{R}}_1(t) &= \exp[-(-t)^{-\alpha}] \text{ for } t < 0, \bar{\mathfrak{R}}_1(t) \\ &= 0 \text{ for } t \geq 0, \alpha > 0 \end{aligned} \quad (16)$$

$$\begin{aligned} \bar{\mathfrak{R}}_2(t) &= 1 \text{ for } t < 0, \bar{\mathfrak{R}}_2(t) \\ &= \exp[-t^\alpha] \text{ for } t \geq 0, \alpha > 0 \end{aligned} \quad (17)$$

$$\bar{\mathfrak{R}}_3(t) = \exp[-\exp[t]] \text{ for } t \in (-\infty, \infty) \quad (18)$$

Reliability of Large Two-State Parallel Systems

The class of limit reliability functions for homogeneous two-state parallel systems may be determined on the basis of the following auxiliary theorem [5–7].

Lemma 2 *If*

- (i) $\mathfrak{R}(t) = 1 - \exp[-V(t)]$ is a nondegenerate reliability function,
- (ii) $\mathbf{R}_n(t)$ is the reliability function of a homogeneous two-state parallel system defined by equation (9), $a_n > 0, b_n \in (-\infty, \infty)$,

then

$$\lim_{n \rightarrow \infty} \mathbf{R}_n(a_n t + b_n) = \mathfrak{R}(t) \text{ for } t \in C_{\mathfrak{R}} \quad (19)$$

if and only if

$$\lim_{n \rightarrow \infty} nR(a_n t + b_n) = V(t) \text{ for } t \in C_V \quad (20)$$

By applying Lemma 2, it is possible to fix the class of limit reliability functions for homogeneous two-state parallel systems. However, it is easier to obtain this result using the duality property of parallel and series systems expressed in the relationship

$$\mathbf{R}_n(t) = 1 - \bar{\mathbf{R}}_n(-t) \text{ for } t \in (-\infty, \infty) \quad (21)$$

which results in the following Lemma ([5–7, 10]).

Lemma 3 *If $\bar{\mathfrak{R}}(t)$ is the limit reliability function of a homogeneous two-state series system with reliability functions of particular components $\bar{R}(t)$, then*

$$\mathfrak{R}(t) = 1 - \bar{\mathfrak{R}}(-t) \text{ for } t \in C_{\bar{\mathfrak{R}}} \quad (22)$$

is the limit reliability function of a homogeneous two-state parallel system with reliability functions of particular components

$$R(t) = 1 - \bar{R}(-t) \text{ for } t \in C_{\bar{R}} \quad (23)$$

At the same time, if (a_n, b_n) is a pair of normalizing constants in the first case, then $(a_n, -b_n)$ is such a pair in the second case.

The application of Lemma 3 and Theorem 1 yields the following result [5–7].

Theorem 2 *The only nondegenerate limit reliability functions of the homogeneous parallel system are*

$$\begin{aligned} \mathfrak{R}_1(t) &= 1 \text{ for } t \leq 0, \mathfrak{R}_1(t) \\ &= 1 - \exp[-t^{-\alpha}] \text{ for } t > 0, \alpha > 0 \end{aligned} \quad (24)$$

$$\begin{aligned} \mathfrak{R}_2(t) &= 1 - \exp[-(-t)^\alpha] \text{ for } t < 0, \mathfrak{R}_2(t) \\ &= 0 \text{ for } t \geq 0, \alpha > 0 \end{aligned} \quad (25)$$

$$\mathfrak{R}_3(t) = 1 - \exp[-\exp[-t]] \text{ for } t \in (-\infty, \infty) \quad (26)$$

Using Lemma 2, it is possible to prove the following fact [10].

Corollary 1 *If components of the homogeneous two-state parallel system have Weibull reliability functions*

$$R(t) = 1 \text{ for } t < 0, R(t) = \exp[-\beta t^\alpha] \text{ for } t \geq 0, \alpha > 0, \beta > 0 \quad (27)$$

and

$$a_n = b_n/(\alpha \log n), b_n = (\log n/\beta)^{1/\alpha} \quad (28)$$

then

$$\mathfrak{R}_3(t) = 1 - \exp[-\exp[-t]], t \in (-\infty, \infty) \quad (29)$$

is its limit reliability function.

Example 1 (A steel rope, durability) Let us consider a steel rope composed of 36 strands used in ship-rope elevator and assume that it has not failed if at least one of its strands is not broken. Under this assumption we may consider the rope as a homogeneous parallel system composed of $n = 36$ basic components. The cross section of the considered rope is presented in Figure 3.

Further, assuming that the strands have Weibull reliability functions with parameters

$$\alpha = 2, \beta = (7.07)^{-6} \quad (30)$$

by equation (9), the rope's exact reliability function takes the form

$$R_{36}(t) = 1 \text{ for } t < 0, R_{36}(t) = 1 - [1 - \exp[-(7.07)^{-6}t^2]]^{36} \text{ for } t \geq 0 \quad (31)$$

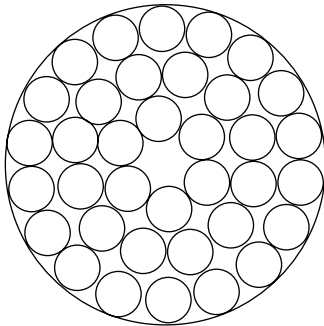


Figure 3 The cross-section of the steel rope

Thus, according to Corollary 1, assuming

$$a_n = (7.07)^3/(2\sqrt{\log 36}), b_n = (7.07)^3\sqrt{\log 36} \quad (32)$$

and applying equation (13), we arrive at the approximate formula for the rope reliability function of the form

$$R_{36}(t) \cong \mathfrak{R}_3((t - b_n)/a_n) = 1 - \exp[-\exp[-0.01071t + 7.167]] \text{ for } t \in (-\infty, \infty) \quad (33)$$

The mean value of the rope lifetime T and its standard deviation, in months, calculated on the basis of the above approximate result and according to the formulae

$$E[T] = Ca_n + b_n, \sigma = \pi a_n/\sqrt{6} \quad (34)$$

where $C \cong 0.5772$ is Euler's constant, are, respectively,

$$E[T] \cong 723, \sigma \cong 120 \quad (35)$$

The values of the exact and approximate reliability functions of the rope are presented in Table 1 and graphically in Figure 4. The differences between them are not large, which means that the mistakes in replacing the exact rope reliability function by its approximate form are practically not significant.

Table 1 The values of the exact and approximate reliability functions of the steel rope

t	$R_{36}(t)$	$\mathfrak{R}_3\left(\frac{t-b_n}{a_n}\right)$	$\Delta = R_{36} - \mathfrak{R}_3$
0	1.000	1.000	0.000
400	1.000	1.000	0.000
500	0.995	0.988	-0.003
550	0.965	0.972	-0.007
600	0.874	0.877	-0.003
650	0.712	0.707	0.005
700	0.513	0.513	0.000
750	0.330	0.344	-0.014
800	0.193	0.218	-0.025
900	0.053	0.081	-0.028
1000	0.012	0.029	-0.017
1100	0.002	0.010	-0.008
1200	0.000	0.003	-0.003

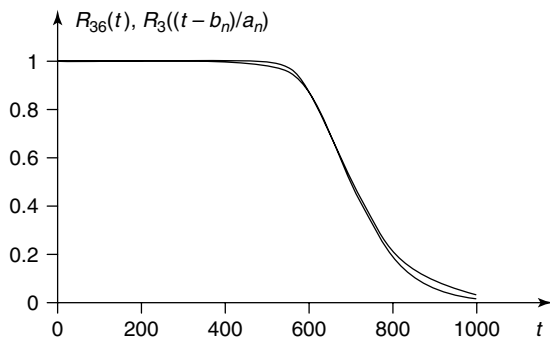


Figure 4 The graphs of the exact and approximate reliability functions of the steel rope

Conclusions

Generalizations of the results on limit reliability functions of two-state homogeneous systems for these and other two-state systems, in case they are nonhomogeneous, are mostly given in [7, 10]. More general and practically important complex systems composed of multistate and degrading in time components are considered in wide literature, for instance, in [11]. An especially important role that they play in the evaluation of technical systems reliability and safety and their operating process effectiveness is described in [10] for large multistate systems with degrading components. The most important results regarding generalizations of the results on limit reliability functions of two-state systems dependent on transferring them to series, parallel, “ m out of n ”, series-parallel and parallel-series multistate systems with degrading components are given in [10]. Some practical applications of the asymptotic approach to the reliability evaluation of various technical systems are available in [10] as well.

The results concerned with the asymptotic approach to system reliability analysis have become the basis for the investigation concerned with domains of attraction for the limit reliability functions of the considered systems. In a natural way, they have led to investigation of the speed of convergence of the system reliability function sequences to their limit

reliability functions. These results have also initiated the investigation of limit reliability functions of “ m out of n ”-series, series-“ m out of n ” systems, systems with hierarchical reliability structures and systems varying in time, their structures and components, reliability functions, as well as investigations on the problems of the system reliability improvement and optimization described briefly in [10].

References

- [1] Fisher, R.A. & Tippett, L.H.C. (1928). Limiting forms of the frequency distribution of the largest and smallest member of a sample, *Proceedings of Cambridge Philosophical Society* **24**, 180–190.
- [2] Frechet, M. (1927). Sur la loi de probabilité de l'écart maximum, *Annals de la Société Polonaise de Mathématique* **6**, 93–116.
- [3] Gumbel, E.J. (1962). *Statistics of Extremes*, Columbia University Press, New York.
- [4] Von Mises, R. (1936). La distribution de la plus grande de n valeurs, *Revue Mathématique de l'Union Interbalkanique* **1**, 141–160.
- [5] Gnienenko, B.W. (1943). Sur la distribution limite du terme maximum d'une série aléatoire, *Annals of Mathematics* **44**, 432–453.
- [6] Barlow, R.E. & Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing: Probability Models*, Holt Rinehart & Winston, New York.
- [7] Kołowrocki, K. (1993). *On a Class of Limit Reliability Functions for Series-Parallel and Parallel-Series Systems*, Monograph, Maritime University Press, Gdynia.
- [8] Smirnow, N.W. (1949). *Predielnyje Zakony Raspredelenija dla Czlenow Wariacjonnoego Riada*, Trudy Matem. Inst. im. W.A. Stieklowa.
- [9] Chernoff, H. & Teicher, H. (1965). Limit distributions of the minimax of independent identically distributed random variables, *Proceedings of American Mathematical Society* **116**, 474–491.
- [10] Kołowrocki, K. (2004). *Reliability of Large Systems*, Elsevier, Amsterdam–Boston–Heidelberg–London–New York–Oxford–Paris–San Diego–San Francisco–Singapore–Sydney–Tokyo.
- [11] Xue, J. & Yang, K. (1995). Dynamic reliability analysis of coherent multi-state systems, *IEEE Transactions on Reliability* **4**(44), 683–688.

KRZYSZTOF KOŁOWROCKI

Reliability Optimization

The design of new products involves the specification of performance requirements, the evaluation and selection, or possibly development, of components to perform clearly defined functions and the determination of a system-level architecture (*see Systems Reliability*). Once a decision has been made to produce a new product, detailed engineering specifications are determined to prescribe minimum levels of reliability, maximum weight, maximum volume, etc. If the design is to be produced economically or within some specified budget, numerous design alternatives must be considered, resulting in a difficult optimization problem.

Reliability Optimization Problems

There are different classes of reliability optimization problems depending on the selected objective function and decision variables. For single objective problems, the primary types of problems are redundancy allocation problem, reliability allocation problem, and reliability–redundancy allocation problem. Kuo *et al.* [1] provides an excellent overview and summary of reliability optimization research.

Redundancy Allocation Problem

The most widely studied reliability optimization problem is the redundancy allocation problem (*see Multistate Systems*). For systems designed using off-the-shelf component types, with fixed cost, reliability, and weight characteristics, system design becomes a combinatorial optimization problem. For each required function identified within the design process, there are often many different component types available with different levels of cost, reliability, weight, and other properties. The problem is, then, to select the optimal combination of parts and levels of redundancy to collectively meet reliability, weight, and other constraints at a minimum cost, or alternatively, to maximize reliability, given constraints on various system properties.

For this problem, there are m_i discrete component choices available for each subsystem ($i = 1, \dots, s$). Figure 1 presents a typical example of a series system with k -out-of- n subsystem reliabilities. If k is one for

each subsystem, then this is a series–parallel system. For each subsystem, a minimum of k_i components must be chosen from among the m_i available choices (with an unlimited supply available for each of the m_i components). It may become necessary to place additional components in parallel to achieve the reliability requirement. Additionally, it may be advantageous to use more levels of redundancy ($> k_i$) of a lower reliability component as an alternative to using a more reliable, but more costly component. As a result, there is a large number of possible solutions even for relatively small problems.

The redundancy allocation problem for a series–parallel system, as depicted in Figure 1, can be formulated to maximize reliability (equation (1)) or to minimize cost (equation (2)), given various constraints. Typical formulations are as follows:

$$\max \quad R(\mathbf{x}) = \prod_{i=1}^s R_i(\mathbf{x}_i | k_i) \quad (1)$$

$$\begin{aligned} \text{subject to:} \quad & \sum_{i=1}^s C_i(\mathbf{x}_i) \leq C_o \\ & \sum_{i=1}^s W_i(\mathbf{x}_i) \leq W_o \\ & k_i \leq \sum_{j=1}^{m_i} x_{ij} \leq n_{\max,i} \quad \forall i = 1, 2, \dots, s \\ & x_{ij} \in \{0, 1, 2, \dots\} \end{aligned}$$

$$\min \quad \sum_{i=1}^s C_i(\mathbf{x}_i) \quad (2)$$

$$\begin{aligned} \text{subject to:} \quad & \prod_{i=1}^s R_i(\mathbf{x}_i | k_i) \geq R_o \\ & \sum_{i=1}^s W_i(\mathbf{x}_i) \leq W_o \\ & k_i \leq \sum_{j=1}^{m_i} x_{ij} \leq n_{\max,i} \quad \forall i = 1, 2, \dots, s \\ & x_{ij} \in \{0, 1, 2, \dots\} \end{aligned}$$

where

R_o : minimum system reliability

C_o : maximum system cost

W_o : maximum system weight

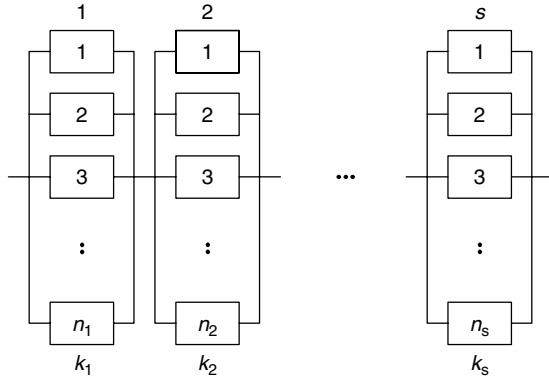


Figure 1 Series-parallel system

s : number of subsystems

$\mathbf{x}_i$: $(x_{i1}, x_{i2}, \dots, x_{i,m_i})$

x_{ij} : number of the j th component used in subsystem i

$n_{max,i}$: maximum number of components used in subsystem i (specified)

$R_i(\mathbf{x}_i|k_i)$: reliability of subsystem i , given k_i

k_i : minimum number of components required for subsystem i to operate

$C_i(\mathbf{x}_i)$: total cost of subsystem i

$W_i(\mathbf{x}_i)$: total weight of subsystem i .

The redundancy allocation problem for series-parallel systems has proved to be difficult. Chern [2] showed that the problem is NP hard. Many different optimization approaches have been used to determine optimal or “very good” solutions. Previous research involving this problem can be classified on the basis of the solution approach as (a) dynamic programming (see **Repair, Inspection, and Replacement Models**), (b) integer programming, (c) mixed integer and nonlinear programming, or (c) metaheuristics.

Bellman [3] and Bellman and Dreyfuss [4, 5] showed that an optimal solution to the redundancy allocation problem could be found using dynamic programming if there was only one component choice for each subsystem. In the original Bellman formulation, the objective was to maximize reliability given a single cost constraint. For each subsystem, the problem was then to identify the optimal levels of redundancy. A well-known disadvantage of dynamic programming formulations is the inability to handle multiple constraints efficiently.

Fyffe *et al.* [6] also used a dynamic programming approach and solved a more difficult design problem.

They considered a system with 14 subsystems and constraints on both cost and weight. For each subsystem in the Fyffe formulation, there are three or four different component choices, each with different reliability, cost, and weight. To accommodate multiple constraints within a dynamic programming formulation, a Lagrangian multiplier is used for the weight constraint within the objective function. The equivalent dynamic programming formulation of the problem is as follows. Their approach to constraint satisfaction uses Lagrangian relaxation. The objective function incorporates the original objective function (system reliability), the constrained system weight and λ , a Lagrangian multiplier.

$$\begin{aligned} \max \quad & \left[\prod_{i=1}^s R_i(\mathbf{x}_i|k_i = 1) \right] e^{-\lambda \sum_{i=1}^s W_i(\mathbf{x}_i)} \\ \text{subject to} \quad & \sum_{i=1}^s C_i(\mathbf{x}_i) \leq C_0 \end{aligned} \quad (3)$$

The dynamic programming recursion formulas are as follows, where the final solution is found at $f_s(C_0)$ for the optimal value of λ .

$$f_1(\alpha) = \max_{\mathbf{x}_1: C_1(\mathbf{x}_1) \leq \alpha} \{ R_1(\mathbf{x}_1|k_1) e^{-\lambda W(\mathbf{x}_1)} \} \quad (4)$$

$$\begin{aligned} f_i(\beta) = \max_{\mathbf{x}_i: \sum C_i(\mathbf{x}_i) \leq \beta} \\ \times \{ R_i(\mathbf{x}_i|k_i) e^{-\lambda W(\mathbf{x}_i)} f_{i-1}(\beta - C_i(\mathbf{x}_i)) \} \\ \text{for } i = 2, \dots, s \end{aligned} \quad (5)$$

To use the Fyffe *et al.* [6] dynamic programming algorithm, it is necessary to assume a Lagrangian multiplier value (λ) and then successively solve recursion formulas starting with the function $f_1(\alpha)$, $f_2(\beta)$ and continuing on until $f_{s-1}(\beta)$. Then, $f_s(C_0)$ is the optimal solution for that particular λ . If $\sum W_i(\mathbf{x}_i) = W_0$ or $\sum C_i(\mathbf{x}_i) = C_0$, and $\sum W_i(\mathbf{x}_i) \leq W_0$, then this is the final solution. If not, then λ must be incrementally changed, and the recursion formulas repeated, until one of the equality conditions is met and the final solution is found. In practice, the equality constraints may not be exactly met because of the integer constraints on the number of components. In this case, it is not always possible to determine the optimal solution.

Nakagawa and Miyazaki [7] showed that the use of a Lagrangian multiplier with dynamic programming can be inefficient. Instead of using Lagrangian

multipliers, they deployed a surrogate constraint combining the cost and weight constraints into one. It was then necessary to solve a series of problem iterations with different surrogate multipliers. A heuristic was presented to successively update the surrogate multipliers. Additionally, a stopping criteria was provided to identify cases when the Nakagawa and Miyazaki (N&M) algorithm would not lead to an optimal solution. The algorithm was demonstrated by solving 33 variations of the Fyffe problem with different weight constraints ($W_0 = 159-191$). Of the 33 problems, the N&M algorithm produced optimal solutions to 30 of these problems. In the other cases, the final solution was not feasible, although feasible solutions to the problem do exist. In general, the N&M algorithm worked best when one of the two constraints had slack.

Many variations of the problem can be transformed into equivalent integer programming problems. This was originally demonstrated and proved by Ghare and Taylor [8], who used a branch-and-bound approach to solve randomly generated redundancy allocation problems with 30 subsystems with 15 constraints, and 99 subsystems with 10 constraints. In all cases, Ghare and Taylor assumed that there was only one component choice for each subsystem.

Bulfin and Liu [9] also used an integer programming approach to the redundancy allocation problem. One heuristic and two exact algorithms were proposed, depending on the problem structure. They formulated the problem as a knapsack problem (KP) and used a surrogate constraints approach analogous to that of Nakagawa and Miyazaki [7]. The surrogate multipliers were approximated as the optimal Lagrangian multipliers as found by subgradient optimization [10]. Bulfin and Liu determined solutions to all 33 problems investigated by Nakagawa and Miyazaki, as well as other examples, with good results.

Other examples of integer programming solutions to the redundancy allocation problem were presented by Misra and Sharma [11], Gen *et al.* [12, 13]. Misra and Sharma presented a very fast and useful algorithm to solve integer programming problems formulated like that of Ghare and Taylor [8]. Gen *et al.* formulated the problem as a multiobjective decision-making problem with distinct goals for reliability, cost, and weight.

The application of metaheuristics in recent years has resulted in impressive results. Coit and Smith [14]

applied genetic algorithms (GA). Their work and that of Levitin *et al.* [15] were among the first papers that considered component mixing for redundancy allocation problem (RAP). Component mixing involves the use of functionally equivalent components within the same subsystem to provide redundancy (e.g., a mechanical gyroscope providing redundancy for an electronic gyroscope in an airplane). Later, Kulturel-Konak *et al.* [16] presented the application of Tabu Search, and Liang and Smith [17] presented the Ant Colony Optimization. The advantage of metaheuristics is that they usually yield good results in a relatively short time. The drawbacks are that they cannot guarantee global optimality, not always yield repeatable results, and generally require customized versions for particular problems.

Most of the RAP solution methodologies only allow active redundancy. Coit [18] presented an integer programming approach to solve the problem with cold-standby redundancy. Later, Coit [19] proposed an approach that allowed for a choice of active or cold-standby redundancy.

Reliability Allocation Problem

The reliability allocation problem is fundamentally different. For this problem, the system structure is fixed and the component reliability values are bounded continuous decision variables. For this problem, there is no general restriction on the system structure. Figure 2 shows an example of a general system. Component cost and other parameters are generally expressed as mathematical functions of the component reliability. Increasing the component reliability (and thus, the system reliability) can therefore increase the cost, weight, and other factors that should be considered when designing a system.

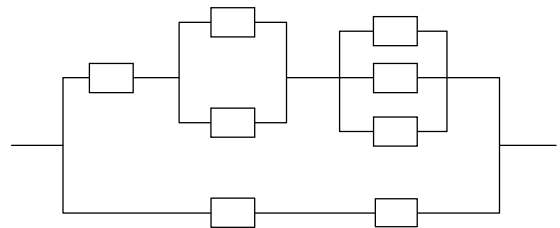


Figure 2 Example of a general system

For the reliability allocation problem, the general mathematical formulation can be represented as

$$\begin{aligned} & \max_{\mathbf{r}} R(\mathbf{r}) \\ & \text{subject to:} \\ & g_p(\mathbf{r}) \leq b_p : \forall p \in \{1, 2, \dots\} \\ & r_i^l \leq r_i \leq r_i^u : \forall i \in \{1, 2, \dots\} \end{aligned} \quad (6)$$

where

$R(\mathbf{r})$: system reliability as a function of $\mathbf{r}$

$\mathbf{r}$: (r_1, r_2, r_3)

r_i : reliability of the component i

r_i^l, r_i^u : lower and upper bound for r_i

$g_p(\mathbf{r})$: p th system performance measure (cost, weight, etc.)

b_p : p th constraint limit (cost budget, maximum allowable weight, etc.).

Since the only decision variables are continuous, forms of nonlinear programming can be used to determine solutions [1]. To accommodate the constraints, Lagrangian multipliers, or penalty, or barrier functions are used. Successful solutions to this problem were developed by Hwang *et al.* [20] on the basis of sequential unconstrained minimization techniques (SUMT) and Kuo *et al.* [21] who used the branch-and-bound method.

Reliability–Redundancy Allocation Problem

The reliability–redundancy allocation problem is the most general problem formulation. The system is composed of one or more “subsystems”. Each subsystem has x_i components in parallel with reliability of r_i and the decision is how to optimally allocate redundancy and reliability to the components of each subsystem so that the overall reliability of the system is maximized. In general, subsystems can be in any arrangement, as depicted in Figure 2.

Increasing the reliability and redundancy levels can increase the cost, weight, and other factors that should be considered when designing a system. The general mathematical formulation can be represented as

$$\begin{aligned} & \max_{\{\mathbf{r}, \mathbf{x}\}} R(\mathbf{r}, \mathbf{x}) \\ & \text{subject to} \\ & g_p(\mathbf{r}, \mathbf{x}) \leq b_p : \forall p \in \{1, 2, \dots\} \\ & l_i \leq x_i \leq u_i : \forall i \in \{1, 2, \dots, s\} \\ & r_i^l \leq r_i \leq r_i^u : \forall i \in \{1, 2, \dots, s\} \\ & x_{ij} \in \{1, 2, \dots, s\} \end{aligned} \quad (7)$$

Tillman *et al.* [22] were among the first to solve the reliability–redundancy allocation problem (RRAP). Their method, known as *THK* (Tillman, Hwang & Kuo), is based on Hooke–Jeeves (H–J) pattern search [23]. Gopal *et al.* [24] developed a heuristic method called *GAG2* (Gopal, Aggarwal & Gupta-2) which adjusts the component reliability of one subsystem in each stage. Having the component reliability for each subsystem, the *GAG* (Gopal, Aggarwal & Gupta) approach [25] was applied to compute subsystem redundancy. Xu *et al.* [26] developed a heuristic which is known as *XXL* (Xu, Kuo & Liu). The algorithm finds the component redundancy by using a heuristic which is combined with an analytic approach to find the optimal reliability level.

Hsieh *et al.* [27] represented a GA approach for RRAP. Hikita *et al.* [28] applied a surrogate constraint method for RRAP in specific structures, including series–parallel and parallel–series systems.

Multistate System Reliability Optimization

More recently, reliability optimization algorithms have been applied for multistate systems (MSS) (*see Imprecise Reliability*). For these systems, component and/or system behavior exhibit varying performance, as opposed to traditional reliability where a component or system is either completely working or completely failed. Maximization of MSS requires efficient methods to estimate reliability. There are generally four methods for MSS reliability assessment: (a) the structure function approach (Brunelle and Kapur [29], Pourret *et al.* [30]), (b) the stochastic processes “Markov” approach (Xue and Yang [31]), (c) Monte Carlo simulation technique (*see Risk Measures and Economic Capital for (Re)insurers; Uncertainty Analysis and Dependence Modeling*) (Ramirez-Marquez and Coit [32]), and (d) the universal moment generating function approach (Levitin and Lisnianski [33], Ushakov [34, 35]).

Research for MSS reliability optimization considering one objective and several constraints have been presented recently. Ramirez-Marquez and Coit [36] proposed a heuristic to solve a multistate series–parallel system with binary capacitated components. In their study, the problem is formulated with the objective of minimizing the total cost associated with a system design constrained by a reliability performance index. Levitin *et al.* [37] used a GA for

solving the multistate RAP, where the system and its components have a range of performance levels. On the basis of the universal moment generating function (UMGF), they determined the system availability. Levitin [38] addressed the multistage expansion problem for multistate series–parallel systems. In this problem, the system-study period is divided into several stages. Later, Levitin [39] solved a redundancy optimization problem for multistate systems with fixed resource requirements and unreliable sources, subject to availability constraints.

Multiple-Objective System Reliability Optimization

Many system design initiatives are truly multiobjective. In addition to the maximization of system reliability, it is generally desirable to consider the minimization of other objectives such as cost, weight, etc. In such situations, the designer faces the problem of optimizing all objectives simultaneously.

As an example, consider a design problem where the objective is to determine the optimal design configuration that maximizes system reliability, minimizes the total cost and minimizes the system weight. The mathematical formulation of the problem is given below:

$$\left\{ \begin{array}{l} \max \left[R = \prod_{i=1}^s R_i(\mathbf{x}_i | k_i) \right], \\ \min \left[C = \sum_{i=1}^s \sum_{j=1}^{m_i} c_{ij} x_{ij} \right], \\ \min \left[W = \sum_{i=1}^s \sum_{j=1}^{m_i} w_{ij} x_{ij} \right] \end{array} \right\} \quad (8)$$

subject to

$$1 \leq \sum_{j=1}^{m_i} x_{ij} \leq n_{\max,i} \quad \forall i = 1, 2, \dots, s$$

$$x_{ij} \in \{0, 1, 2, \dots\}$$

where

c_{ij}, w_{ij} : cost and weight for the j th available component for subsystem i .

Multiple-objective versions of system reliability optimization problems have also been studied. Sakawa [40] considered a series system with

four objectives: maximization of system reliability, minimization of cost, weight, and volume. Sakawa proposed a method to derive Pareto-optimal solutions by optimizing the composite objective functions that were obtained as linear combinations of the four objective functions. The Lagrangian function for each composite problem was decomposed into parts and optimized by applying both the dual decomposition method and the surrogate worth trade-off method.

Dhingra [41] used a combination of goal programming and the goal attainment method to generate Pareto-optimal solutions in a four-stage series system to maximize system reliability and minimize cost, weight, and volume. Kulturel-Konak *et al.* [42] solved the problem using Tabu Search method with three objective functions: maximize reliability, minimize the cost of the system, and minimize the weight of the system. Coit *et al.* [43] proposed a multicriteria formulation to maximize the system reliability estimate and minimize the associated variance.

Multicriteria formulations using GAs can also be found. Busacca *et al.* [44] proposed a multiobjective GA approach that was applied to a safety system of a nuclear power plant. Huang *et al.* [45] proposed a new method of system reliability multiobjective optimization using GAs. They considered a reliability optimization model which contains two objective functions, simultaneously maximizing system reliability and minimizing system cost, subject to limits on weight and volume.

Recently, Taboada and Coit [46] and Taboada *et al.* [47] formulated the redundancy allocation problem as a multiobjective problem with the system reliability to be maximized, and cost and weight of the system to be minimized. The Pareto-optimal set was initially obtained using the fast elitist nondominated sorting genetic algorithm (NSGA-II) originally proposed by Deb *et al.* [48]. Then, the decision-making stage was performed by applying two proposed pruning methods to reduce the size of the Pareto-optimal set and obtain a smaller representation of the multiobjective design space.

References

- [1] Kuo, W., Prasad, V., Tillman, F. & Hwang, C.L. (2000). *Optimal Reliability Design: Fundamentals and Applications*, Cambridge University Press.

- [2] Chern, M.S. (1992). On the computational complexity of reliability redundancy allocation in a series system, *Operations Research Letters* **11**, 309–315.
- [3] Bellman, R.E. (1957). *Dynamic Programming*, Princeton University Press, Princeton.
- [4] Bellman, R.E. & Dreyfuss, E. (1958). Dynamic programming and reliability of multicomponent devices, *Operations Research* **6**, 200–206.
- [5] Bellman, R.E. & Dreyfuss, E. (1962). *Applied Dynamic Programming*, Princeton University Press, Princeton.
- [6] Fyffe, D.E., Hines, W.W. & Lee, N.K. (1968). System reliability allocation problem and a computational algorithm, *IEEE Transaction on Reliability* **17**, 64–69.
- [7] Nakagawa, Y. & Miyazaki, S. (1981). Surrogate constraints algorithm for reliability optimization problems with two constraints, *IEEE Transaction on Reliability* **30**, 175–180.
- [8] Ghare, P.M. & Taylor, R.E. (1969). Optimal redundancy for reliability in series system, *Operations Research* **17**, 838–847.
- [9] Bulfin, R.L. & Liu, C.Y. (1985). Optimal allocation of redundant components for large systems, *IEEE Transactions on Reliability* **R-34**(3), 241–247.
- [10] Fisher, M. (1981). The Lagrangian relaxation method for solving integer programming problems, *Management Science* **27**, 1–18.
- [11] Misra, K.B. & Sharma, U. (1991). An efficient algorithm to solve integer programming problems arising in system reliability design, *IEEE Transactions on Reliability* **R-40**, 81–91.
- [12] Gen, M., Ida, K., Tsujimura, Y. & Kim, C.E. (1993). Large-scale 0–1 fuzzy goal programming and its application to reliability optimization problem, *Computers and Industrial Engineering* **24**, 539–549.
- [13] Gen, M., Ida, K. & Lee, J.U. (1990). A computational algorithm for solving 0–1 goal programming with GUB structures and its application for optimization problems in system reliability, *Electronics and Communications in Japan, Part 3* **73**, 88–96.
- [14] Coit, D.W. & Smith, A.E. (1996). Reliability optimization for series-parallel systems using a genetic algorithm, *IEEE Transactions on Reliability* **45**(4), 895–904.
- [15] Levitin, G., Lisnianski, A. & Elmakis, D. (1997). Structure optimization of power system with different redundant elements, *Electric Power Systems Research* **43**(1), 19–27.
- [16] Kulturel-Konak, S., Smith, A.E. & Coit, D.W. (2003). Efficiently solving the redundancy allocation problem using Tabu search, *IIE Transactions* **35**(6), 515–526.
- [17] Liang, Y.C. & Smith, A.E. (2004). An ant colony optimization algorithm for the redundancy allocation problem (RAP), *IEEE Transactions on Reliability* **53**(3), 417–423.
- [18] Coit, D. (2001). Cold-standby redundancy optimization for nonrepairable systems, *IIE Transactions* **33**(6), 471–478.
- [19] Coit, D. (2003). Maximization of system reliability with a choice of redundancy strategies, *IIE Transactions* **35**(6), 535–544.
- [20] Hwang, C., Lai, K., Tillman, F. & Fan, T. (1975). Optimization of system reliability by the sequential unconstrained minimization technique, *IEEE Transactions on Reliability* **R-24**, 133–135.
- [21] Kuo, W., Lin, H., Xu, Z. & Zhang, W. (1987). Reliability optimization with the Lagrange multiplier and branch-and-bound technique, *IEEE Transactions on Reliability* **R-36**, 624–630.
- [22] Tillman, F.A., Hwang, C.L. & Kuo, W. (1977). Determining component reliability and redundancy for optimum system reliability, *IEEE Transaction on Reliability* **R-26**(3), 162–165.
- [23] Hooke, R. & Jeeves, T.A. (1961). Direct search solution of numerical and statistical problems, *Journal of the Association of Computing Machinery* **8**(2), 212–229.
- [24] Gopal, K., Aggarwal, K.K. & Gupta, J.S. (1980). A new method for solving reliability optimization problem, *IEEE Transactions on Reliability* **R-29**, 36–38.
- [25] Gopal, K., Aggarwal, K.K. & Gupta, J.S. (1978). An improved algorithm for reliability optimization, *IEEE Transactions on Reliability* **R-27**, 325–328.
- [26] Xu, Z., Kuo, W. & Lin, H. (1990). Optimization limits in improving system reliability, *IEEE Transactions on Reliability* **39**, 51–60.
- [27] Hsieh, Y.C., Chen, T.C. & Bricker, D.L. (1998). Genetic algorithms for reliability design problems, *Microelectronics Reliability* **38**(10), 1599–1605.
- [28] Hikita, M., Nakagawa, Y., Nakashima, K. & Narihisa, H. (1992). Reliability optimization of systems by a surrogate-constraints algorithm, *IEEE Transactions on Reliability* **41**(3), 473–480.
- [29] Brunelle, R.D. & Kapur, K.C. (1998). Continuous-state system-reliability: an interpolation approach, *IEEE Transactions on Reliability* **47**(2), 181–187.
- [30] Pourret, O., Collet, J. & Bon, J.-L. (1999). Evaluation of the unavailability of a multistate-component system using a binary model, *Reliability Engineering and System Safety* **64**(1), 13–17.
- [31] Xue, J. & Yang, K. (1995). Dynamic reliability analysis of coherent multistate systems, *IEEE Transactions on Reliability* **44**(4), 683–688.
- [32] Ramirez-Marquez, J. & Coit, D. (2005). A Monte-Carlo simulation approach for approximating multistate two-terminal reliability, *Reliability Engineering and System Safety* **87**(2), 253–264.
- [33] Levitin, G. & Lisnianski, A. (2001). A new approach to solving problems of multi-state system reliability optimization, *Quality and Reliability Engineering International* **17**(2), 93–104.
- [34] Ushakov, I. (1986). A universal generating function, *Soviet Journal of Computer and System Sciences* **24**(5), 37–49.
- [35] Ushakov, I. (1988). Reliability analysis of multistate systems by means of a modified generating function, *Journal of Information Process* **24**(3), 131–135.

-
- [36] Ramirez-Marquez, J. & Coit, D. (2004). A heuristic for solving the redundancy allocation problem for multistate series-parallel systems, *Reliability Engineering and System Safety* **83**(3), 341–349.
 - [37] Levitin, G., Lisnianski, A., Ben-Haim, H. & Elmakis, D. (1998). Redundancy optimization for series-parallel multi-state systems, *IEEE Transactions on Reliability* **47**(2), 165–172.
 - [38] Levitin, G. (2000). Multi-state series-parallel system expansion-scheduling subject to availability constraints, *IEEE Transactions on Reliability* **49**(1), 71–79.
 - [39] Levitin, G. (2001). Redundancy optimization for multi-state system with fixed resource-requirements and unreliable sources, *IEEE Transactions on Reliability* **50**(1), 52–59.
 - [40] Sakawa, M. (1981). Optimal reliability-design of a series-parallel system by a large-scale multiobjective optimization method, *IEEE Transactions on Reliability* **30**, 173–174.
 - [41] Dhingra, A.K. (1992). Optimal apportionment of reliability and redundancy in series systems under multiple objectives, *IEEE Transactions on Reliability* **41**(4), 576–582.
 - [42] Kulturel-Konak, S., Coit, D.W. & Baheranwala, F. (2005). *Pruned Pareto-Optimal Sets for the System Redundancy Allocation Problem Based on Multiple Prioritized Objectives*, Rutgers University, ISE Working Paper 05–009.
 - [43] Coit, D., Jin, T. & Wattanapongsakorn, N. (2004). System optimization considering component reliability estimation uncertainty: a multi-criteria approach, *IEEE Transactions on Reliability* **53**(3), 369–380.
 - [44] Busacca, P.G., Marseguer, M. & Zio, E. (2001). Multi-objective optimization by genetic algorithms: application to safety systems, *Reliability Engineering and System Safety* **72**, 59–74.
 - [45] Huang, H.-Z., Qu, J. & Zuo, M.J. (2006). A new method of system reliability multi-objective optimization using genetic algorithms, *Proceedings of the Reliability and Maintainability Symposium (RAMS)*, Newport Beach, CA.
 - [46] Taboada, H. & Coit, D.W. (2007). Data clustering of solutions for multiple objective system reliability optimization problems, *Quality Technology and Quantitative Management Journal* **4**(2), 191–210.
 - [47] Taboada, H., Baheranwala, F., Coit, D.W. & Wattanapongsakorn, N. (2006). Practical solutions of multi-objective optimization: an application to system reliability design problems, *Reliability Engineering and System Safety* **91**(9), 992–1007.
 - [48] Deb, K., Agrawal, S., Pratap, A. & Meyarivan, T. (2002). A fast and elitist multi-objective genetic algorithm: NSGA-II, *IEEE Transactions on Evolutionary Computation* **6**(2), 182–197.

DAVID W. COIT

Remote Sensing

Remote sensing can help inform risk assessment in several regards. It is especially advantageous when used in conjunction with geographic information systems (GISs), both of which support spatial specificity in scenarios of risk assessment (*see Spatial Risk Assessment*). Several realms of risk-relevant information can be addressed at least in part through these tandem technologies [1, 2].

One such aspect of risk-related information pertains to *exposure* to the agent(s) of concern. This type of information is indicative of the likelihood that a unit of interest will come under the influence of the agent(s) or stressor(s) being assessed. A second informational aspect pertains to the *sensitivity* or vulnerability of the unit(s) that may potentially be subject to the agent(s) of concern. This type of information is indicative of the degree of influence expected when a unit of interest is subjected to an agent of concern, as a function of level of exposure. A third informational aspect of interest concerns *condition* of units that are potentially involved in exposure to the impacting agent(s). This may be expressed variously in terms of value, health, functionality, diversity, impairment, etc. A fourth factor of analytical assessment addresses *resiliency* or propensity to recover from adverse influence after being subjected to an agent of concern. There are undoubtedly other aspects of assessment that can be aided by remote sensing, but these should be sufficient to induce investigative interest [3].

Sensors, Signals, and Synthesis

The traditional technology of remote sensing receives radiant signals *via* special sensors placed in perspective position at advantageous altitudes and then structures synoptic image information as digital data [4–6]. Systems can be configured to provide low spatial resolution over large areas, or high spatial resolution over smaller areas.

Standard signals are selected sections of the electromagnetic energy spectrum spanning visible and infrared wavelengths. Simultaneously sensed signals from several spectral sectors become *bands* of image information. It makes a substantial difference what spectral sections are selected. Figure 1

is a relatively low-resolution but expansive Landsat Thematic Mapper (TM) band spanning wavelengths of 0.63–0.69 μm in the red region of the spectrum for the risk-relevant context of forest fires and their after effects in China. This is a spectral section in which green plants absorb energy for photosynthesis, thus darkening the image in heavily vegetated areas. The large lighter patches in the lower-left quadrant represent areas where vegetation has been extensively impacted by large forest fires. This may be compared with Figure 2, which shows the thermal band (10.4–12.5 μm) of the same digital data that differentiates primarily between certain cold clouds, cool forests, and warmer exposed land surface areas.

Pseudosignals

Sometimes an individual spectral band is directly informative regarding conditions of interest, and sometimes it is not. Quite often an environmental condition of interest will express itself differently in particular bands, with the combination of expressions being fairly definitive [7, 8]. This is the case with green vegetation, which absorbs red radiation to support photosynthesis but strongly reflects infrared radiation in wavelengths just beyond the visible range. In this case, using some mathematical formulation for the difference between infrared and red that is normalized by the total of red and infrared will be indicative of the presence or absence of vegetation. Such formulations are conventionally called *vegetation indicators* or *vegetation indices*, often abbreviated as VIN. Figure 3 shows an image of one such formulation for the China fire digital image data.

It can be observed that the image of the vegetation indicator signal is somewhat darkened in the areas affected by the fires, indicating lesser cover of green vegetation. Likewise, the areas of clouds and the river show an absence of vegetation.

There is considerable scope for innovative formulation of such pseudosignals for incorporation into a multiband digital image dataset. Since the signal data are in digital form, there are no intrinsic constraints to such statistical synthesis. Owing to the typically large size of digital image datasets, however, special parsimonious data formats are used. Special software systems for image analysis are needed to access and analyze such data. Therefore, a partnership between remote sensing image analyst and risk assessment

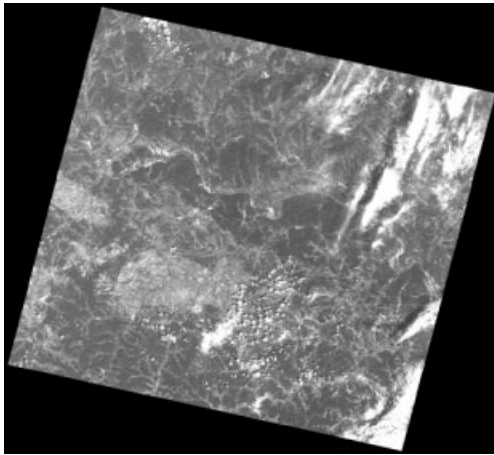


Figure 1 Landsat TM red band image of areas in China affected by large forest fires

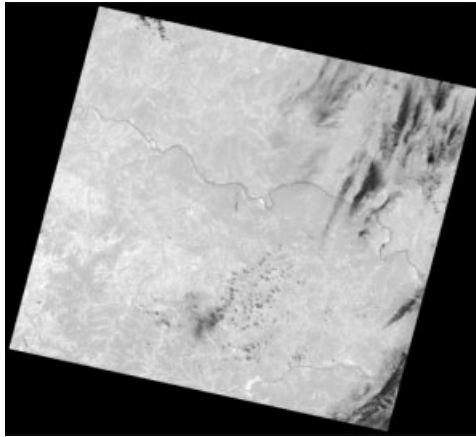


Figure 2 Landsat TM thermal image of areas in China affected by large forest fires

analyst is advantageous. It may also be noted that pseudosignals can also be synthesized from nonspectral sources such as digital elevation data to augment the analyses.

Compositional Classification

One of the standard statistical scenarios entails multivariate methods of categorical classification to produce maps of compositional components such as land cover classes. Land cover classification is routinely relevant to risk, since it is very likely to affect

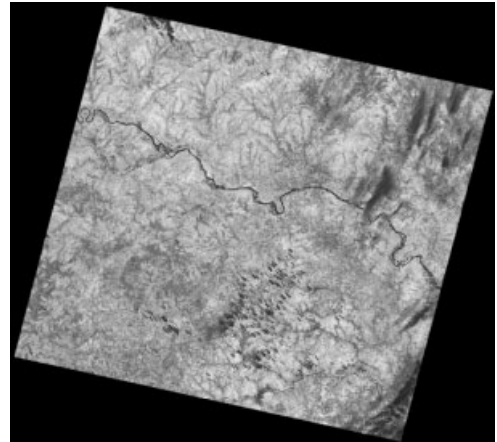


Figure 3 Vegetation indicator pseudosignal for scene of China forest fire

exposure, sensitivity, condition, and resiliency. There are two major modes of mapping for such purposes, with multiple minor modifications on each [9].

One of the major modes of classificatory mapping is known in the image analysis jargon as *supervised classification*. This incorporates the intent of discriminant analysis by designating samples of each category as *training sets* to serve as prototypes for statistical pattern recognition processing. The supervision thus consists of designating the target classes in terms of training sets, with the mapping legend being predetermined.

The second major mode of classificatory mapping is called *unsupervised classification*, which is considerably more empirical since it does not require advance selection of class prototypes. This is initiated by conducting a statistical cluster analysis on the image data using one of a large variety of clustering strategies. A category name is then assigned to each of the resulting clusters on the basis of either existing information for some areas or data collected for the specific purpose of labeling clusters as the need arises. It is also common to refine the level of categorical detail using ancillary map information available in a GIS.

Modeling Spatial Interaction of Signals

As discussed with regard to pseudosignals, it is often the joint variation or interaction of different signal bands that is particularly informative with

respect to landscape patterns comprising the scene. The unsupervised approach to scene composition can be considerably extended to achieve modeling of the patterns of signal interaction at different levels of detail. This can be accomplished by compound statistical segmentation of a scene as described by Myers and Patil [10] to obtain *polypatterns*.

Polypatterns are derived by a four-stage segmentation sequence, which produces bi-level polypatterns that are contained in two bytes per image element (picture element = pixel). Since multiband digital image datasets normally use at least one byte per band, the polypatterns also accomplish image compression or informational condensation to occupy less computer memory resources. The aggregated A level of patterns is arranged as ordered overtones that can be rendered directly as a gray-tone image. The finer B level must be decompressed to be usable with generic image analysis software. The polypatterns serve as a model of the multiband image by enabling an approximated version of the original image dataset to be generated. The fidelity of the approximation can be checked and improved if desired. The A-level model contains 250 segments that can be analyzed in the manner of tabulated multivariate statistical data. The opportunities for statistical analysis are thus considerably expanded, since the original image data only support analysis on an element-by-element basis with elements typically numbering in the millions.

An image of the A-level model for the China fire context is shown in Figure 4. This model draws information from all bands and allows the major

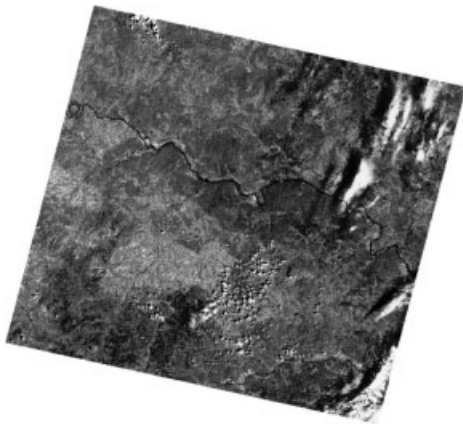


Figure 4 Image overtones of A-level polypattern model for China forest fire context

variants of band interaction to be located. This image model captures the major patterns of joint variation for the image bands using only one byte of computer storage capacity per image element. In effect, this is a shaded map of segment locations with the shades being ordered in the same manner as the overall intensities of the segments. A particular advantage of such an A-level model is that it can be used in a GIS in the manner of a raster (cellular) map.

The nature of the four stages of development provides a bit more insight into the polypatterns. The first stage takes multiband image data and produces 250 pattern prototypes without specifying where the patterns occur in the scene. The second stage scans the image data relative to the pattern prototypes and produces a provisional pattern map. The third stage partially disaggregates the provisional patterns by successive splitting to obtain the base B-level model of patterns. The fourth stage (re)groups the B-level patterns into A-level aggregated patterns of the type shown in Figure 4.

Similar polypattern processing can be conducted on pseudosignals in the same manner as for spectral signals. This lends considerable flexibility to the analysis of spatial patterns for risk assessment.

Components of Change

The patterns and progress of landscape changes are very important in relation to risk assessment, and this is another area in which remote sensing is relevant. It is fairly intuitive that repeating images over time with the same sensor and examining differences in the respective bands should be revealing with respect to patterns of change. A straightforward extension from single-band difference to multiband difference is obtained by Euclidean distance, which is the square root of the sum of squared differences for the respective bands. This is equal to the distance moved by an image element in spectral space between imaging occasions, and is therefore usually called *change vector* analysis. Spectral space (or spectral feature space) consists of band values as axes on a graph having as many dimensions as there are bands. This reduces to the single-band difference if there is only one band, and is similar to the quadratic mean change in lending emphasis to the bands having the larger changes. There are likewise several other ways in which indicators of spectral change could be

formulated, as for example, using Manhattan distance instead of Euclidean distance. Manhattan distance is effectively the sum of the band differences, and therefore does not give disproportionate emphasis to large changes in band values *versus* smaller changes.

It is also possible to make what might be called a *multi-indicator change image* by including several different indicators as if they were bands in a multiband digital image. For example, this might include differences for the respective bands along with a change vector. Multi-indicator change images tend to be complicated to interpret by conventional image analysis, but lend themselves nicely to modeling spatial interaction of signals as outlined above. The change interaction model is again much more parsimonious than its parent change image dataset, and can be interpreted at the segment level. This idea of image modeling for multiband composites can likewise be extended to multitemporal investigations involving more than two imaging occasions. Change indicators can be compiled for each successive pair of occasions, and then composited in the manner of multiband digital data for modeling. Patterns in interaction obtained in modeling thus reveal patterns of temporal change.

The polypattern approach to image modeling confers still further advantage for change analysis in a rather different manner. Spatial, as opposed to spectral, matching of A-level patterns can be done between imaging occasions. Each pattern in one date thus has a counterpart pattern in the other date. Inconsistencies in occurrences of counterparts between dates are indicative of change. This approach does not presume that the bands recorded by the sensors on the two occasions are exactly the same.

Color Coding

It is worthy to note the important role of tricolor coding images in remote sensing. Human vision obtains sense of color as mixtures of red, green, and blue light. This allows color renditions to be obtained

from three image bands by coloring one in shades of red, one in shades of green, and one in shades of blue (RGB). Tricolor composites are essential to take full advantage of remote sensing imagery for visual interpretation. Unfortunately, color images are also expensive to publish and have been considered an unaffordable luxury for this writing.

References

- [1] Bolstad, P. (2005). *GIS Fundamentals – A First Text on Geographic Information Systems*, Eider Press, White Bear Lake, p. 411.
- [2] Chrisman, N. (2002). *Exploring Geographic Information Systems*, John Wiley & Sons, New York, p. 305.
- [3] Walsh, S. & Crews-Meyer, K. (2002). *Remote Sensing and GIS Applications for Linking People, Place and Policy*, Kluwer Academic Publishers, Boston.
- [4] Cracknell, A. & Hayes, L. (1991). *Introduction to Remote Sensing*, Taylor & Francis, New York, p. 303.
- [5] Gibson, P. & Power, C. (2000). *Introductory Remote Sensing: Principles and Practices*, Taylor & Francis, New York, p. 208.
- [6] Lillesand, T. & Kiefer, R. (1999). *Remote Sensing and Image Interpretation*, 3rd Edition, John Wiley & Sons, New York, p. 750.
- [7] Frohn, R. (1998). *Remote Sensing for Landscape Ecology: New Metric Indicators for Monitoring, Modeling and Assessment of Ecosystems*, Lewis Publishers, Boca Raton, p. 99.
- [8] Jensen, J. (2000). *Remote Sensing of the Environment: An Earth Resource Perspective*, Prentice-Hall, Upper Saddle River, p. 544.
- [9] Gonzales, R. & Woods, R. (2002). *Digital Image Processing*, 2nd Edition, Prentice-Hall, Upper Saddle River, p. 544.
- [10] Myers, W. & Patil, G.P. (2006). *Pattern-Based Compression of Multi-Band Image Data for Landscape Analysis*, Springer, New York, p. 186.

Related Articles

Hotspot Geoinformatics

WAYNE MYERS AND GANAPATHI P. PATIL

Repair, Inspection, and Replacement Models

Maintenance of stochastically deteriorating systems (see **Reliability Growth Testing**) has received much attention, as systems used in production of goods and delivery of services constitute the vast majority of most industries' capital. An optimal maintenance policy (see **Role of Risk Communication in a Comprehensive Risk Management Approach**) is important to prevent the occurrence of system failure, improve system reliability, and reduce maintenance costs. A maintenance policy is a function whose range is the set of possible maintenance actions, the action space, and whose domain is the set of possible realizations in the system's history, the information space. Repair, inspection, and replacement are the main elements of the action space. An optimal maintenance policy results from optimization of a mathematical maintenance model in which both costs and benefits of maintenance are quantified. Results from early developments of maintenance models are summarized by McCall [1] and Barlow and Proschan [2].

Although the industry recognized the importance of maintenance, it took some time before it adopted a scientific approach toward maintenance. The reasons might be that the effects of poor maintenance only become manifest in the medium to long term, and that the direct benefit of effective maintenance is difficult to measure: "the true value of maintenance is given in terms of events that do not occur, which makes the assessment of the value of these measures at best speculative" [3].

Over the years, the complexity of maintenance models (see **Role of Risk Communication in a Comprehensive Risk Management Approach; Structural Reliability; Maintenance Modeling and Management**) has increased in response to the more complex maintenance problems in real life. Many of the important early maintenance models, such as age and block-replacement models possessed an elegance and simplicity that led to easily implementable policies. These models were later generalized to describe real situations more accurately. Some of the new results are complex and require powerful computers for implementation. However, to appreciate these complex models, understanding of the

basic models is required. Before introducing some basic maintenance models for single-unit and multiunit systems in the sections "Single-Unit Maintenance Models" and "Multiunit Maintenance Models", we discuss the general characteristics of maintenance models in the section "Model Characteristics". For more information about generalizations of these basic models, we refer to excellent overview papers in the literature [1, 4–10].

Model Characteristics

We discuss some important general characteristics of maintenance models.

States of the System

Consider a system consisting of $n(\geq 1)$ units and denote by $X_i(t)$, $i = 1, \dots, n$, $t \geq 0$, the state of the i th unit at time t . Then the state of the system at time t is given by the random vector $\mathbf{X}(t) = (X_1(t), X_2(t), \dots, X_n(t))$. The state space Ω_i of the i th unit is the set of all its possible states, and the state space of the system is $\Omega = \prod_{i=1}^n \Omega_i$. Maintenance policies are conditioned on the state of the system as they describe the action that should be taken for each $\omega \in \Omega$. Several quantities can describe the state of a unit, e.g., deterioration level, age, or number of spare units, and similar quantities can describe the system.

Available Actions

In general, three main actions are considered in maintenance models: repair, replace, or do nothing. Which action should be taken depends on several factors, such as the state of the system at the time of action, the objective function, and the time horizon. If the state of the system is not known, then inspection of the system is also a possible action. McCall [1] termed maintenance models where the state of the system subject to stochastic failure is always known with certainty *preventive* maintenance models, and models where the state of the system subject to stochastic failure is unknown unless either inspection or replacement is carried out *preparedness* maintenance models.

Several different levels of repair, replacement, and inspection characterize different maintenance models. For example, if repair or replacement of a failed

unit restores function to the system, but the system's failure rate remains as it was just before failure, then the action is called *minimal* repair. If there is economies of scale in the repair or replacement activity, and the repair or replacement of a failing unit provides an opportunity for preventive maintenance of other units, it is called *opportunistic* repair or replacement. The system can be *continuously* monitored, in which case the state of the system is known at every moment, or inspection can only take place at certain *discrete* times. These actions are explained in greater detail in the sections titled "Single-Unit Maintenance Models" and "Multiunit Maintenance Models".

Time Horizon

It is important to make a distinction between maintenance models with finite and infinite time horizons, the latter have received more attention in the literature as the corresponding optimization problems tend to be easier. Optimal policies are obtained by optimization of an objective function, possibly subject to a set of constraints. In case of an infinite time horizon, the optimal policy is *stationary* (i.e., constant over time). For a finite time horizon, the optimal policy is recalculated at each replacement or repair time by optimizing the objective function over the remaining time. Such *sequential* policies appear to be most important in industries where equipment is subject to rapid technological change. The end point of a finite horizon can be fixed or random, the latter, e.g., if it depends on a finite number of spare parts [11].

Knowledge of the System

On the basis of the knowledge of the system, one can make a distinction between maintenance models with complete and partial/incomplete information, e.g., on the system's failure time distribution, its current state, or its cost structure [8]. If the current state of the system is unknown, inspection is required, which can be perfect or imperfect. McCall [1] suggests four different ways of dealing with incomplete information: (a) minimax maintenance policies if one has virtually no information, (b) maintenance policies based on bounding techniques if one possesses some information regarding the stochastic behavior of the system, e.g., the expected value of the failure time or that its failure rate is increasing, (c) Bayes

adaptive models (*see Bayesian Statistics in Quantitative Risk Assessment*) if one has some relevant subjective information and statistical information will accrue over time, and (d) estimated parametric maintenance models if one has detailed knowledge of the failure distribution and enough data to estimate the parameters. Maintenance models with incomplete information have not received much attention in the literature. Recently, adaptive nonparametric predictive maintenance models have been presented, where uncertainty about the failure time of a future unit is quantified on the basis of its exchangeability with n observed failure times, and where optimal policies adapt to available process information [12, 13].

Model Types

Maintenance models can be classified according to *discrete* or *continuous* time being used. In the first case, system monitoring and actions only take place at discrete time points. Another possible distinction considers the deterioration process of the system, leading to *deterministic* and *stochastic* maintenance models. We restrict attention to stochastically deteriorating systems. Maintenance models can also be classified with regard to the type of system considered, e.g., *single-unit* or *multiunit* systems, and whether or not multiple units are dependent. Interactions between units can be classified into three different types [6]:

1. **Economic dependency**
The cost structure of replacement and maintenance has dependencies between the units. Either costs can be saved when several units are jointly maintained (economies of scale), or the opposite happens if simultaneous downtime of the units is undesirable and hence maintenance must be spread out as much as possible.
2. **Structural dependency**
Some working units have to be replaced or dismantled to replace or repair failed units.
3. **Probabilistic dependency**
The state of one unit influences the failure time distribution of other units, or common-cause failures occur that bring about simultaneous failures and hence correlate the lifetimes.

If all units in a system are economically, structurally, and stochastically independent, maintenance policies for single-unit systems can be applied to

the multiunit system. But if any units in a system are dependent upon each other, an optimal repair or replacement decision for one unit is not necessarily optimal for the whole system, so interactions cannot be neglected in maintenance policies. The section “Single-Unit Maintenance Models” presents maintenance models for single-unit systems, and the section “Multiunit Maintenance Models”, for multiunit systems.

Optimality Criteria

Optimal maintenance policies (*see Economic Criteria for Setting Environmental Standards*) obviously depend on the chosen objective function. When considering an infinite time horizon, common criteria are minimization of long-run expected (discounted) costs or long-run expected downtime, or maximization of long-run expected availability, all per unit time. If the process considered does not change over the infinite horizon and the replacement units all have the same failure time distribution, then the renewal reward theorem can be applied. This implies, e.g., that the long-run expected costs per unit time are equal to the expected costs per cycle divided by the expected cycle length, where a cycle is the time between consecutive renewals. In practice, even though such a long period for the same process is not realistic, the renewal reward optimality criterion is often considered attractive and reasonable, partly because of its mathematical simplicity.

During the last decade [14–16], there has been increasing interest in adaptive replacement strategies (*see Imprecise Reliability*), where instead of assuming complete knowledge of a system’s failure time distribution, one uses information from the process to update the distribution used for deriving the optimal replacement strategy for the next system. Such an adaptive approach undermines the use of the renewal reward criterion, as one explicitly does not use the same replacement strategy over a very long period of time. In [14, 17] the use of a “one-cycle” optimality criterion is proposed, aiming at minimal expected costs per unit time for the single cycle corresponding to the system considered. Comparisons of the renewal reward and one-cycle criteria are given in [18, 19].

When considering a finite time horizon, optimal sequential maintenance policies can be obtained, e.g., by minimizing the expected total (discounted) costs or downtime of the system over the whole time

horizon. Unfortunately, analytic identification and practical implementation of such a sequential policy tends to be prohibitively complex. Hence, asymptotic approximations (e.g., by using the renewal reward criterion) are more frequently employed.

Methods of Solutions

The primary optimization techniques used to derive optimal policies include (non) linear programming, dynamic programming (*see Role of Alternative Assets in Portfolio Construction; Reliability Optimization; Optimal Stopping and Dynamic Programming*), Markov decision methods, decision analysis techniques, search techniques, and heuristic approaches. Policies that require detailed descriptions of each possible realization of the system or that have complicated forms are often analytically difficult and costly to implement. Therefore, attention is normally confined to classes of policies that possess a simple form, which often determines the appropriate optimization techniques.

Single-Unit Maintenance Models

We consider maintenance models (*see Reliability Integrated Engineering Using Physics of Failure; Quantitative Reliability Assessment of Electricity Supply*) for systems consisting of only a single unit: repair models, inspection models, shock models, and data-based models. We present the foundations of the models and mention some generalizations; see [1, 4, 7, 8, 10] for overviews, including more generalizations. We should emphasize that not all maintenance models can be classified into one of these categories.

Repair Models

The basic period preventive maintenance model was introduced by Derman [20] and extended by Klein [21] (see also [2]). Consider a unit that is observed at discrete time points, and let $X(n)$ be the deterioration level of the unit at time n . The state space of the unit is the set $\Omega = \{0, 1, \dots, L\}$, where state 0 denotes a new or completely renovated unit and state L , an inoperative or failed unit. It is assumed that (a) the unit deteriorates stochastically according to a time-homogeneous Markov chain with

one-step transition probabilities $p_{ij} = P\{X(n+1) = j | X(n) = i\}$, $i, j \in \Omega$, $n = 0, 1, \dots$; (b) without replacement or repair, the unit does not return to state 0, so $p_{i0} = 0$ for $i \in \Omega \setminus \{L\}$; (c) the unit must be replaced upon failure, i.e., $p_{L0} = 1$; (d) the unit will eventually fail, the underlying Markov chain $\{X(n), n = 0, 1, \dots\}$ has a single ergodic class, and the steady-state distribution exists, i.e., there is an n such that $p_{iL}(n) > 0$ for each $i \in \Omega \setminus \{L\}$, where $p_{ij}(n) = P\{X(n+m) = j | X(m) = i\}$ are the n -step transition probabilities.

The unit can be replaced preventively, at cost c_p , to avoid failure or further deterioration, whereas replacement upon failure costs $c_f > c_p$. Let $N_p(t)$ and $N_f(t)$ be the number of preventive and corrective replacements during $[0, t)$, respectively. The objective is to minimize the long-run expected costs per unit time,

$$\lim_{t \rightarrow \infty} \frac{c_p E[N_p(t)] + c_f E[N_f(t)]}{t} \quad (1)$$

Replacement decisions are summarized by a set of decision rules based on the entire history of the process. Derman [22] established conditions that enable restriction to stationary nonrandomized rules, showing that if $\sum_{j=k}^L p_{ij}$ is nondecreasing in $i \in \Omega \setminus \{L\}$, for each $k \in \Omega$, then a *control-limit* rule is optimal, i.e., there is a state $i^* \in \Omega$ such that the unit should be replaced if and only if the observed state is $k \geq i^*$. The additional condition implies that the Markov chain is increasing in failure rate (IFR), i.e., without replacements the probability of deterioration increases as the initial state increases.

This basic model has received much attention in the literature and many generalizations, with respect to, for example, the state space and cost structure, have been presented. See the review papers mentioned earlier for overviews of generalizations.

One of the best known repair models with maintenance not restricted to a set of discrete time points is the *age-replacement model* (see **Imprecise Reliability; Reliability Data**) [2], where a unit is replaced upon failure or upon reaching the age T , whichever occurs first. Let F be the *known* cumulative distribution function (cdf) of the unit's failure time, f its probability density function (pdf), then the hazard rate of the unit at age x is $h(x) = f(x)/(1 - F(x))$, where $h(x) dx$ can be interpreted as the probability that a unit of age x will fail in the small interval $(x, x + dx]$.

When considering an infinite time horizon, Barlow and Proschan [2] proved that the optimal age-replacement policy is nonrandom if the underlying failure time distribution is continuous. Application of the renewal reward theorem implies that the long-run expected costs per time unit for replacement strategy $T > 0$ are equal to the expected costs per cycle divided by the expected cycle length, where a cycle is the time between two consecutive replacements, so equation (1) can be written as

$$C(T) = \frac{c_p(1 - F(T)) + c_f F(T)}{\int_0^T (1 - F(x)) dx} \quad (2)$$

Let the optimal strategy be denoted by T^* (which might be infinite), then the first-order necessary condition for optimality of a finite T^* [2] is

$$h(T) \int_0^T (1 - F(x)) dx - F(T) = \frac{c_p}{c_f - c_p} \quad (3)$$

For monotonically strictly increasing $h(T)$, Barlow and Proschan [2] proved that if $h(\infty) > c_f/((c_f - c_p)E[X])$, with $E[X]$ the first moment of F , then equation (3) has a unique finite solution T^* , else it has no finite solution, and the unit should only be replaced on failure.

Example 1 (Age-replacement model) Suppose that the failure time of a unit has a Weibull distribution with shape parameter 2 and scale parameter 1 (notation: $W(2,1)$), so $f(t) = 2t e^{-t^2}$, $t \geq 0$ and $h(t) = 2t$, and let $c_p = 1$ and $c_f = 10$. We would like to find the optimal age-replacement strategy T^* that minimizes the long-run expected costs per unit time as given in equation (2). As the hazard rate $h(t)$ is monotonically strictly increasing in t and $h(\infty) > c_f/((c_f - c_p)E[X])$, it follows that equation (3) has a unique finite solution $T^* = 0.3365$, with $C(T^*) = 6.0561$. If no preventive replacement were carried out, the long-run expected costs per unit time would be $C(\infty) = 11.2838$.

When considering a finite time horizon, the problem becomes much more difficult. If F is continuous, then for any finite time horizon, there exists a minimum cost age-replacement policy [2], but no general analytic formula is available for its calculation. In [2] it is shown how to calculate the optimal policy for a particular case.

Many variations and generalizations of the age-replacement model have been presented, some involving refined cost functions. Nakagawa and Osaki [23] considered age replacement for a two-unit redundant system, where one of the identical units is in operation and the other is in standby and immune to failure or aging effects. Dekker and Dijkstra [24] considered opportunity-based age replacement, with preventive replacement at the first opportunity after reaching a predetermined threshold age.

The class of *minimal repair models* (see **Repairable Systems Reliability**) is also important. There are many instances in which complex systems with several components are regarded as single units for maintenance purposes. However, the performance of complex systems depends on the individual components. At system failure, a decision is required on whether to replace the system or to repair (replace) the failed component and reset the system to operation. If a repair or replacement of the failed component restores function to the entire system but the system's failure rate remains as it was just before failure, then it is called *minimal repair*. Since the failure rate of most complex systems increases with age, it would become increasingly expensive to maintain operation by minimal repair. The question is then, when is it optimal to replace the entire system instead of performing minimal repair?

Barlow and Hunter [25] first discussed this problem, using a periodic replacement model with minimal repairs between replacements. It is assumed that there is only minimal repair after failures, so without any effect on the system's increasing failure rate $h(t)$. Replacement occurs at times $T, 2T, 3T, \dots$, and the cost of a minimal repair c_m is less than the cost of replacing the entire system c_s . The problem is to minimize $C(T)$, the long-run expected costs per unit time under policy $T > 0$, which, using the renewal reward theorem is given by

$$C(T) = \lim_{t \rightarrow \infty} \frac{c_m E[N_m(t)] + c_s E[N_s(t)]}{t} = \frac{c_m \int_0^T h(x) dx + c_s}{T} \quad (4)$$

where $N_m(t)$ and $N_s(t)$ are the number of minimal repairs and system replacements during $[0, t)$, respectively. Let the optimal strategy be denoted by T^* ,

then the first-order necessary condition for optimality [2] is

$$\int_0^T [h(T) - h(x)] dx = \frac{c_s}{c_m} \quad (5)$$

If there is an interval $[0, b)$, with $0 < b \leq \infty$, on which $h(t)$ is continuous and unbounded, then equation (5) has a solution, which is unique if $h(t)$ is strictly increasing and differentiable [2]. Generalizations of this model include, for example, minimal repair costs that depend on the number of such repairs since the last replacement or on the age of the system [26, 27]. In [28], both the replacement and repair costs may vary.

For a finite time horizon, the class of *sequential replacement policies* (see **Product Risk Management: Testing and Warranties**) is important. Under a sequential policy, the next planned replacement is selected to minimize expected expenditure during the remaining time. For an analysis of a sequential replacement model, see [2].

Inspection Models

Inspection models consider situations in which faults are discovered only by inspection, often after some time has elapsed since a failure occurred. The goal is to determine inspection policies that minimize the total expected costs resulting from both inspection and failure. At each decision moment, one must decide which maintenance action to take and when next to inspect the system.

In the basic model [29], it is assumed that upon detection of failure the problem ends, that is, no replacement or repair takes place. Each inspection incurs a cost, and there is a penalty cost associated with the lapsed time between occurrence and detection of a failure. We assume that (a) system failure is known only through inspection, (b) inspection takes negligible time and does not affect the system, (c) the system cannot fail while being inspected, (d) each inspection entails a fixed cost c_i , (e) undiscovered system failure costs c_s per unit of time, and (f) inspection ceases upon discovery of a failure. Let $N_i(t)$ be the number of inspections during $[0, t]$, γ_t the interval between failure and its discovery if failure occurs at time t , and F the cdf of the failure time. Then the expected costs are

$$C = \int_0^\infty (c_i(E[N_i(t)] + 1) + c_s \gamma_t) dF(t) \quad (6)$$

An optimum inspection policy specifies successive inspection times $x_1 < x_2 < \dots$ for which C is minimized. It has been proved that the optimal policies are nonrandom [29]. If we define $x_0 = 0$ and require that the sequence $\{x_n\}$ be such that the support of F is contained in $[0, \lim_{n \rightarrow \infty} x_n)$, to exclude the possibility of undetected failure occurring with positive probability, then equation (6) becomes

$$C = \sum_{k=0}^{\infty} \int_{x_k}^{x_{k+1}} (c_i(k+1) + c_s(x_{k+1} - t)) dF(t) \quad (7)$$

Barlow *et al.* [29] presented several important results. An optimal inspection schedule exists if F is continuous with finite first moment. If it is known that failure must occur by time $T > 0$, so $F(T) = 1$, then if $F(t) \leq (1 + (c_s/c_i)(T - t))^{-1}$ for $0 \leq t \leq T$ the optimal inspection schedule consists of a single inspection performed at time T . Otherwise, in addition to the inspection at time T , one or more earlier inspections are required. They also presented an algorithm for computing the optimum inspection policy under certain conditions. Beichelt [30] generalized this basic model by allowing repair upon detection of failure. He used a minimax approach to find the optimal inspection policy in case of partially known failure time distributions. Luss [31] considered the basic repair model of Derman [20] in which both corrective and preventive replacement are allowed, but with the state of deterioration only known through inspection. He developed a simple iterative algorithm for the optimal control-limit policy and corresponding optimal inspection intervals in all states, minimizing the long-run expected costs per unit time. Many further generalizations have been presented, including imperfect or unreliable inspections.

Christer and Waller [32] presented models for optimal inspection and replacement using *delay time analysis* (see **Accelerated Life Testing; Availability and Maintainability**). The delay time h of a fault is the time lapse from when a fault could first be noticed until the time when its repair can be delayed no longer because of unacceptable consequences. A repair may be undertaken any time within this period. They assume that (a) an inspection takes place every T time units, costs I , and requires d time units, (b) inspections are perfect, (c) defects identified at an inspection will be repaired within the inspection period, (d) the initial instant at which a defect may be assumed to first arise is uniformly distributed over

time since the last inspection, independent of h , and faults arise at the rate of k per unit time, (e) the pdf f of the delay time is known, or can be estimated. The basic model is as follows: Suppose that a fault arising within the period $[0, T)$ has a delay time in the interval $[h, h + dh)$. This fault will be repaired as a breakdown repair if the fault arises in period $(0, T - h)$, otherwise as an inspection repair. The probability of a fault arising as a breakdown, for an inspection policy of period T , is

$$b(T) = \int_0^T \left(\frac{T - h}{T} \right) f(h) dh \quad (8)$$

If the average downtime for breakdown repair is d_b and the average breakdown and inspection repair costs are c_b and c_i , respectively, then the expected downtime per unit time for inspection policy T is

$$D(T) = \frac{1}{T + d} \{kT d_b b(T) + d\} \quad (9)$$

with total expected costs per unit time equal to

$$C(T) = \frac{1}{T + d} \{kT [c_b b(T) + c_i (1 - b(T))] + I\} \quad (10)$$

The optimal inspection policy can be calculated easily, as the cost model is simple.

Example 2 (Delay time model) Consider an inspection model for a unit with delay time pdf $f(h) = 0.05 e^{-0.05h}$ for $h \geq 0$. Assume that faults arise at a rate of 0.1 faults per hour, and that the downtime for breakdown repair (d_b) and for inspection and subsequent repairs (d) are 0.5 and 0.35 hours, respectively. The average costs of a breakdown repair $c_b = 10$ and of inspection repair $c_i = 5$. The cost of inspection $I = 2$. The probability of a fault arising as a breakdown and the expected downtime per unit time, for inspection policy T , are

$$b(T) = 1 - \frac{20}{T} (1 - e^{-0.05T}) \quad \text{and} \quad D(T) = \frac{1}{T + 0.35} (0.05T + e^{-0.05T} - 0.65) \quad (11)$$

and the total expected costs per hour are

$$C(T) = \frac{1}{T + 0.35} (T - 8 + 10 e^{-0.05T}) \quad (12)$$

Minimizing the total expected costs per hour yields $T^* = 14.99$, so inspection should take place every 15 hours with minimal total expected costs per hour of 0.764. The corresponding probability of a fault arising as a breakdown is 0.296, while the corresponding expected downtime per hour equals 0.03729 hours.

If the objective is to minimize the expected downtime per hour, we obtain optimal $\tilde{T} = 24.02$, so inspection should take place every 24 hours, and $D(\tilde{T}) = 0.03496$. In this case, the total expected costs per hour are 0.781 and the probability of a breakdown repair during an inspection period equals 0.418. Note that although the expected downtime per hour under policy $\tilde{T}$ is smaller than under policy T^* , the probability of a breakdown repair has increased owing to the longer inspection period.

Several generalizations of this basic delay time model have been presented, including delay time models with imperfect inspection. An important result with respect to the assumption of known delay time distribution is obtained by Coolen and Dekker [33]. They considered a model in which a system is subject to several failure modes (competing risks), but there exists an intermediate state to failure that can be detected by inspection only for one failure mode. This intermediate state corresponds to the delay time for that failure mode. They showed that, in practice, the first two moments of the delay time distribution are sufficient to obtain the optimal inspection policy so that assumption (e) can be relaxed, and also that competing risks do not need to be considered in determining the optimal inspection policy.

Shock Models

We consider systems that are subject to shocks, where each shock causes an amount of damage that accumulates additively until replacement or failure. The time between shocks and the damage caused by a shock are random variables that may depend on $X(t)$, the accumulated damage at time t . The system is replaced upon failure at a cost $c(\Delta)$, where Δ denotes the state of failure, but it may also be replaced before failure at a cost $c(x) \leq c(\Delta)$ if the damage level at the time of replacement is x . It is assumed that the replacement cost function $c(\cdot)$ is a nondecreasing function of the accumulated damage, that replacements take a negligible amount of time,

and that damaged systems are replaced by new, identical systems. Let ξ be the failure time and T the replacement time. Then, using the renewal reward theorem, the long-run expected costs per unit time are

$$C(T) = \frac{P\{T < \xi\}E[c(X(T))] + P\{T = \xi\}c(\Delta)}{E[T \cap \xi]} \quad (13)$$

Most shock models give conditions under which the optimal policy takes the form of a control-limit policy, i.e., the system is replaced when the accumulated damage exceeds a critical level α or at failure, whichever occurs first. In this case, the replacement time T is

$$T = \min\{\inf\{t > 0 : X(t) > \alpha\}, \xi\} \quad (14)$$

Many generalizations of this model have been presented, including shock models where the cumulated damage is a nondecreasing semi-Markov process and failures can occur at any time [34], and models where system deterioration occurs both continuously and due to shocks [35] (*see Mathematics of Risk and Reliability: A Select History*).

Data-Based Models

We present two maintenance models for which the failure time distribution is unknown, but for which past realizations of the failure time are available. According to McCall [1], one way to circumvent the difficulty of incomplete information is by means of *Bayes adaptive models*. There are a few earlier papers known in literature concerning Bayesian analysis of maintenance models, but Mazzuchi and Soyer [14] were the first to provide a fully Bayesian analysis of the age and block-replacement models with minimal repair (see section titled “Group Maintenance Models”). In their approach to age replacement, let t_A be the replacement age, T the time to failure, and c_f and c_p the corrective and preventive replacement costs. They model the time to failure with a Weibull distribution with unknown scale and shape parameters α and β . The length of the time between successive replacements, and the costs per unit time, are random variables defined by

$$\begin{aligned} L(T, t_A) &= \begin{cases} t_A & \text{if } T \geq t_A \\ T & \text{if } T < t_A \end{cases} \quad \text{and} \\ C(T, t_A) &= \begin{cases} \frac{c_p}{t_A} & \text{if } T \geq t_A \\ \frac{c_f}{T} & \text{if } T < t_A \end{cases} \end{aligned} \quad (15)$$

The uncertainty about α and β is expressed by prior distributions, based on expert judgments and obtained by elicitation. These can be updated when failure and replacement data become available. The optimal policy is obtained as the value t_A that minimizes $E_{\alpha,\beta}[E[C(T, t_A)|\alpha, \beta]]$. Note that they do not use the renewal reward theorem, as its use is not appropriate in case of an adaptive policy.

If the failure distribution is unknown, but failure data are available, one can consider maintenance policies within an *adaptive nonparametric predictive inference (NPI) framework* (see **Imprecise Reliability**), as presented in [12, 13, 17] for adaptive (opportunity-based) age replacement. NPI is based on Hill's assumption A_n [36, 37]. Denoting n ordered, observed failure times by $x_1 < x_2 < \dots < x_n$, A_n specifies direct probabilities for a future failure time X_{n+1} by

$$P\{X_{n+1} \in (x_j, x_{j+1})\} = \frac{1}{n+1}, \quad j = 0, \dots, n \quad (16)$$

where $x_0 = 0$ and $x_{n+1} = \infty$, or $x_{n+1} = r$ if we can safely assume a finite upper bound r for the support of X_{n+1} . This approach is suitable if there is hardly any knowledge about the random quantities of interest, other than the n observations. It does not provide precise probabilities for all possible events of interest, but it does provide optimal bounds for all probabilities by application of De Finetti's Fundamental Theorem of Probability (see **Subjective Probability**) [38]. Such bounds are lower and upper probabilities within the theory of interval probability [37]. We illustrate this approach with the age-replacement problem based on the renewal reward criterion as described in the section titled "Repair Models" [12]; opportunity-based age replacement and age replacement with a one-cycle criterion are presented in [13] and [17].

Denote by $S(x) = 1 - F(x)$ the survival function. Instead of assuming a known probability distribution F for the time to failure of a unit, imprecise predictive survival functions for the time to failure of the next unit are used based on A_n . The maximum lower bound $\underline{S}(x)$ ("lower survival function") and the minimum upper bound $\bar{S}(x)$ ("upper survival function") are obtained by shifting the probability mass in the interval in which x lies to

the left and right end point of the interval respectively,

$$\underline{S}_{X_{n+1}}(x) = S_{X_{n+1}}(x_{j+1}) = \frac{n-j}{n+1} \quad \text{for } x \in (x_j, x_{j+1}], \quad j = 0, \dots, n \quad (17)$$

and

$$\bar{S}_{X_{n+1}}(x) = S_{X_{n+1}}(x_j) = \frac{n+1-j}{n+1} \quad \text{for } x \in [x_j, x_{j+1}), \quad j = 0, \dots, n \quad (18)$$

The cost function $C(T)$ (equation (2)) is decreasing as function of $S(\cdot)$, in the sense that $C(T)$ decreases if $S(x)$ increases for $x \in (0, T]$. This implies that these lower and upper survival functions for X_{n+1} straightforwardly lead to the following lower bounds $\underline{C}$ and upper bounds $\bar{C}$ for the cost functions,

$$\begin{aligned} \underline{C}_{X_{n+1}}(x_j) &= \frac{jc_f + (n+1-j)c_p}{(n+1-j)x_j + \sum_{l=1}^j x_l}, \\ &\quad j = 1, \dots, n+1 \\ \underline{C}_{X_{n+1}}(T) &= \frac{jc_f + (n+1-j)c_p}{(n+1-j)T + \sum_{l=1}^j x_l}, \\ &\quad T \in (x_j, x_{j+1}), \quad j = 0, \dots, n \\ \bar{C}_{X_{n+1}}(x_j) &= \frac{jc_f + (n+1-j)c_p}{(n+1-j)x_j + \sum_{l=1}^{j-1} x_l}, \\ &\quad j = 1, \dots, n+1 \\ \bar{C}_{X_{n+1}}(T) &= \frac{(j+1)c_f + (n-j)c_p}{(n-j)T + \sum_{l=1}^j x_l}, \\ &\quad T \in (x_j, x_{j+1}), \quad j = 0, \dots, n \end{aligned} \quad (19)$$

Coolen-Schrijner and Coolen [12] showed that the global minimum of $\bar{C}_{X_{n+1}}(\cdot)$ is assumed in one of the points x_j , $j = 1, \dots, n$, and the minimum of $\underline{C}_{X_{n+1}}(\cdot)$ on $(0, x_n]$ is assumed in one of the x_j^- , $j = 1, \dots, n$, with x_j^- , meaning "just before x_j ".

Table 1 NPI adaptive age replacement

	$T_{\text{low},n+1}^*$	$\Delta_{\text{low},n+1}$	$T_{\text{low},n+2}^*$	$\Delta_{\text{low},n+2}$	$T_{\text{up},n+1}^*$	$\Delta_{\text{up},n+1}$	$T_{\text{up},n+2}^*$	$\Delta_{\text{up},n+2}$
Mean	0.3553	0.0313	0.3529	0.0310	0.3694	0.0315	0.3673	0.0312
Median	0.3453	0.0152	0.3433	0.0151	0.3588	0.0157	0.3564	0.0155
SD	0.0926	0.0414	0.0917	0.0409	0.0948	0.0412	0.0941	0.0409

If we assume a known upper bound r for the support of X_{n+1} , $\underline{C}_{X_{n+1}}(\cdot)$ is also strictly decreasing on (x_r, r) , so we must also consider $\underline{C}_{X_{n+1}}(r^-)$ for finding the global minimum on $(0, r]$. If this minimum is attained in r^- , then it is better not to replace preventively.

Minimization of the lower and upper cost functions with regard to age replacement of unit $n+1$ may lead to two different optimal replacement times, $T_{\text{low},n+1}^*$ and $T_{\text{up},n+1}^*$. We now present a general theory for which it is not relevant to make the distinction, in which case we use T_{n+1}^* as generic notation for an optimal age-replacement time that was used for unit $n+1$. Under the optimal age-replacement strategy for X_{n+1} , the observation for X_{n+1} is either a failure time less than T_{n+1}^* , in which case the unit was replaced correctively, or a right-censored observation at T_{n+1}^* , in which case the unit was replaced preventively. To study the adaptive nature of this procedure, we now want to consider the optimal age-replacement strategy for unit $n+2$. If unit $n+1$ fails before T_{n+1}^* , we can directly apply the results presented before on the $n+1$ failure times, with A_n replaced by A_{n+1} to obtain the optimal age-replacement time for unit $n+2$. If unit $n+1$ is preventively replaced at $T_{n+1}^* = x_k$ for some $k \in \{1, \dots, n\}$, the relevant data set now consists of the n original failure times together with a right-censored observation at x_k . For NPI based on such data, the assumption rc- A_{n+1} [39] can be applied, leading to upper and lower survival functions for X_{n+2} , which consequently lead to lower and upper cost functions for X_{n+2} . This updating process is presented in [12], and illustrated by the following example.

Example 3 (NPI age-replacement model) A simulation study of 10 000 runs was carried out, in which in each run 100 failure times were simulated from a W(2,1) distribution, enabling us to compare the optimal replacement times corresponding to the NPI lower and upper cost functions with the theoretical optimal replacement time $T^* = 0.3365$ from Example 1, assuming the same costs as in that example. For each run, we calculated the optimal age-replacement

times for units $n+1$ and $n+2$, the latter after a further observation for unit $n+1$ had been simulated and its optimal replacement strategy applied, by minimizing the lower and upper cost functions for X_{n+1} and X_{n+2} . Table 1 gives some results with the following notation: $T_{\text{low},n+i}^* = \arg\min \underline{C}_{X_{n+i}}(T)$, $T_{\text{up},n+i}^* = \arg\min \overline{C}_{X_{n+i}}(T)$, $\Delta_{\text{low},n+i} = (C(T_{\text{low},n+i}^*) - C(T^*)) / C(T^*)$, $\Delta_{\text{up},n+i} = (C(T_{\text{up},n+i}^*) - C(T^*)) / C(T^*)$, $i = 1, 2$, where $C(\cdot)$ is the cost function defined in equation (2). These Δ 's indicate how good the NPI optimum replacement times are compared to the theoretical optimum, judged by comparing the loss in long-run expected costs per unit time if the NPI optimum is used instead of the theoretical optimum, as fraction of the long-run expected costs per unit time in the theoretical optimum.

The means of the $T_{\text{low},n+i}^*$ and $T_{\text{up},n+i}^*$, $i = 1, 2$, from the simulations are all larger than the theoretical T^* . However, as the distributions of the values of $T_{\text{low},n+i}^*$ and $T_{\text{up},n+i}^*$, $i = 1, 2$ are all skewed to the right, the medians may be better indications of performance of the NPI method. We also see that the mean, median, and standard deviation of $T_{\text{low},n+1}^*$, $T_{\text{up},n+1}^*$, $\Delta_{\text{low},n+1}$, and $\Delta_{\text{up},n+1}$ are all larger than the corresponding quantities at time $n+2$, which reflects that the NPI-based optimal replacement times adapt well to the process data.

Multiunit Maintenance Models

In multiunit systems, units may depend on each other economically, structurally, or stochastically, as described in the section titled “Model Characteristics”. Consequently, the optimal maintenance policy is not derived by considering each unit separately, and maintenance decisions are not independent. For such situations, we discuss group maintenance, repair/inventory, inspection/maintenance, and cannibalization models. For overviews of maintenance models for multiunit systems, see e.g. [4, 8–10].

Group Maintenance Models

When considering an infinite time horizon, group maintenance models can be divided into three categories [9]: (a) Group corrective maintenance models, (b) group preventive maintenance models, and (c) group opportunistic maintenance models.

Group Corrective Maintenance Models. In such models, units are only correctively maintained, and a failed unit can be left until its corrective maintenance is carried out jointly with that of other failed units, so this requires some redundancy to be available in the systems. If there exists economies of scale, simultaneous corrective repair is advantageous, but leaving units in a failed condition for some time increases the risk of costly production losses. The basic group maintenance model [9] for a system consisting of n identical parallel units assumes that (a) the number of failed units and the age of nonfailed units is always known; (b) repair results in an “as good as new” unit; (c) repairs can start at any time; (d) the functionality of the nonfailed units is not affected by a failed unit (no probabilistic dependency). The costs consist of two components, one concerned with loss of productivity at a rate $C_1(k)$ per unit of time when k units are down, the other related to the cost $C_2(k)$ of repair per unit when k units are in repair. The goal is to minimize the long-run expected costs per unit time. Several results have been presented for specific functions $C_1(k)$ and $C_2(k)$.

Group Preventive Maintenance Models. An important class of group preventive maintenance models is the class of *block-replacement models*, where a unit is replaced at failure *and* all units of the group are replaced simultaneously at times kT , $k = 1, 2, \dots$, independent of the system’s failure history.

Let F be the cdf of the failure time and $M(T) = E[N(T)]$ the expected number of failures in $[0, T]$, and assume that the corresponding pdf’s, f and $m(T)$, exist. The cost of corrective repair is c_f and the cost of the block replacement is c_b . For block-replacement policy with interval T over an infinite time horizon, using the renewal reward theorem, the long-run expected costs per unit time are

$$C(T) = \frac{c_f M(T) + c_b}{T} \quad (20)$$

If F is continuous, $C(T)$ has a minimum for $0 < T \leq \infty$. A necessary condition for a finite

value T^* to minimize equation (20) is $T^*m(T^*) - M(T^*) = c_b/c_f$ [2], and the corresponding minimal costs are $c_fm(T^*)$.

Example 4 (Block-replacement model) Suppose that a unit’s failure time has pdf $f(t) = \lambda^2 t e^{-\lambda t}$ with $\lambda = 0.1$, for $t \geq 0$, and let $c_f = 10$ and $c_b = 2$. The corresponding expected number of failures during $[0, T]$ is

$$M(T) = \frac{\lambda T}{2} - \frac{1}{4} (1 - e^{-2\lambda T}) \quad (21)$$

The optimal block-replacement time T^* is obtained by solving

$$\left(\frac{\lambda T}{2} + \frac{1}{4} \right) e^{-2\lambda T} = \frac{1}{20} \quad (22)$$

which yields $T^* = 14.97$ with $C(T^*) = 0.475$, so the optimal block-replacement policy is to replace a unit at failure, while also replacing all units at $15k$ for all $k = 1, 2, \dots$. The expected number of failures during each such 15-units interval equals 0.511.

The block-replacement model assumes a fixed group of units. Another group maintenance problem concerns the optimal grouping of the units for such cases, in which case the combinatorial aspects complicate the problem [9].

Group Opportunistic Maintenance Models. In opportunistic maintenance models, both preventive and corrective maintenance can be used to obtain economies of scale, with preventive maintenance possible if opportunities occur. In such cases, the unpleasant event of a failing unit may provide an opportunity for preventive maintenance of other units. Besides models where the process that generates opportunities depends on the failure process, models where these processes are independent have also been presented [40].

For a finite time horizon, dynamic group maintenance models are more appropriate. In such models, short-term information on deterioration of units or unexpected opportunities is taken into account. Dynamic models can be categorized according to *finite* or *rolling* time horizon [9]. In finite horizon models, it is assumed that the system is not used afterwards. Rolling horizon models also use finite horizons, but they do so repeatedly and based on a long-term plan: once decisions based on a finite

horizon are implemented, or when new information becomes available, a new horizon is considered and a tentative plan based on the long term is adapted according to short-term circumstances. Rolling horizon models aim to bridge the gap between finite and infinite time horizon models, while combining advantages of both [9].

Other Maintenance Models

Maintenance models for multiunit systems are difficult to handle mathematically, in particular, in case of stochastic dependence among the units. In such cases, few analytical results are obtainable on operating characteristics and optimal activities. Often, the only way to deal with such problems is by simulation or approximations. For some classes of models, some theoretical results have been presented.

1. Repair/inventory maintenance models

Most maintenance models discussed above assume that the replacement units are always available. However, in real life situations, availability of spare units is an important factor in determining an optimal maintenance policy for a system. Besides decisions on optimal times and levels of repair and replacement, other main decisions to be made include the decision on optimal times and levels of replenishment of spare units.

2. Inspection/maintenance models

In preparedness maintenance models, the current states of the units of the system are unknown unless an inspection is performed. Assaf and Shanthikumar [41] present an inspection/maintenance model considering a system consisting of N units, each having exponential failure time with constant rate. A failed unit can be repaired at any time, and a repaired unit is considered as good as new. It is assumed that the number of failed units is unknown, unless an inspection is carried out. Upon an inspection, a decision is required on whether to repair the failed units or not, based on the number of failed units in the system. The cost function consists of inspection costs, repair costs, and penalty costs associated with production losses due to failed units. Under the criterion of minimal long-run expected costs per unit time, they derived a necessary and sufficient condition for an optimal policy, together with its general form.

Barker and Newby [42] consider a multiunit system in which the deterioration of each component is modeled by a Wiener process. They derive an optimal inspection and maintenance policy for a model with two thresholds ξ and L ($\xi < L$), such that if the system's deterioration exceeds the first threshold ξ , maintenance is performed, while if it reaches the second threshold L , the system fails. The first threshold ξ ensures a minimum level of reliability.

3. Cannibalization models

In cannibalization models, units of the same type are used in different subsystems. Cannibalization simply involves interchanging units between subsystems, to increase the number of functioning subsystems, whenever some units in the system are inoperative. Such a replacement policy can be used when there are no spare units available.

There is a lot of ongoing research on maintenance models for multiunit systems, in particular because multiunit systems can be so complex that each model is system specific. For more information about maintenance models for multiunit systems, we refer the reader to the overview papers [1, 4–6, 8–10].

References

- [1] McCall, J.J. (1965). Maintenance policies for stochastically failing equipment: a survey, *Management Science* **11**, 493–524.
- [2] Barlow, R.E. & Proschan, F. (1965). *Mathematical Theory of Reliability*, John Wiley & Sons, New York.
- [3] Van der Duyn Schouten, F.A. & Tapiero, C.S. (1995). OR models for maintenance management and quality control, *European Journal of Operational Research* **82**, 211–213.
- [4] Pierskalla, W.P. & Voelker, J.A. (1976). A survey of maintenance models: the control and surveillance of deteriorating systems, *Naval Research Logistics Quarterly* **23**, 353–388.
- [5] Sherif, Y.S. & Smith, M.L. (1981). Optimal maintenance models for systems subject to failure – a review, *Naval Research Logistics Quarterly* **28**, 47–74.
- [6] Thomas, L.C. (1986). A survey of maintenance and replacement models for maintainability and reliability of multi-item systems, *Reliability Engineering* **16**, 297–309.
- [7] Valdez-Flores, C. & Feldman, R.M. (1989). A survey of preventive maintenance models for stochastically deteriorating single-unit systems, *Naval Research Logistics* **36**, 419–446.

- [8] Cho, D.I. & Parlar, M. (1991). A survey of maintenance models for multi-unit systems, *European Journal of Operational Research* **51**, 1–23.
- [9] Dekker, R., Wildeman, R.E. & Van der Duyn Schouten, F.A. (1997). A review of multi-component maintenance models with economic dependence, *Mathematical Methods of Operations Research* **45**, 411–435.
- [10] Wang, H. (2002). A survey of maintenance policies of deteriorating systems, *European Journal of Operational Research* **139**, 469–489.
- [11] Derman, C., Lieberman, G.J. & Ross, S.M. (1984). On the use of replacements to extend system life, *Operations Research* **32**, 616–627.
- [12] Coolen-Schrijner, P. & Coolen, F.P.A. (2004). Adaptive age replacement strategies based on nonparametric predictive inference, *Journal of the Operational Research Society* **55**, 1281–1297.
- [13] Coolen-Schrijner, P., Coolen, F.P.A. & Shaw, S. (2006). Nonparametric adaptive opportunity-based age replacement strategies, *Journal of the Operational Research Society* **57**, 63–81.
- [14] Mazzuchi, T.A. & Soyer, R. (1996). A Bayesian perspective on some replacement strategies, *Reliability Engineering and System Safety* **51**, 295–303.
- [15] Sheu, S.H., Yeh, R.H., Lin, Y.B. & Juan, M.G. (1999). A Bayesian perspective on age replacement with minimal repair, *Reliability Engineering and System Safety* **65**, 55–64.
- [16] Sheu, S.H., Yeh, R.H., Lin, Y.B. & Juan, M.G. (2001). A Bayesian approach to an adaptive preventive maintenance model, *Reliability Engineering and System Safety* **71**, 33–44.
- [17] Coolen-Schrijner, P. & Coolen, F.P.A. (2007). Nonparametric adaptive age replacement with a one-cycle criterion, *Reliability Engineering and System Safety* **92**, 74–84.
- [18] Ansell, J., Bendell, A. & Humble, S. (1984). Age replacement under alternative cost criteria, *Management Science* **30**, 358–367.
- [19] Coolen-Schrijner, P. & Coolen, F.P.A. (2006). On optimality criteria for age replacement, *Journal of Risk and Reliability* **220**, 21–29.
- [20] Derman, C. (1962). On sequential decisions and Markov chains, *Management Science* **9**, 16–24.
- [21] Klein, M. (1962). Inspection-maintenance-replacement schedules under Markovian deterioration, *Management Science* **9**, 25–32.
- [22] Derman, C. (1970). *Finite State Markovian Decision Processes*, Academic Press, New York.
- [23] Nakagawa, T. & Osaki, S. (1974). The optimal repair limit replacement policies, *Operations Research Quarterly* **25**, 311–317.
- [24] Dekker, R. & Dijkstra, M.C. (1992). Opportunity-based age replacement: exponentially distributed times between opportunities, *Naval Research Logistics* **39**, 175–190.
- [25] Barlow, R.E. & Hunter, L.C. (1960). Optimum preventive maintenance policies, *Operations Research* **8**, 90–100.
- [26] Boland, P.J. & Proschan, F. (1982). Periodic replacement with increasing minimal repair costs at failure, *Operations Research* **30**, 1183–1189.
- [27] Boland, P.J. (1982). Periodic replacement when minimal repair costs vary with time, *Naval Research Logistics Quarterly* **29**, 541–546.
- [28] Aven, T. (1983). Optimal replacement under a minimal repair strategy – a general failure model, *Advances in Applied Probability* **15**, 198–211.
- [29] Barlow, R.E., Hunter, L.C. & Proschan, F. (1963). Optimum checking procedures, *Journal of the Society for Industrial and Applied Mathematics* **4**, 1078–1095.
- [30] Beichelt, F. (1981). Minimum inspection strategies for single unit systems, *Naval Research Logistics Quarterly* **28**, 375–381.
- [31] Luss, H. (1976). Maintenance policies when deterioration can be observed by inspection, *Operations Research* **24**, 359–366.
- [32] Christer, A.H. & Waller, W.M. (1984). Delay time models of industrial inspection maintenance problems, *Journal of the Operational Research Society* **35**, 401–406.
- [33] Coolen, F.P.A. & Dekker, R. (1995). Analysis of a 2-phase model for optimization of condition-monitoring intervals, *IEEE Transactions on Reliability* **44**, 505–511.
- [34] Aven, R. & Gaarder, S. (1987). Optimal replacement in a shock model: discrete time, *Journal of Applied Probability* **24**, 281–287.
- [35] Hordijk, A. & Van der Duyn Schouten, F.A. (1983). Average optimal policies in Markov decision drift processes with applications to a queueing and a replacement model, *Advances in Applied Probability* **15**, 274–303.
- [36] Hill, B.M. (1968). Posterior distribution of percentiles: Bayes's theorem for sampling from a population, *Journal of the American Statistical Association* **63**, 677–691.
- [37] Augustin, T. & Coolen, F.P.A. (2004). Nonparametric predictive inference and interval probability, *Journal Statistical Planning and Inferences* **124**, 251–272.
- [38] De Finetti, B. (1988). *Theory of Probability*, John Wiley & Sons, London.
- [39] Coolen, F.P.A. & Yan, K.J. (2004). Nonparametric predictive inference with right-censored data, *Journal of Statistical Planning and Inference* **126**, 25–54.
- [40] Dekker, R. & Smeitink, E. (1991). Opportunity-based block replacement: the single component case, *European Journal of Operational Research* **53**, 46–63.
- [41] Assaf, D. & Shanthikumar, J.G. (1987). Optimal group maintenance policies with continuous and periodic inspections, *Management Sciences* **33**, 1440–1452.
- [42] Barker, C.T. & Newby, M. (2006). Optimal inspection and maintenance to meet prescribed performance standards, in *Safety and Reliability for Managing Risk*, C. Guedes Soares & E. Zio, eds, Taylor & Francis, London, pp. 475–481.

Repairable Systems Reliability

Models for repairable systems must describe the random occurrence of events in time. After a repairable system fails, it is repaired. Quite often, we ignore the repair time or we do not count the repair time as part of the cumulative operating time. The assumptions we make about the effect of a repair often dictate the type of model we choose. For example, if we assume (in addition to a few other technical assumptions) that the repair is minimal, that is, it brings the system back to the condition just before the failure, then we are led to the Poisson process. If we assume that a repair brings a unit to a like-new condition, then we are led to the renewal process (see **Extreme Event Risk; Mathematics of Risk and Reliability: A Select History**). Under the piecewise exponential process, a repair effects a change in the level of the intensity. This change could be positive, leading to a higher intensity and a shorter expected time to the next failure, or negative, leading to a lower intensity and a longer expected time to the next failure. For the modulated Poisson process model, the effect of a repair is somewhere between minimal repair (same-as-old) and the renewal process (same-as-new).

These models, together with statistical inference for them, are covered in more detail in the next sections.

The Nonhomogeneous Poisson Process

Let $N(t)$ denote the number of failures through time t , and more generally, let $N(a, b]$ denote the number of failures in the interval $(a, b]$, with similar definitions for $N(a, b)$, $N[a, b)$, and $N[a, b]$. Thus, $N(t) = N(0, t]$. Suppose we make the following assumptions about $N(\cdot)$:

1. $N(0) = 0$,
2. if $a < b \leq c < d$, then $N(a, b]$ and $N(c, d]$ are independent random variables,
3. there exists a function $\lambda(t)$ such that

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(N(t, t + \Delta t] = 1)}{\Delta t}$$

4.
$$\lim_{\Delta t \rightarrow 0} \frac{P(N(t, t + \Delta t] \geq 2)}{\Delta t} = 0$$

When these assumptions hold, we say that the process is a nonhomogeneous Poisson process (NHPP).

A few comments about these assumptions are in order. The first assumption is basically for accounting; we begin with no failures. The second is called the *independent increments property*; the numbers of failures in nonoverlapping intervals are independent. The function $\lambda(t)$ in the third assumption is called the *intensity function*. This intensity function measures how likely failures are as a function of the cumulative operating time t . If $\lambda(t)$ is increasing, then we will have reliability deterioration; that is, we expect failures to occur more frequently. If $\lambda(t)$ is decreasing, then we have reliability improvement; that is, we expect failures to occur less frequently. The fourth assumption excludes the possibility of simultaneous failures. The four assumptions together imply that the number of failures in the interval $(a, b]$ has a Poisson distribution with mean $\int_a^b \lambda(t) dt$.

We often make assumptions about the functional form of $\lambda(t)$. One common assumption is that $\lambda(t)$ takes the form

$$\lambda(t) = \frac{\beta}{\theta} \left(\frac{t}{\theta} \right)^{\beta-1}, \quad t > 0 \quad (1)$$

This process is called the *power law process* (see **Product Risk Management: Testing and Warranties, Reliability Growth Testing**), or sometimes the Weibull process. The intensity function has the same form as the hazard function of the Weibull distribution, although the intensity function and the hazard function are conceptually different. Given data in the form of the successive failure times, we usually would like to estimate the parameters of the process, β and θ in the case of the power law process. Before proceeding with this, we must identify how data are collected. If the data collection stops at a predetermined time, then the data are said to be time truncated. If the data collection stops after a fixed number of failures, then the data are said to be failure truncated.

Inference for the NHPP: Failure Truncated Case

Suppose that we observe a repairable system whose failure times are governed by an NHPP with intensity function $\lambda(t)$. Let $T_1 < T_2 < \dots < T_n$ denote the (random) failure times measured from the initial startup of the system. The number of failures, n , is

predetermined. The joint probability density function (pdf) of the n failure times is

$$f(t_1, t_2, \dots, t_n) = \left(\prod_{i=1}^n \lambda(t_i) \right) \exp \left(- \int_0^{t_n} \lambda(t) dt \right) \quad (2)$$

For the special case of the power law process, where $\lambda(t) = \frac{\beta}{\theta} \left(\frac{t}{\theta} \right)^{\beta-1}$, the joint pdf, which can also be viewed as the likelihood function, is

$$\begin{aligned} f(t_1, t_2, \dots, t_n | \beta, \theta) \\ = L(\beta, \theta) = \frac{\beta^n}{\theta^{n\beta}} \left(\prod_{i=1}^n t_i \right)^{\beta-1} \exp \left(- \left(\frac{t_n}{\theta} \right)^\beta \right) \end{aligned} \quad (3)$$

From this, the maximum-likelihood estimators (MLEs) are found to be

$$\hat{\beta} = \frac{n}{\sum_{i=1}^{n-1} \ln(t_n/t_i)} \quad (4)$$

and

$$\hat{\theta} = \frac{t_n}{n^{1/\hat{\beta}}} \quad (5)$$

There must be at least $n = 2$ failures in order for the MLEs to exist. For details of the derivation of the MLEs, see Crow [1] or Rigdon and Basu [2].

In the power law process, the parameter β determines whether the intensity is increasing (deterioration) or decreasing (reliability growth). If $\beta = 1$, then the intensity function is equal to the constant $1/\theta$ (independent of t). This special case is the homogeneous Poisson process. Confidence intervals and tests of hypotheses for β are based on the result that $2n\beta/\hat{\beta}$ has a χ^2 distribution with $2(n-1)$ degrees of freedom. A $100 \times (1 - \alpha)\%$ confidence interval for β is

$$\frac{\chi_{1-\alpha/2}^2(2(n-1)) \hat{\beta}}{2n} < \beta < \frac{\chi_{\alpha/2}^2(2(n-1)) \hat{\beta}}{2n} \quad (6)$$

Since $\beta = 1$ is the boundary between decreasing and increasing intensity functions, it is often of interest to see whether 1 is contained in this confidence interval.

Hypothesis tests for β can also be based on the pivotal quantity $2n\beta/\hat{\beta}$. To test $H_0 : \beta = \beta_0$ against the two-sided alternative $H_a : \beta \neq \beta_0$, we would reject the null hypothesis if

$$\hat{\beta} < \frac{2n\beta_0}{\chi_{\alpha/2}^2(2(n-1))} \quad \text{or} \quad \hat{\beta} > \frac{2n\beta_0}{\chi_{1-\alpha/2}^2(2(n-1))} \quad (7)$$

We are often interested in testing $H_0 : \beta = 1$ against the alternative $H_a : \beta \neq 1$.

For the failure truncated case, confidence intervals and hypothesis tests for θ can be based on the tables of Finkelstein [3]. The entries in this table were determined analytically, so the confidence intervals and critical regions should have the nominal level.

In the situation of reliability growth, the important quantity is frequently the value of the intensity function when the testing stops. If no further improvements are made to the system, it is reasonable to assume that the failure process will be a homogeneous Poisson process whose constant intensity is equal to $\hat{\lambda}(t_n)$, the value of the intensity when testing stops. The MLE of $\lambda(t_n)$ is

$$\hat{\lambda}(t_n) = \frac{\hat{\beta}}{\hat{\theta}} \left(\frac{t_n}{\hat{\theta}} \right)^{\hat{\beta}-1} = \frac{n\hat{\beta}}{t_n} \quad (8)$$

Rigdon and Basu [2] give tables for constructing a confidence interval for $\lambda(t_n)$.

Goodness-of-fit tests for the power law process can be based on one of two transformation of the failure times $T_1 < T_2 < \dots < T_n$. One is the ratio-power transformation

$$\hat{R}_i = \left(\frac{t_i}{t_n} \right)^{\hat{\beta}} \quad (9)$$

where

$$\bar{\beta} = \frac{n-2}{\sum_{i=1}^{n-1} \ln(t_n/t_i)} \quad (10)$$

is an unbiased estimator for β . The rationale for this test is based on the following result. If the failure times are governed by a NHPP with intensity function $\lambda(t)$ and mean function $\Lambda = E(N(t)) = \int_0^t \lambda(u) du$,

then

$$R_i = \frac{\Lambda(t_i)}{\Lambda(t_n)} = \left(\frac{t_i}{t_n} \right)^\beta, \quad i = 1, 2, \dots, n-1 \quad (11)$$

are distributed as $n-1$ order statistics from the uniform distribution on the interval $(0,1)$. Using the unbiased estimator $\bar{\beta}$ in place of β gives $\hat{R}_i$ as defined above. Once this transformation is made, any goodness-of-fit test for the uniform distribution can be applied. Crow [1] suggests, and gives tables for, the Cramér–von Mises test which has test statistic

$$C_R^2 = \frac{1}{12(n-1)} + \sum_{i=1}^{n-1} \left(\hat{R}_i - \frac{2i-1}{2(n-1)} \right)^2 \quad (12)$$

The other transformation that leads to a goodness-of-fit test is

$$U_i = \ln \left(\frac{t_n}{t_{n-i}} \right), \quad i = 1, 2, \dots, n-1 \quad (13)$$

Rigdon and Basu [2] show that $U_1 < U_2 < \dots < U_{n-1}$ are distributed as $n-1$ order statistics from an exponential distribution with mean $1/\beta$. After this transformation is made, any goodness-of-fit test for the exponential distribution with unknown mean can be based on the U_i 's. See Rigdon [4] for a comparison of various tests for testing the power law process using the U_i 's.

A Bayesian analysis can be conducted on the power law process. The process follows the usual Bayesian paradigm.

1. Choose a prior distribution for (θ, β) that reflects prior belief about the parameters.
2. Collect data.
3. Inference is based on the posterior distribution of (θ, β) .

Guida *et al.* [5], Calabria *et al.* [6, 7], and Bar-Lev *et al.* [8] presented the details of a Bayesian analysis of the power law process and gave guidelines on the selection of a prior.

As an example of the power law process, consider the data from Duane [9] that is given in Table 1.

The MLEs of the parameters are

$$\hat{\beta} = \frac{14}{\sum_{i=1}^{13} \ln \left(\frac{4596}{t_i} \right)} = 0.483 \quad (14)$$

Table 1 Failure times of aircraft generator^(a)

Failure number i	Failure time t_i
1	10
2	55
3	166
4	205
5	341
6	488
7	567
8	731
9	1308
10	2050
11	2453
12	3115
13	4017
14	4596

^(a) Reproduced from [9]. © IEEE, 1964

and

$$\hat{\theta} = \frac{4596}{14^{1/0.483}} = 19.474 \quad (15)$$

Clearly, the times between failure are getting larger, indicating reliability improvement, so we would expect the estimate of β to be smaller than 1. Confidence intervals (95%) using the formulas from this section along with the appropriate tables are

$$0.239 < \beta < 0.723$$

and

$$0.21 < \theta < 161$$

Inference for the NHPP: Time Truncated Case

Suppose now that testing stops at a predetermined time t . Inference for this case is similar to the failure truncated case. For the failure truncated case, the number of failures n is fixed and the time when testing stops t_n is random. For the time truncated case, the reverse is true: the number of failures N is random and the time t when testing stops is fixed.

The likelihood function is

$$\begin{aligned} f(n, t_1, t_2, \dots, t_n | \beta, \theta) &= L(\beta, \theta) \\ &= \frac{\beta^n}{\theta^{n\beta}} \left(\prod_{i=1}^n t_i \right)^{\beta-1} \exp \left(- \left(\frac{t}{\theta} \right)^\beta \right) \end{aligned} \quad (16)$$

From this, the MLEs are found to be

$$\hat{\beta} = \frac{n}{\sum_{i=1}^n \ln(t/t_i)} \quad (17)$$

and

$$\hat{\theta} = \frac{t}{n^{1/\hat{\beta}}} \quad (18)$$

The number of failures must be $n \geq 1$ for the MLEs to exist. Conditioned on $N = n$, the statistic $2n\beta/\hat{\beta}$ has a χ^2 distribution with $2n$ degrees of freedom. This result leads to the confidence interval

$$\frac{\chi_{1-\alpha/2}^2(2n)\hat{\beta}}{2n} < \beta < \frac{\chi_{\alpha/2}^2(2n)\hat{\beta}}{2n} \quad (19)$$

Similarly, the null hypothesis $H_0 : \beta = \beta_0$ is rejected in favor of $H_a : \beta \neq \beta_0$ if

$$\hat{\beta} < \frac{2n\beta_0}{\chi_{\alpha/2}^2(2n)} \text{ or } \hat{\beta} > \frac{2n\beta_0}{\chi_{1-\alpha/2}^2(2n-1)} \quad (20)$$

The MLE of the intensity at the conclusion of the testing phase, that is, at time t , is

$$\hat{\lambda}(t) = \frac{\hat{\beta}}{\hat{\theta}} \left(\frac{t}{\hat{\theta}} \right)^{\hat{\beta}-1} = \frac{n\hat{\beta}}{t} \quad (21)$$

Bain and Engelhardt [10] discuss further the inference for the time truncated power law process.

Goodness-of-fit tests can be constructed in a manner analogous to the failure truncated case. The appropriate transformations are

$$\hat{R}_i = \left(\frac{t_i}{t} \right)^{\hat{\beta}} \quad (22)$$

and

$$W_i = \ln(t/t_{n-i+1}), \quad i = 1, 2, \dots, n \quad (23)$$

The Renewal Process

If the repair process is such that it brings failed/ repaired items to a like-new condition, then the times between successive failures should be independent and identically distributed (i.i.d.). If we

let $X_1, X_2, \dots, X_n$ denote the times between failure, then the failure times are

$$\begin{aligned} T_1 &= X_1 \\ T_2 &= X_1 + X_2 \\ T_3 &= X_1 + X_2 + X_3 \\ &\vdots \\ T_n &= X_1 + X_2 + \dots + X_n \end{aligned} \quad (24)$$

If we let $\Lambda(t) = E(N(t))$, as we did for the NHPP, then we can define

$$\mu(t) = \Lambda'(t) = \text{rate of change in expected number of failures} \quad (25)$$

The function $\mu(t)$ is called the *rate of occurrence of failures* (ROCOFs). (Note that for the NHPP, the ROCOF is equal to the intensity function. For the renewal process, the intensity function makes no sense, but the ROCOF does, and we will study its properties.) It can be shown [2] that if $g_k(t)$ is the pdf of T_k , then

$$\mu(t) = \sum_{k=1}^{\infty} g_k(t) \quad (26)$$

For example, if the times between failure are i.i.d. gamma random variables with parameters $\theta = 0.1$ and $\kappa = 20$, having mean $\mu = \theta\kappa = 2$ and variance $\sigma^2 = \theta^2\kappa = 0.2$, then the ROCOF function is as shown in Figure 1. The pdf's for the failure times $T_1 < T_2 < \dots$ are shown in Figure 1(a), and the ROCOF is shown in Figure 1(b). The ROCOF is initially very low because the system starts out with a brand new component that is unlikely (under this model) to fail early in its life. Around time $t = 2$ is when we would expect the item to fail for the first time, so the ROCOF is large there, after which it decreases. Figure 1(b) suggests that the ROCOF will bounce up and down around the times of the expected failures, but after a while, the randomness in the failure times will make the ROCOF nearly constant. This suggests the following result. (The proof is long and complicated, but is given in Rigdon and Basu [2].) For a renewal process with times between failures $X_1, X_2, \dots$, each a continuous random variable having support $[0, \infty)$ and having mean η and finite variance,

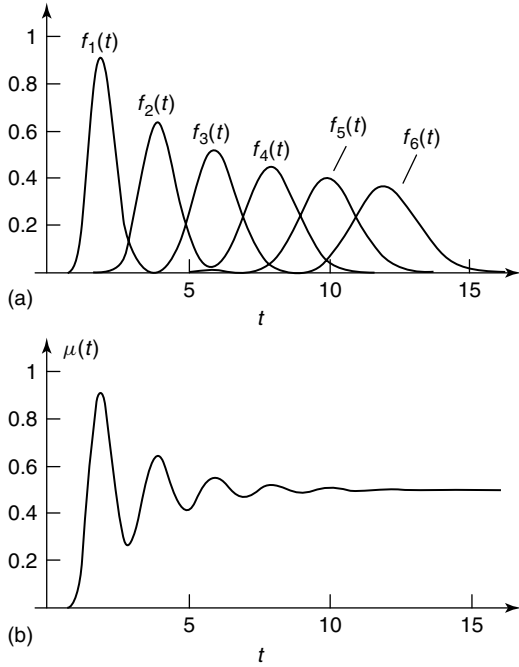


Figure 1 Pdf's and ROCOF for a gamma renewal process

$$\lim_{t \rightarrow \infty} \frac{\Lambda(t)}{t} = \frac{1}{\eta} \quad (27)$$

$$\lim_{t \rightarrow \infty} \mu(t) = \frac{1}{\eta} \quad (28)$$

Since, for the renewal process, the times between failures are i.i.d., analyzing data from a repairable system modeled by the renewal process is the same as inference for nonrepairable systems whose lifetimes are i.i.d.

The Piecewise Exponential Model

The piecewise exponential (PEXP) model was proposed by Sen and Bhattacharyya [11]. Under this model, the times between failure are independent exponential random variables, but each has a different mean, so times between failure are not i.i.d. Under their model, the mean of the i th time between failures is

$$E(X_i) = \frac{\delta}{\mu} i^{\delta-1}, \quad i = 1, 2, \dots \quad (29)$$

The parameter μ acts like a scale parameter, and the parameter δ measures the degree of deterioration. If $\delta = 1$, then the times between failure are i.i.d. exponential and therefore the process is the homogeneous Poisson process. If $\delta < 1$, then $E(X_i)$ is decreasing in i , indicating reliability improvement. If $\delta > 1$, then $E(X_i)$ is increasing in i , indicating reliability deterioration. Under the PEXP model, reliability improvement (or deterioration) is effected at the time of a failure, as compared to the NHPP where the reliability improvement (or deterioration) is continuous.

Inference for the PEXP can be found in Sen and Bhattacharyya [11] and Sen [12].

The Modulated Power Law Process

The modulated power law process (MPLP) is a compromise between the NHPP (same-as-old) and the renewal process (same-as-new) models. This model is a special case of the modulated gamma process of Berman [13] and was proposed as a model for the reliability of repairable systems by Lakey and Rigdon [14]. One way to develop the MPLP is to envision a power law process with parameters θ and β , except that a failure occurs not at every event in this power law process, but rather every κ th event, where κ is a positive integer. Under such a model, if $\beta > 1$, then there will be overall reliability deterioration, but after each failure, the subsequent repair would make the system better than it was just before the failure because the number of events would be reset to zero, and another κ events would be required before the next failure. Thus, a system can be deteriorating, but still be improved at the time of a failure/repair.

If the likelihood is developed under the above assumptions, it can be seen that this is a valid likelihood even when κ is not an integer. Another way of developing the MPLP (not shown here because it is rather complicated) does not make the assumption that κ is a positive integer, only that $\kappa > 0$.

Inference for the MPLP was developed by Black and Rigdon [15] and is summarized by Rigdon and Basu [2].

Inference for Multiple Systems

So far, we have assumed that inference has been for a single system. If there are multiple copies

of the system, the type of inference depends on the assumptions we are willing to make about the similarity of the systems. Here are a number of possibilities.

1. No assumptions

If no assumptions are made about the various systems, then data from each system should be analyzed separately.

2. All systems identical

If this is the case, then the likelihood is easily developed for most models described here. Crow [1] described how this is done for the power law process.

3. Systems with the same deterioration parameter, but possibly different scale parameters

In the context of the power law process, this would imply that all systems have the same β parameter, but different θ 's. This case was also described by Crow [1].

4. Systems with the same deterioration parameter, but different scale parameters that can be thought of as being selected from some joint (prior) distribution

Again, in the context of the power law process, the β parameter would be the same for all systems, but we would view $\theta_1, \theta_2, \dots, \theta_n$ as being a random sample (i.i.d.) from some distribution $g(\theta)$. At this point, we could apply an empirical Bayes method, or, if we place a prior on the parameters of $g(\theta)$ as well as β , we could apply a fully Bayesian analysis.

5. Systems with different deterioration parameters and different scale parameters, but each can be thought of as being selected from some (prior) distribution

For the power law process, this would mean that $(\theta_1, \beta_1), (\theta_2, \beta_2), \dots, (\theta_n, \beta_n)$ are selected from some (prior) distribution $g(\theta, \beta)$. Again, we could perform an empirical Bayes analysis or, if we place priors on all parameters, we would perform a fully Bayesian analysis.

Systems with Covariates

For systems whose failure times are observed along with covariates (or concomitant variables), we can use Poisson regression (*see Risk from Ionizing Radiation; Reliability Growth Testing; Cohort*

Studies; Repeated Measures Analyses; Competing Risks) to study the relationship between these covariates and the intensity function. Lawless [16] suggested a proportional intensity model (analogous to the proportional hazards model) that has the form

$$\lambda_{\mathbf{x}}(t) = \lambda_0(t)g(\mathbf{x}; \beta), \quad t > 0 \quad (30)$$

Here $\mathbf{x}$ is a vector of covariates and β is a vector of parameters. The function $\lambda_0(t)$ is called the *baseline intensity*. Since $\lambda_{\mathbf{x}}(t)$ must be a positive quantity, we require that $g(\mathbf{x}; \beta) > 0$ for all $\mathbf{x}$. A convenient and flexible choice for $g(\mathbf{x}; \beta)$ is

$$g(\mathbf{x}; \beta) = \exp(\mathbf{x}'\beta) \quad (31)$$

Lawless [16] considered a number of cases. If $\lambda_0(t)$ is an arbitrary function, then the method is called *semiparametric*. If $\lambda_0(t)$ has a parametric form, with some unknown parameters, then the method is called (fully) parametric. A common form for $\lambda_0(t)$ is the power law intensity function $\lambda_0(t) = \frac{\beta}{\theta} \left(\frac{t}{\theta}\right)^{\beta-1}$. Lawless also considers cases where there are random effects. He derived the likelihood equations for a number a number of different cases. Ciampi *et al.* [17] have developed a computer program to find the maximum-likelihood estimates of the regression parameters β and the parameters of the baseline intensity function.

Summary

Various models for the reliability of repairable systems have been discussed. Usually, the assumptions we make about the effect of a repair dictate the type of model that is selected. For analyzing multiple systems, the assumptions concerning the similarities of the various copies of the system lead us to a method of data analysis and inference. There are many other models for repairable systems. Ascher and Feingold [18] discusses a number of models that have not been covered here.

References

- [1] Crow, L.R. (1974). Reliability analysis for complex systems, in *Reliability and Biometry*, F. Proschan & R.J. Serfling, eds, SIAM, Philadelphia, pp. 379–410.
- [2] Rigdon, S.E. & Basu, A.P. (2000). *Statistical Methods for the Reliability of Repairable Systems*, John Wiley & Sons, New York.

- [3] Finkelstein, J.M. (1976). Confidence bounds on the parameters of the Weibull process, *Technometrics* **18**, 115–117.
- [4] Rigdon, S.E. (1989). Testing goodness-of-fit for the power law process, *Communications in Statistics A-18*, 4665–4676.
- [5] Guida, M., Calabria, R. & Pulcini, G. (1989). Bayes inference for a nonhomogeneous Poisson process with power law intensity, *IEEE Transactions on Reliability* **38**, 603–609.
- [6] Calabria, R., Guida, M. & Pulcini, G. (1990). Bayes estimation of prediction intervals for a power law process, *Communications in Statistics: Theory and Methods* **19**, 3023–3035.
- [7] Calabria, R., Guida, M. & Pulcini, G. (1992). A Bayes procedure for estimation of current system reliability, *IEEE Transactions on Reliability* **41**, 616–619.
- [8] Bar-Lev, S., Lavi, I. & Reiser, B. (1992). Bayesian inference for the power law process, *Annals of the Institute of Statistical Mathematics* **44**, 632–639.
- [9] Duane, J.T. (1964). Learning curve approach to reliability monitoring, *IEEE Transactions on Aerospace* **2**, 563–566.
- [10] Bain, L.J. & Engelhardt, M. (1980). Inferences on the parameters and current system reliability for a time truncated Weibull process, *Technometrics* **22**, 421–426.
- [11] Sen, A. & Bhattacharyya, G.K. (1993). A piecewise exponential model for reliability growth and associated inferences, in *Advances in Reliability*, A.P. Basu, ed, Elsevier, North Holland, pp. 331–355.
- [12] Sen, A. (1998). Estimation of current reliability in a Duane-based reliability growth model, *Technometrics* **40**, 334–344.
- [13] Berman, M. (1981). Inhomogeneous and modulated gamma processes, *Biometrika* **68**, 143–152.
- [14] Lakey, M. & Rigdon, S.E. (1992). The modulated power law process, *Transactions of the 46th Annual Quality Congress* 559–563.
- [15] Black, S.E. & Rigdon, S.E. (1996). Statistical inference for a modulated power law process, *Journal of Quality Technology* **28**, 81–90.
- [16] Lawless, J.F. (1987). Regression methods for Poisson process data, *Journal of the American Statistical Association* **82**, 808–815.
- [17] Ciampi, A., Dougherty, G., Lou, Z., Negassa, A. & Grondin, J. (1992). NHPPREG: A computer program for the analysis of nonhomogeneous Poisson process data with covariates, *Computer Methods and Programs in Biomedicine* **38**, 37–48.
- [18] Ascher, H. & Feingold, H. (1984). *Repairable Systems Reliability: Modeling, Inference, Misconceptions and their Causes*, Marcel Dekker, New York.

STEVEN E. RIGDON

Repeated Measures Analyses

Repeated measurements occur frequently in a variety of disciplines [1–4]. Repeated measures data emerge in studies where the same response variable is recorded on each subject under two or more conditions or occasions. They can be collected from experimental studies (e.g., controlled clinical trials) *see* **Randomized Controlled Trials** or from observational studies (e.g., epidemiological cohort studies; *see* **Cohort Studies**), either prospectively or retrospectively (e.g., using medical records).

Repeated measurements in risk assessment studies include profit and loss, environmental consequences, medical outcomes, and risk perception data [5]. Proper analysis of repeated measures risk data can help decision makers understand the impact of uncertainty and the consequences of different decisions (*see* **Scientific Uncertainty in Social Debates Around Risk**). For examples, repeated measures analysis has been applied in comparative risk assessment of environmental risk [6] and in the study of assessing longitudinal changes in mammographic density associated with breast cancer risk [7].

Repeated measurements can be quantitative or categorical, univariate or multivariate. In repeated measures experimental studies, the subjects are often assigned to the competing treatment groups using *randomization* procedures. The treatment for any particular group of subjects might remain constant throughout the study, or vary from one occasion to another, as in the case of a *crossover design*. Repeated measures data often arise in longitudinal studies where the measurements are made at different time points, and the number and the spacing of the repeated measurements may vary from one subject to another, consequently, resulting in unbalanced data. The repeated measures data may be unbalanced because of the study design or because of data which is unintentionally missing. Missing data are very common for studies involving humans since it is often difficult for every subject to follow an exact schedule.

The repeated measures design provides opportunities for investigators to examine the pattern of response variable under different measurement conditions or occasions, to understand the relationship between repeated outcomes and covariates, and to

discern the dependence pattern among the repeated measurements. A special feature of repeated measures data is that measurements from the same unit are usually correlated, and thus statistical analysis often requires assumptions concerning the within-subject correlation. The analysis methods often differ across disciplines, such as between medicine and psychology [2], owing to various assumptions and conventions, different study objectives and designs, quantitative or categorical responses, and different approaches to dealing with missing data [8].

There have been many recent advancements with new interest in analysis of repeated measures. Analytical approaches for repeated measurements have been developed ranging from the simple to the complex. This paper provides a review of the analysis methods for repeated measures, including simple graphic display techniques, univariate summary methods, analysis of variance for balanced normal responses, parametric random-effect approaches, marginal generalized estimating equation methods, and modern semiparametric and nonparametric approaches. For brevity, we focus on longitudinal studies. For interested readers, sources for more technical details and theoretical developments are excellent monographs by Crowder and Hand [9], Lindsey [10], Diggle *et al.* [11], Davis [12], Little and Rubin [8], Ruppert *et al.* [13], Fitzmaurice *et al.* [14], and many others.

Data Visualization and Graphical Display

Graphical display techniques are useful for investigators to examine and visualize important patterns hidden in the data [2]. While exploring repeated measures data, an effective graphical display may reveal both cross-sectional and longitudinal patterns. Some useful guidelines for exploring repeated measures data have been suggested in Diggle *et al.* [11]. Next, a few techniques for graphical display are illustrated using some real data examples.

Example 1 Vitamin E and weight growth of guinea pigs: Crowder and Hand [9] discussed a study of the effects of various doses of diet supplement vitamin E on the growth of guinea pigs. A total of 15 guinea pigs were randomly assigned to control, low dose, or high dose groups. The animals were given growth-inhibiting substance during week 1 and then vitamin E diet interventions were initiated at week 5. Body-weight measurements of each animal

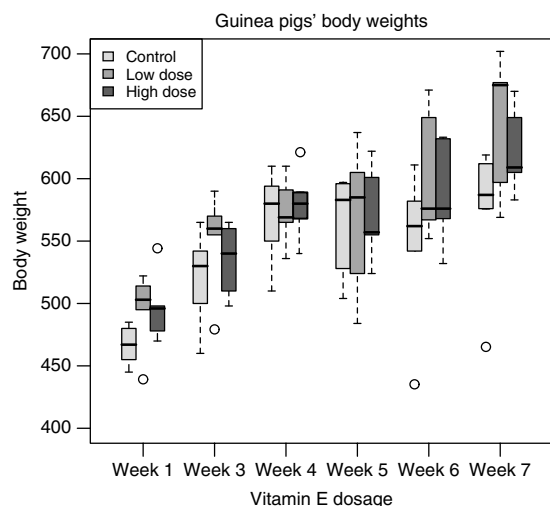


Figure 1 Box plots of body weights of guinea pigs for various Vitamin E dose groups at different time points

were made at the end of weeks 1, 3, 4, 5, 6, and 7. Box plots of the distribution of the measurements for each treatment group at each time point are presented in Figure 1. The side-by-side box plots at a specific time point show the cross-sectional differences of treatment groups; at the six time points, the three treatment box plots display the longitudinal pattern of body-weight growth. The box plot is effective and convenient to present the simple five-number summaries of quantitative data. It can be used in almost any available statistical package and also can give the indication of outliers (e.g., using circles) as illustrated in Figure 1.

Example 2 Halothane and blood pressure of rats with heart attack induction: Crepeau *et al.* [15] described a study on the effect of halothane on responses to irreversible myocardial ischemia and subsequent infarction. The study measured the blood pressure of 43 rats exposed to different concentrations of halothane (0, 0.25, 0.5, and 1%) after inducing heart attacks. The repeated measurements were observed at nine fixed time points, with some missing data. Using the available data, the averages of the blood pressure at each time point were calculated for four of the treatment groups; the results are presented in Figure 2. The curves of average blood pressure in different concentration groups are compared, and it is evident that the group with the

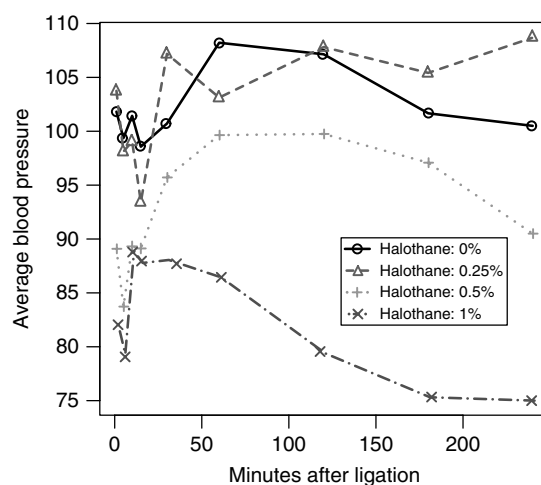


Figure 2 Mean profiles of blood pressure for four groups of rats treated with different doses of halothane

highest concentration of halothane shows the lowest blood pressures. However, one should be cautious that the values of average blood pressures in Figure 2 are calculated using the completely observed data only, and thus, such analysis may be biased (see more discussion in the section on missing data).

Example 3 Auranofin and self-assessment of arthritis: Fitzmaurice and Lipsitz [16] described a randomized trial on the effects of auranofin for patients with arthritis. A total of 51 patients were randomized into a treatment group or a placebo group. Each patient had at most five binary self-assessments on arthritis status (poor = 0, good = 1). The measurements were scheduled at baseline and subsequent weeks 1, 5, 9, and 13. In the treatment group with auranofin, the treatment changed from placebo to auranofin from week 5, and the placebo group remained on the same placebo treatment throughout the study. Because the response variable of self-assessment is categorical, the outcomes at each time point for each group are summarized using the proportion of good self-assessments, and the summary is presented using bar plots. In addition, twice the standard errors of the proportions are plotted to show variability of the proportions. In Figure 3, the side-by-side bar graphs show how the proportions of patients with good arthritis self-assessments vary over the five occasions, and different groupings of bars highlight the differences in male and female patients in the treatment groups.

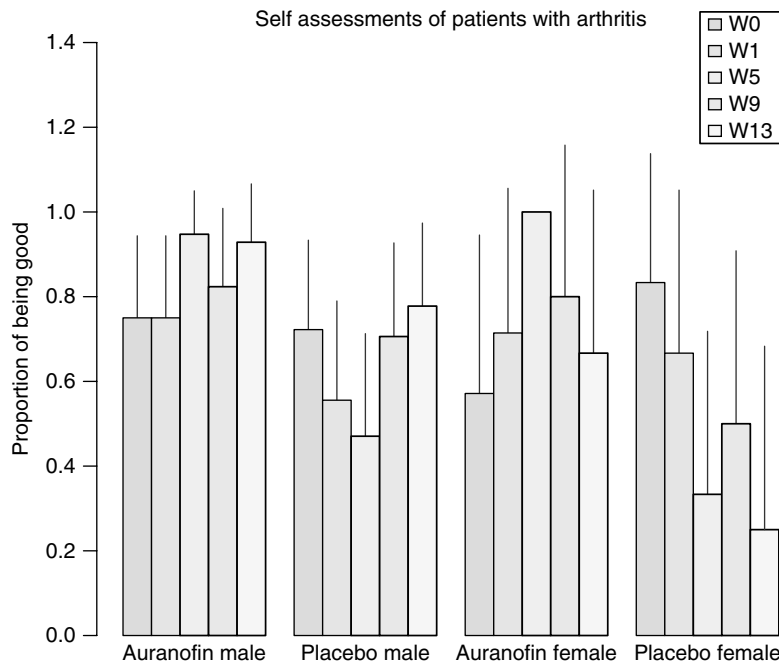


Figure 3 Bar plots of proportions of self-assessments of arthritis patients

Univariate Summary Methods

In some cases, a simple and effective way to explore repeated measures data is to condense repeated measurements into some summary measurements, and then to model each summary variable (e.g., mean, median, range, change, or slope). This approach is also called *two-stage or derived-variable analysis*. The major task of this approach is to select one or more powerful and representative summary statistics to best characterize the main features of repeated measurements. Matthews *et al.* [17] provide a list of potentially useful summary measures according to the study objectives. For example, in the case of *crossover* studies, the average of two measurements may be used to determine the better order of administration of treatments. In the case of longitudinal data showing a linear trend, the individual intercept and slope may be representative of overall information in the data. One may fit a separate regression model using each individual's data, and the resulting estimates can be used as the derived variables for further analysis. In addition, this approach suggests the idea of using subject-specific random effects to explain the heterogeneity across the subjects (see the section

titled “Mixed-Effects Models”). The use of summary measures in the analysis of repeated measures data has several advantages as discussed in [2, 17]. However, the derived-variable method can be inefficient with highly unbalanced data and unfeasible when the covariates of interest change over measurement occasions.

Methods for Balanced Repeated Measures Data

Traditional multivariate methods, such as Hotelling T^2 and profile analysis can be used to analyze balanced repeated measures data if they are normally distributed (or after appropriate transformations). The Hotelling T^2 , an analog of univariate t -test, can test whether mean vectors are different between two groups. The multivariate normal distribution with an unstructured covariance matrix is a common method for modeling the joint distribution of repeated measures, especially in modeling growth curves and in psychology [18, 19]. The unstructured multivariate approach does not require any assumption on the covariance matrix at the cost of many degrees of

freedom in estimating covariance parameters. When the dimension of multivariate data is large, this can lead to low power to detect mean differences of interest.

Another proposal for modeling correlated multivariate Gaussian responses is the multivariate analysis of variance (MANOVA). The MANOVA is a multivariate analog of univariate ANOVA, and assumes an additive model where the response variable is related to a set of discrete covariates and the subject-specific random effects. Thus, the models assume a structured covariance matrix, in which the correlation among repeated measurements is assumed to attribute to a subject-specific random effect. The MANOVA models focus on comparing mean responses of units along different measurement occasions.

When normality of the repeated measures data is not assumed, nonparametric methods and other techniques may be used. Nonparametric alternatives that require less model assumptions include the Wilcoxon signed-rank test for the case of two repeated measures per subject and the Friedman test to analyze the treatment effect in the randomized complete block design. Also, it is practically rare to find balanced data sets in longitudinal studies and the response variable can be continuous or discrete. Therefore, it is necessary to use some alternative techniques that are essentially regression models to handle unbalanced data and relax distributional assumptions.

Methods for Unbalanced Repeated Measures Data

Unbalanced repeated measures data arise when the number of repeated measures per subject is not constant and/or the measurement timing varies across subjects. In addition, the response variable can be either continuous or categorical, and the explanatory variables can be either time independent (e.g., baseline) or time dependent. Various extensions of the univariate *linear regression models*, *generalized linear models*, and *quasi-likelihood* methods have been proposed for the analysis of unbalanced repeated measurements. In this article, we will classify the analysis methods for unbalanced data into three categories [20]—mixed-effects models, marginal models, and transition models.

Mixed-Effects Models

One can view the mixed-effects model as a generalization of the univariate (generalized) *regression model* with correlated errors. The linear mixed-effects model allows explicit specifications on the association between the responses and population-level factors and subject-specific covariates. It can accommodate complex measurement patterns in the observation of repeated measures. In mixed-effects models, the subject-specific random-effects account for between-subject heterogeneity. More specifically, let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{im_i})$ be the response of subject i at m_i measurement occasions, $i = 1, \dots, n$. The mixed-effects model is specified as follows. Given a $q \times 1$ vector of the subject-specific random effects, b_i , the conditional mean of the response Y_{ij} at the j th occasion, is given by

$$g\{E(Y_{ij}|b_i, X_{ij}, Z_{ij})\} = \beta'X_{ij} + b_i'Z_{ij} \quad (1)$$

where $g(\cdot)$ is a monotone continuous link function, and β denotes a $p \times 1$ vector of fixed effects. The covariates of interest, X_{ij} and Z_{ij} , can be identical, overlapped, or mutually exclusive. In addition, the responses are assumed to be independent, conditioning on the random effects. The random effects, b_i , are assumed to be independent and identically distributed with a parametric distribution function.

For continuous responses, the link function is often set to be an identity function and $\text{Var}(Y_{ij}|b_i) = \sigma_e^2$. Thus, equation (1) becomes a linear mixed-effects model [21–23]. Assuming normality for the random effects and error terms, the likelihood function for Gaussian repeated measures data can be expressed in an explicit form. One can estimate the regression parameters and covariance parameters using the *maximum-likelihood* (ML) method or the restricted maximum-likelihood (REML) approach [21, 24]. For general categorical responses with nonlinear link functions, evaluation of the likelihood requires numerical methods in most cases (e.g., see Gaussian quadrature method or the first-order Taylor series approximation). Breslow and Clayton [25] proposed an approximation to the integrand so that its integrals of the likelihood function have closed forms. This method gives effective estimates of the fixed effects. However, it can yield biased results for the parameters associated with the distribution of random effects.

Marginal Models

When the objective of the study is focused on the marginal, population-averaged associations between responses and covariates, marginal models can directly address this aim. The generalized linear regression models specify the marginal mean as

$$g\{E(Y_{ij}|X_{ij})\} = g\{\mu_{ij}\} = \beta'_M X_{ij} \quad (2)$$

where $g(\cdot)$ is a known link function, and β_M characterizes how the cross-sectional response distribution depends on the covariates. In this context, the marginal regression coefficients are of primary interest. Semiparametric and parametric approaches have been proposed for the inference of these marginal models.

The Generalized Estimating Equations Approach.

Liang and Zeger [26] proposed a generalized estimating equation (GEE1) approach, which can be viewed as a multivariate generalization of the quasi-likelihood estimation method. The GEE method is a flexible and powerful tool in the analysis of repeated measures data, especially for categorical responses. The method is semiparametric in nature because it does not require a full specification of the joint distribution of the multivariate responses. The approach yields consistent estimates as long as the marginal mean model is correctly specified, regardless of specifications of the within-subject associations.

More specifically, the generalized estimating equation for β_M is specified as

$$U(\beta) = \sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta_M} \right)' [V_i(\alpha)]^{-1} (\mathbf{Y}_i - \mu_i) = 0 \quad (3)$$

where $\mathbf{Y}_i$ and μ_i are vectors consisting of corresponding components of Y_{ij} and μ_{ij} , $V_i(\alpha) = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$, ϕ is a variance scale parameter, $R_i(\alpha)$ is a working correlation matrix, and A_i is a diagonal matrix consisting of marginal variance of Y_{ij} . Liang and Zeger [26] proposed to replace ϕ and α by their consistent estimates and developed an iterative algorithm to estimate β_M . The GEE is constructed on the basis of the first sample moment, so that the inferences about β_M are insensitive to the specification of V_i . Several choices for the working correlation matrix have also been suggested by Liang and Zeger [26]. The simplest choice is the independent working correlation matrix. It works reasonably

well when the number of subjects is large relative to the number of repeated observations since the correlation influence can be small in this situation. Other choices include exchangeable working correlation matrix, the first-order autoregressive model, and unstructured model.

Zhao and Prentice [27] extended the original GEE method to allow the joint estimation of the mean and covariance parameters (GEE2). When both the marginal mean model and covariance structure are correctly specified, the GEE2 approach leads to more efficient estimates; otherwise, it can be biased.

The Marginalized Likelihood Approaches

The GEE method described above has some advantages in robustness since it does not require the correct specification on the within-subject correlations. But sometimes, its efficiency can be a concern. The marginalized likelihood approaches not only specify the marginal mean regression model, similar to the one in the GEE method, but also specify the joint multivariate distribution through higher-order assumptions [28–32]. Therefore, one can carry out efficient likelihood-based inference and individual level prediction for repeated measures data when the target of inference is on the marginal, population-averaged covariate effects. In addition, the marginal inference is insensitive to the higher-order model misspecifications [33]. However, the marginalized likelihood approaches can be computationally intensive owing to the iterative conversion between the marginal models and higher-order models, and the evaluation of the likelihood function.

Semiparametric Regression Models

The models described above assume that the mean model has a parametric linear regression expression. In practice, however, the parametric linear assumption may be restrictive, and the analysis is susceptible to misspecifications of the parametric model. In the semiparametric and nonparametric regression models of Lin and Ying [34] for continuous longitudinal data, the marginal means of continuous responses are directly modeled using a semiparametric regression model,

$$E\{Y_i(t)|X_i(t)\} = \alpha_0(t) + \beta' X_i(t) \quad (4)$$

where $\alpha_0(t)$ is an unspecified baseline function. The model does not require any assumption on the stochastic distribution of the responses and has also been further generalized to accommodate *time-varying* coefficients, i.e., $E\{Y_i(t)|X_i(t)\} = \alpha_0(t) + \beta'(t)X_i(t)$. In addition, the method of Lin and Ying can allow the recurrent observation process to be dependent on covariates that can be incorporated using a semiparametric multiplicative rate model [35]. Discussions on efficient semiparametric marginal estimation models for longitudinal data can be found in [36–38]. Other general semiparametric regression models for various types of repeated measures data are discussed in Ruppert *et al.* [13].

Transition Models

Transition models for repeated measures characterize the conditional distribution of each response Y_{ij} as an explicit function of covariates and the history of $(Y_{i1}, \dots, Y_{i(j-1)})$. For example, a conditional generalized linear model can be specified as

$$\begin{aligned} g\{E(Y_{ij}|Y_{i1}, \dots, Y_{i(j-1)}, X_{ij})\} \\ = \beta'X_{ij} + \gamma'f(Y_{i1}, \dots, Y_{i(j-1)}) \end{aligned} \quad (5)$$

where $f(\cdot)$ is a known function of the past observations. One useful transition model is the k -order Markov model, where the conditional distribution of the current response depends only on the k prior observations. Tsay [39] proposed the Markov linear model for continuous responses. For binary responses, Korn and Whittemore [40] and Zeger *et al.* [41] proposed a logistic first-order Markov chain model. Zeger and Qaqish [42] developed a first-order Markov chain Poisson model for count data. Transition models are similar to the *autoregressive* models for *time-series* data. Note that repeated measures data usually consist of a large number of shorter sequences, whereas time-series data typically arise from a single or a few long extending sequences.

Comparisons among the Three Types of Models for Unbalanced Data

Under the situation of linear models for continuous responses, the above-discussed three classes of models can yield the same type of interpretation of the regression coefficients. Coefficients from the linear

mixed-effects models and the transition models can both give the marginal interpretations. However, in the case of categorical response with a nonlinear link function, these three models yield different interpretations for the regression coefficients such that each may be suitable for a specific objective.

The regression coefficients of the mixed-effects models have a subject-specific interpretation; that is, the conditional effects of the explanatory covariates conditioning on the subject-specific random effects. This type of model is more appropriate when the scientific focus of the repeated measures study is on the individual's responses. Nevertheless, it can be awkward for interpreting the population effects. For example, it can be difficult to interpret gender difference given the same values of the random effects, because it is unnatural to assume a man and a woman have the same values of the random effects.

The regression coefficients in transition models have an interpretation of the effects of covariates adjusted for the individual's response history. The transition models are more appropriate when the target of inference is on the stochastic process of the repeated measurements given the covariates and the history, but they can be inappropriate for the estimation of population-level and time-independent covariate effects.

In contrast, marginal models yield interpretations of the regression coefficients as the population-averaged effects. They are suitable for studies focusing on the population-level inference, such as comparing treatment groups. In addition, marginal models separate the specification for marginal mean regression models from the assumption on the associations among repeated measures. Hence, the interpretation of regression parameters is invariant with respect to different assumptions for the within-subject correlations. On the other hand, mixed-effects models and transition models specify the covariate effects and within-subject association in one model. In summary, it is useful to consider the differences among various models when choosing appropriate models for practical applications.

Analysis Methods for Missing Data

Repeated measures data are often incomplete for reasons such as scheduled measurements not being taken or not being available. Missing measurements in the

repeated measures data result in unbalanced data and entail technical difficulties to statistical analysis. The missing pattern can be classified into two broad groups as monotone missing and intermittent missing. If the missingness occurs in the monotone fashion, all subsequent observations would be missing, e.g., patient dropouts or death in clinical trials. Intermittent missing patterns include situations where the subjects return to the study after missing one or more scheduled observations. It is also important to raise the deeper conceptual questions on the reasons for the missing measurements. The missingness is often classified into three categories: missing completely at random; missing at random; and nonignorable missing as discussed in Little and Rubin [8].

The simplest analysis to handle repeated measures data with missing measurements is to only use the subjects with complete observations. However, this approach obviously can waste a large amount of information by discarding all incomplete observations. More seriously, it may potentially introduce bias if the missing mechanism is related to the responses. Another way to handle repeated measures data with missingness is the *last observation carried forward*, which is an imputation method by extrapolating the last observed measurements for all subsequent missing observations. This method is primarily used for testing the null hypothesis of no difference between treatment groups. It is argued that, under the null hypothesis, the subsequent extrapolation of the last observed values represents the inherent feature of random outcome of a given subject and thus, the test remains valid without requiring an explicit modeling assumption.

Most statistical methods remain valid when data are missing completely at random. Thus, it is of interest to test whether the data missing mechanism is completely at random. Diggle [43] proposed a method to test for monotone dropout completely at random. Chen and Little [44] developed tests for general missing completely at random. Likelihood-based inference approaches without modeling assumptions on missing mechanism are usually valid for data missing at random. However, statistical inference for nonignorable missing data is a challenge. Indeed, to obtain valid estimates about the complete responses, the relationship between responses and the missing mechanism must be specified. One must also note that any assumption made about nonignorable missingness is, in general, unverifiable using the observed

data. Without external information about the underlying missing reasons, the observed data provide no information for determining the validity of assumptions on nonignorable missing data. In the literature, three broad classes of models have been extensively discussed for repeated measures with nonignorable missing: selection models [45–47]; pattern-mixture models [48–50]; and latent variable models [51–53]. Some authors have also provided excellent overviews of recent work in this area [54, 55].

Statistical Software and Summarizing Remarks

For the analysis of repeated measures data, most of the currently available statistical software are adequate for simple graphical displays, analysis at individual time points, or using the summary statistic approach. Also, a large number of statistical packages have the capacity of performing the analysis of variance procedures, the mixed-effects model, and the generalized estimating equation model. It is worth mentioning that, recently, more and more statistical algorithms developed using “R” are freely available, providing the capacity to fit complex statistical models that allow for missing data and a variety of covariance structures [56, 57].

In summary, repeated measures arise frequently in many studies. Many methods for statistical analysis have been suggested including simple *t*-tests at each separate occasion, summary statistics, multivariate analysis of variance, parametric random-effects methods, generalized estimating equation approaches, or modern semiparametric and nonparametric models. The field is still evolving rapidly, however, it is safe to say that no single “optimal method” is applicable to all cases, simply because such repeated measures can be collected in many settings, with various study designs, and different aims. A number of analysis methods are discussed in this article, but a thorough review of the field is beyond the scope of this article. More references and technical details can be found in the many cited recent monographs and review papers.

Acknowledgments

The authors would like to thank the Stony Wold–Herbert Fund for financial support.

References

- [1] Singer, J.D. & Willett, J.B. (2003). *Applied Longitudinal Data Analysis : Modeling Change and Event Occurrence*, Oxford University Press, Oxford.
- [2] Everitt, B.S. (1995). The analysis of repeated-measures – a practical review with examples, *Statistician* **44**(1), 113–135.
- [3] Gulati, R. (1995). Social structure and alliance formation patterns: a longitudinal analysis, *Administrative Science Quarterly* **40**(4), 619–652.
- [4] Albert, P.S. (1999). Longitudinal data analysis (repeated measures) in clinical trials, *Statistics in Medicine* **18**(13), 1707–1732.
- [5] McNeil, A.J., Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton Series in Finance, Princeton University Press, Princeton.
- [6] Ding, C.G., Woo, Y.Y., Sheu, H.J., Chien, H.C. & Shen, S.F. (1996). An effective statistical approach for comparative risk assessment, *Risk Analysis* **16**(3), 411–419.
- [7] Vachon, C.M., Pankratz, V.S., Scott, C.G., Maloney, S.D., Ghosh, K., Brandt, K.R., Milanese, T., Carston, M.J. & Sellers, T.A. (2007). Longitudinal trends in mammographic percent density and breast cancer risk, *Cancer Epidemiology, Biomarkers and Prevention* **16**(5), 921–928.
- [8] Little, R.J.A. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, 2nd Edition, John Wiley & Sons, New York.
- [9] Crowder, M.J. & Hand, D.J. (1990). *Analysis of Repeated Measures*, Monographs on Statistics and Applied Probability 41, 1st Edition, Chapman & Hall, London.
- [10] Lindsey, J.K. (1993). *Models for Repeated Measurements*, Oxford Statistical Science Series 10, Clarendon Press, Oxford University Press, Oxford.
- [11] Diggle, P., Heagerty, P.J., Liang, K.-Y. & Zeger, S.L. (2002). *Analysis of Longitudinal Data*, Oxford Statistical Science Series 25, 2nd Edition, Clarendon Press; Oxford University Press, Oxford.
- [12] Davis, C.S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*, Springer Texts in Statistics, Springer, New York.
- [13] Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semi-parametric Regression*, Cambridge University Press, Cambridge.
- [14] Fitzmaurice, G.M., Laird, N.M. & Ware, J.H. (2004). *Applied Longitudinal Analysis*, Wiley Series in Probability and Statistics, Wiley-Interscience, Hoboken.
- [15] Crepeau, H., Koziol, J., Reid, N. & Yuh, Y.S. (1985). Analysis of incomplete multivariate data from repeated measurement experiments, *Biometrics* **41**(2), 505–514.
- [16] Fitzmaurice, G.M. & Lipsitz, S.R. (1995). A model for binary time-series data with serial odds ratio patterns, *Journal of the Royal Statistical Society, Series C: Applied Statistics* **44**(1), 51–61.
- [17] Matthews, J.N., Altman, D.G., Campbell, M.J. & Royston, P. (1990). Analysis of serial measurements in medical research, *British Medical Journal* **300**(6719), 230–235.
- [18] Rao, C.R. (1965). Theory of least squares when parameters are stochastic and its application to analysis of growth curves, *Biometrika* **52**, 447–458.
- [19] Potthoff, R.F. & Roy, S.N. (1964). Generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika* **51**(3–4), 313–326.
- [20] Ashby, M., Neuhaus, J.M., Hauck, W.W., Bacchetti, P., Heilbron, D.C., Jewell, N.P., Segal, M.R. & Fusaro, R.E. (1992). An annotated bibliography of methods for analysing correlated categorical data, *Statistics in Medicine* **11**(1), 67–99.
- [21] Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**(4), 963–974.
- [22] Ware, J.H. (1985). Linear-models for the analysis of longitudinal-studies, *American Statistician* **39**(2), 95–101.
- [23] Mclean, R.A., Sanders, W.L. & Stroup, W.W. (1991). A unified approach to mixed linear-models, *American Statistician* **45**(1), 54–64.
- [24] Jennrich, R.I. & Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices, *Biometrics* **42**(4), 805–820.
- [25] Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**(421), 9–25.
- [26] Liang, K.Y. & Zeger, S.L. (1986). Longitudinal data-analysis using generalized linear-models, *Biometrika* **73**(1), 13–22.
- [27] Zhao, L.P. & Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model, *Biometrika* **77**(3), 642–648.
- [28] Fitzmaurice, G.M. & Laird, N.M. (1993). A likelihood-based method for analyzing longitudinal binary responses, *Biometrika* **80**(1), 141–151.
- [29] Molenberghs, G. & Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate Plackett distribution, *Journal of the American Statistical Association* **89**(426), 633–644.
- [30] Heagerty, P.J. & Zeger, S.L. (1996). Marginal regression models for clustered ordinal measurements, *Journal of the American Statistical Association* **91**(435), 1024–1036.
- [31] Heagerty, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data, *Biometrics* **55**(3), 688–698.
- [32] Heagerty, P.J. & Zeger, S.L. (2000). Marginalized multi-level models and likelihood inference, *Statistical Science* **15**(1), 1–19.
- [33] Heagerty, P.J. & Kurland, B.F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models, *Biometrika* **88**(4), 973–985.

- [34] Lin, D.Y. & Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data, *Journal of the American Statistical Association* **96**(1), 103–113.
- [35] Lin, D.Y., Wei, L.J., Yang, I. & Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events, *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **62**, 711–730.
- [36] Lin, X.H. & Carroll, R.J. (2001). Semiparametric regression for clustered data, *Biometrika* **88**(4), 1179–1185.
- [37] Lin, X.H. & Carroll, R.J. (2001). Semiparametric regression for clustered data using generalized estimating equations, *Journal of the American Statistical Association* **96**(455), 1045–1056.
- [38] Wang, N., Carroll, R.J. & Lin, X.H. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data, *Journal of the American Statistical Association* **100**(469), 147–157.
- [39] Tsay, R.S. (1984). Regression-models with time-series errors, *Journal of the American Statistical Association* **79**(385), 118–124.
- [40] Korn, E.L. & Whittemore, A.S. (1979). Methods for analyzing panel studies of acute health-effects of air-pollution, *Biometrics* **35**(4), 795–802.
- [41] Zeger, S.L., Liang, K.Y. & Self, S.G. (1985). The analysis of binary longitudinal data with time-independent covariates, *Biometrika* **72**(1), 31–38.
- [42] Zeger, S.L. & Qaqish, B. (1988). Markov regression-models for time-series – a quasi-likelihood approach, *Biometrics* **44**(4), 1019–1031.
- [43] Diggle, P.J. (1989). Testing for random dropouts in repeated measurement data, *Biometrika* **45**(4), 1255–1258.
- [44] Chen, H.Y. & Little, R. (1999). A test of missing completely at random for generalised estimating equations with missing data, *Biometrika* **86**(1), 1–13.
- [45] Diggle, P. & Kenward, M.G. (1994). Informative drop-out in longitudinal data-analysis, *Applied Statistics-Journal of the Royal Statistical Society, Series C* **43**(1), 49–93.
- [46] Baker, S.G. (1995). Marginal regression for repeated binary data with outcome subject to non-ignorable nonresponse, *Biometrics* **51**(3), 1042–1052.
- [47] Fitzmaurice, G.M., Laird, N.M. & Zahner, G.E.P. (1996). Multivariate logistic models for incomplete binary responses, *Journal of the American Statistical Association* **91**(433), 99–108.
- [48] Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data, *Journal of the American Statistical Association* **88**(421), 125–134.
- [49] Little, R.J.A. (1994). A class of pattern-mixture models for normal incomplete data, *Biometrika* **81**(3), 471–483.
- [50] Hogan, J.W. & Laird, N.M. (1997). Mixture models for the joint distribution of repeated measures and event times, *Statistics in Medicine* **16**(1–3), 239–257.
- [51] Wu, M.C. & Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process, *Biometrics* **44**(1), 175–188.
- [52] Schluchter, M.D. (1992). Methods for the analysis of informatively censored longitudinal data, *Statistics in Medicine* **11**(14–15), 1861–1870.
- [53] Wulfsohn, M.S. & Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error, *Biometrics* **53**(1), 330–339.
- [54] Kenward, M.G. & Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data, *Statistical Methods in Medical Research* **8**(1), 51–83.
- [55] Hogan, J.W., Roy, J. & Korkontzelou, C. (2004). Handling drop-out in longitudinal studies, *Statistics in Medicine* **23**(9), 1455–1497.
- [56] Everitt, B. & Hothorn, T. (2006). *A Handbook of Statistical Analyses Using R*, Chapman & Hall/CRC, Boca Raton.
- [57] Pinheiro, J.C. & Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*, Statistics and Computing, Springer, New York.

MENGLING LIU AND YONGZHAO SHAO

Risk and the Media

The news media helps to shape public discourse surrounding risk issues by its own portrayal and representations of risks and uncertainties in news stories [1]. The public typically receives a variety of information through the news media – from environment and health risks and benefits, to advances in science and technology, and government policy decisions. At the same time, the media plays a complex role in the portrayal of risks and associated uncertainties [1–4]. Media reporting can influence the process by which a risk becomes more prominent or “amplified” by the public [3, 5]. Often as the frontline popular communicators of science, the media takes the necessary scientific information from the risk assessors and makes it accessible to the public [5]. Exploring the processes of how the media communicate and present risk information is important (*see Considerations in Planning for Successful Risk Communication*).

Shaping Agendas: How the Media Report Risk Issues

Prominently covered health issues regularly include discussions of risk [6], whereas a “risk” category is often incorporated when reporting environmental issues [7]. When both health and environmental risks are involved, these typically appear as front page news [8], elevating the importance of the risk(s) being presented. However, little is known about how the media selects and structures risk stories and how individuals use these messages to form their own opinions [7].

The mass media has been recognized as one of the key players (and gatekeepers) that help build [9] and shape a policy agenda for those issues that gain public attention and those that do not [10]. Its influence is often short-lived because the media frequently jumps from issue to issue [11]. Issues being presented in the media have ways of grabbing public attention when they may have been virtually ignored weeks or months before. Focusing events (e.g., *via* public opinion polls, media attention to a risk event, etc.) can prompt an issue to move higher up a political agenda, or at least register on the public or political “radar screen”, for a short period of time [12]. Often, it is through the use of language or symbols that

issues can be manipulated on or off the agenda [13]. A common finding in agenda-setting research is that “issues emerge and recede from the public agenda without important changes in the nature of the issues themselves” [12, p. 47]. In other words, even when the issue is not resolved, the attention dedicated to the issue, by the media, the public, or decision makers, may wane.

The media, government (represented by policy agendas), and public agendas have varying influences over one another. The extent of media influence on public opinions can be assumed by the length of time that an issue appears on the media’s agenda. The longer the issue appears in the media, the more likely it is that the public views it as an important issue. Although the public has some influence over what appears on the policy agenda, the converse is less true. The government’s influence over the public has long been noted as coming *via* the media [14]. However, in light of technological advancements, like the internet, the public can now be directly informed of policy decisions, without interpretation from the media. In this manner, the government has developed some influence on the public’s agenda.

Examining of the agenda-setting role of the mass media has been documented in the literature, most notably in political studies about voter preferences during elections (e.g. [15]), in studies of environment and/or health risk issues (e.g. [6, 16, 17]; *see Environmental Health Risk*), and in edited volumes on the role of the media in communicating risks and science to a public audience (e.g. [18]). In summarizing a 25-year evolution of agenda-setting issues, McCombs and Shaw [15] argue that a common thread within these different studies exists: the media not only tells us *what to think about* but may also influence *how we think about it*. In reporting some components of a story over others, journalists and editors portray the *newsworthiness* of a story to the public [19]. The media’s tendency to present the novel or dramatic [3, 5, 7, 16] over the more common (and often more serious) risks [3] only serves to further distance expert *versus* lay understandings of risk issues.

Slovic [2] indicates that few journalists have the scientific background required to make sense of the wealth of data that is available to them to understand risks. This can result in scientific uncertainties or limitations being inadequately communicated to the public, particularly when “quick” answers are required

[20]. Few studies have been done to determine the extent to which journalists understand the science behind a risk, or probability theory, which is the process by which risks are assessed [7]. As the public seeks out additional information, especially about health risks, it is possible it will become more media-dependent [14]. This is particularly dominant if its only “understandable” source of information is through the media [17]. Dunwoody and Peters [16] argue that when specifically looking at risk research, the media reflects less of what the “experts” consider objective research, but more of the political and social processes that highlight what is occurring at that time. This may have a possible agenda-setting influence based on what the media is presenting [21].

The Media as Amplifiers of Risk

Public perception or understanding of a specified risk is influenced by both media presentation of a risk and how dependent the public is on the media for information. Kasperson and colleagues [3] argue that the media act as amplifiers of risk, or “stations”, constructing and communicating information about that particular risk. First introduced in 1988 by Kasperson and colleagues, the social amplification of risk framework is designed to bring together risk perception and risk communication research [3].

Bennett [22] summarizes the key “triggers” that will capture media (and hence public) attention: blame, conspiracy stories, identifiable victims where many people are exposed, embroiled in conflict, links to other high-interest stories, has a high signal value (portends future ills), is strongly visual, and may be linked to sex or crime. These triggers could amplify a risk, particularly when more than one is present.

The amplification causes a rippling effect, where the risk may be augmented throughout society on the basis of how stories are presented. The media serves as a channel of communication, whereby the information about that risk, and thus its mediated interpretation, flows through [3]. Frewer [23] suggests the framework can also indicate why a small risk with little physical impacts attracts substantial attention when a larger risk with considerable impacts does not. Frewer [23] argues that the extent of media influence over the amplification, or the rippling process, may be related to how the public trusts the information source and the information itself. Thus, greater trust in the source can intensify believability

of the nature of risk, whereas less trust in the source can attenuate the belief about the hazard.

The way in which risk *events* are covered by the media is equally important. Termed as *what-a-story* [24], it occurs when an issue is so consequential, dramatic, and unusual that media outlets drop everything else to report on this event [25]. When the story breaks, it is broadcast throughout the region, state, and possibly country. “What-a-story” events often contain the elements needed to create a satisfying media story – villains, victims, heroes, and tragic circumstances – where one can imagine the verbal and nonverbal gestures and excitement that accompany the announcement of a story in a newsroom [24]. When a major risk issue occurs, the typical reporting pattern is one of greater coverage within the first 10 days [26]. The media tends to provide blanket coverage of the event, such as was found during the media coverage of the Walkerton, Ontario, Canada, contamination of local drinking water supplies that resulted in 2300 people becoming ill and seven deaths, in May 2000 [27]. This initial reporting period plays an important role in establishing a mental image of the risk issue for the general public [26], and could also serve to amplify risk. After this 10-day period, the number and frequency of stories typically drops.

The Use of Frames by Journalists

Frames provide a way of packaging a story so that specific themes or subject matter are emphasized to help define what the issue is really about [28]. The ways in which stories are framed may have an impact on the ways in which the public interprets the story [29]. Frames act as a powerful agent in manipulating information that is disseminated to the public. It is important that risk assessors or risk managers are aware of how their communication of a risk may influence the story frames used by the media.

For example, Leiss [30] found that the initial frames used to describe dioxins (*see Dioxin Risk*), as chemical warfare, back in the early 1970s, put dioxins at the foreground of public attention as a very harmful substance. Over time, this initial framing remained dominant in media stories about dioxins forming a boundary around what aspects are (or not) discussed [31]. A study conducted by Singer and Endreny [32] illustrates that a rare hazard is more newsworthy than a common hazard; a new hazard is more newsworthy

than an old one; and a dramatic hazard that kills many people mysteriously is more newsworthy than one of chronic or familiar illnesses. Similarly, Driedger and Eyles [31] argue that in order to generate a sense of “outrage” and to receive the greatest level of resonance, for a risk story, they must be framed both at the macro level (statistical risk) and at the micro level (individual level).

Both Dunwoody [7] and Potter [33] suggest several modalities by which journalists frame their stories. Dunwoody [7] indicates it is not the job of the journalist to educate the public. A journalist’s role is to inform, but not necessarily with comprehensive details. Historically, journalists have presented the “facts” with very little interpretation about what these facts mean. It was only in the twentieth century that presenting the facts was considered insufficient: to provide the bigger picture facts needed to be contextualized or interpreted for public consumption [34]. Potter [33] suggests that journalists frame their stories so that they are easily recognizable to their audience members. By making the necessary, and purposeful, selections of the information they wish to present, journalists may be able to manipulate the readers to incorporate only the presented items in their understanding of the issue.

Framing is not without its problems. Selection processes limit precise accounts of a risk, thereby making it difficult for the public and decision makers to plan on how to deal with the risk. Social scientists have documented media coverage being often at odds with the scientific assessments of risk. Slovic [2] indicates that content analysis of media reporting on specific hazards has shown a great deal of misinformation and content distortion. As such, the public receives information that is not completely value-free or information rich. For example, there have been several studies concerning media coverage of the accident at Three Mile Island. Repeated stories about Three Mile Island redirects public attention away from competing news stories [5, 35]. At the same time as Three Mile Island, one of the worst airline accidents in American history (at that time) occurred, killing 274 people. Media focus was on Three Mile Island [35], mobilizing fears about risks associated with nuclear power and the possibility of similar nuclear accidents occurring more frequently elsewhere in the country [5].

Summary

It is important for risk assessors to be aware of how risk estimates may be presented in the news. Media outlets (reporters, editors, and owners) play a role in building and shaping public and policy agendas. Media presentations of risk could influence how risk issues are amplified in a public domain. As a result, it is important for risk communication strategies to have a user-based approach such that decisions are made to present what needs to be communicated to address lay concerns based on an understanding of what knowledge is held by a public audience. In other words, it is important to determine what a public knows and balance this against what a communicator needs them to know – not necessarily what a communicator wants to tell them.

References

- [1] Beck, U. (2000). Foreword, in *Environmental Risks and the Media*, S. Allan, B. Adam & C. Carter, eds, Routledge, New York, pp. xii–xiv.
- [2] Slovic, P. (2000). Informing and educating the public about risk, in *The Perception of Risk*, P. Slovic, ed, Earthscan Publications, Sterling, pp. 182–198.
- [3] Kasperson, R.E., Renn, O., Slovic, P., Brown, H.S., Emel, J., Goble, R., Kasperson, J.X. & Ratick, S. (2000). The social amplification of risk, in *The Perception of Risk*, P. Slovic, ed, Earthscan Publications, Sterling, pp. 232–245.
- [4] Berry, D.C. (2004). *Risk, Communication and Health Psychology*, Open University Press, New York.
- [5] Kasperson, R.E., Jhaveri, N. & Kasperson, J.X. (2001). Stigma and the social amplification of risk: toward a framework of analysis, in *Risk, Media and Stigma: Understanding Public Challenges to Modern Science and Technology*, J. Flynn, P. Slovic & H. Kunreuther, eds, Earthscan Publications, Sterling, pp. 9–27.
- [6] Anderson, A. (1997). *Media, Culture and the Environment*, UCL Press, Bristol.
- [7] Dunwoody, S. (1992). The media and public perceptions of risk: how journalists frame risk stories, in *The Social Response to Environmental Risk*, D. Bromley & K. Segerson, eds, Kluwer Academic Publishers, Norwell, pp. 75–101.
- [8] Renn, O. (2004). Perception of risks, *Toxicology Letters* **149**, 405–413.
- [9] Lang, A., Potter, D. & Grabe, M.E. (2003). Making news memorable: applying theory to the production of local television news, *Journal of Broadcasting and Electronic Media* **47**(1), 113–123.
- [10] Hilgartner, S. (2000). *Science on Stage: Expert Advice as Public Drama*, Stanford University Press, Stanford.

- [11] Kingdon, J.W. (2003). *Agendas, Alternatives, and Public Policies*, Little, Brown, Toronto.
- [12] Baumgartner, F.R. & Jones, B.D. (1993). *Agendas and Instability in American Politics*, Chicago University Press, Chicago.
- [13] Stone, D. (1989). Causal stories and the formation of policy agendas, *Political Science Quarterly* **104**, 281–300.
- [14] Soroka, S. (2002). *Agenda-Setting Dynamics in Canada*, UBC Press, Toronto.
- [15] McCombs, M.E. & Shaw D.L. (1993). The evolution of agenda-setting research: twenty-five years in the marketplace of ideas, *Journal of Communication* **43**(2), 58–67.
- [16] Dunwoody, S. & Peters, H.P. (1993). The mass media and risk perception, in *Risk is a Construct*, B. Rück, ed, Kneesebeck, Munich, pp. 293–317.
- [17] Cooper, C.P. & Roter, D.L. (2000). “If it bleeds it leads”? Attributes of TV health news stories that drive viewer attention, *Public Health Reports* **115**(4), 331–338.
- [18] Flynn, J., Slovic, P. & Kunreuther, H. (eds) (2001). *Risk, Media and Stigma: Understanding Public Challenges to Modern Science and Technology*, Earthscan Publications, London and Sterling.
- [19] Hetherington, A. (1985). *News, Newspapers and Television*, Macmillan, London.
- [20] Keating, M. (1993). *Covering the Environment: A Handbook on Environmental Journalism*, National Round Table on the Environment.
- [21] Watt, J.H., Mazza, M. & Snyder, L. (1993). Agenda-setting effects of television news coverage and the effects decay curve, *Communication Research* **20**(3), 408–435.
- [22] Bennett, P. (1999). Understanding responses to risk: some basic findings, in *Risk Communication and Public Health*, P. Bennett & K. Calman, eds, Oxford University Press, Oxford, pp. 3–19.
- [23] Frewer, L.J. (2003). Trust, transparency and social context: implications for social amplification of risk, in *The Social Amplification of Risk*, N. Pidgeon, R.E. Kasperson & P. Slovic, eds, Cambridge University Press, Cambridge, pp. 123–138.
- [24] Tuchman, G. (1973). Making news by doing work: routinizing the unexpected, *The American Journal of Sociology* **79**(1), 110–131.
- [25] Frank, R. (2003). ‘These crowded circumstances’: when pack journalists bash pack journalism, *Journalism* **4**(4), 441–458.
- [26] Frewer, L.J., Rowe, G., Hunt, S. & Nilsson, Å. (1998). *The 10th Anniversary of the Chernobyl Accident: The Impact of Media Reporting of Risk on Public Risk Perceptions in Five European Countries*, CEC Project RISKPERCOM.
- [27] Driedger, S.M. (2007). Risk and the media: a comparison of print and televised news stories of a Canadian drinking water risk event, *Risk Analysis* **27**(3), 775–786.
- [28] Menashe, C.L. & Siegel, M. (1998). The power of frame: an analysis of newspaper coverage of tobacco issues – United States, 1985–1996, *Journal of Health Communication* **3**, 307–325.
- [29] Rogers, C.L. (1999). The importance of understanding audiences, in *Communicating Uncertainty: Media Coverage of New and Controversial Science*, S.M. Friedman, S. Dunwoody & C.L. Rogers, eds, Lawrence Erlbaum Associates, Publishers, Mahwah, pp. 179–200.
- [30] Leiss, W. (2001). Dioxins, or chemical stigma, in *Risk, Media and Stigma: Understanding Public Challenges to Modern Science and Technology*, J. Flynn, P. Slovic & H. Kunreuther, eds, Earthscan Publications, Sterling, pp. 257–267.
- [31] Driedger, S.M. & Eyles, J. (2003). Different frames, different fears: communicating about chlorinated drinking water and cancer in the Canadian media, *Social Science and Medicine* **56**, 1279–1293.
- [32] Singer, E. & Endreny, P.M. (1993). *Reporting on Risk: How the Mass Media Portray Accidents, Diseases, Disasters, and Other Hazards*, Russell Sage Foundation, New York.
- [33] Potter, W.J. (2004). *Theory of Media Literacy: A Cognitive Approach*, Sage Publications, Thousand Oaks.
- [34] Iggers, J. (1999). *Good News, Bad News: Journalism Ethics and the Public Interest*, Westview Press, Boulder.
- [35] Mazur, A. (1984). The journalists and technology: reporting about Love Canal and Three Mile Island, *Minerva* **22**, 45–66.

Related Articles

Evaluation of Risk Communication Efforts Stakeholder Participation in Risk Management Decision Making

BHAVNITA MISTRY AND S. MICHELLE
DRIEDGER

Risk Attitude

Risks Everywhere

Whether it be endogenous or exogenous (as in natural risks), risk may be more or less accepted or abhorred by individuals and populations. History offers innumerable examples where risk is well accepted and even searched for: Egyptians not only knew of the floods near the Nile, but also knew that they would have better agricultural crops there and *intentionally* settled in the area close to the river. Our technological and financial civilization does the same by *intentionally* looking for new products and/or processes in the hope of increasing welfare, but knowingly facing possible adverse consequences in doing so. *Risk attitude* is therefore a fundamental concept, characterizing the degree to which the individual or the organization or a society at large accepts to face a probability mass of adverse consequences in view of desired potential effects. A similar point can be made about the degree of protection we may want to seek against some risk *imposed* on us by natural or social events over which we have no control. As a result, a “risk situation” involves four factors: (a) consequences, (b) their probabilities and distribution (c) individual attitudes preferences, and (d) collective and sharing effects. These are relevant to a broad number of fields motivated by *individual, organizational, and/or “markets” attitudes toward risk*, as well as by *psychological needs and the need to deal with fundamental economic problems*, which implies taking or accepting some risks [1]. As can be intuitively perceived, risk attitudes are an essential part of our life and their quantitative measurement and pricing is closely linked to the individual and collective rationality households or firms may exhibit when facing risky situations [2–4].

Risk Attitudes: Concepts and EU Theory

Theories based on the valuation of decisions made under uncertainty, both expected and non-expected utility (EU) theories (rank-dependent utility (RDU) theory, regret theory, etc.), have been devised, highlighting and contrasting attitudes toward risks. We turn now to the precise expressions that risk attitudes may take as a function of the “utility” (or “global

goal”) function (*see Utility Function*) under EU rationality. In this section we shall see that there are different approaches and definitions of what may be called *risk attitude* and how it is quantified and measured. The section titled “Generalized Risk Aversion: The Rank-dependent Model” examines risk attitudes within a larger rationality framework. In the section titled “Risk Attitudes and Financial Markets”, we also provide an approach, finance based, which has embedded risk attitudes in assets pricing as well, thereby allowing the exchange of risks between persons, investors, and traders of different risk attitudes.

Expected Utility and the Arrow–Pratt Analysis

The EU is the standard way that economists use to evaluate a random prospect or discrete “lottery” L described by the vector $(x_1, p_1; x_2, p_2; \dots; x_n, p_n)$, whereby it is meant that the consequence x_1 is obtained with probability p_1 , etc. Such a discrete lottery will designate hereunder any institutional kind of risk (investment, industrial performance, etc.). From the point of view of an individual endowed with a utility function $u(\cdot)$ and under EU (*see Utility Function; Subjective Expected Utility*), the lottery is worth $\sum_1^n p_i u(x_i)$. The seminal works by Pratt [5] on the one hand and Arrow [6, 7] on the other hand have shown that one version of the attitude toward risk can be recovered from the explicit specification of the utility function.

Assume an individual endowed with a constant (nonrisky) asset C as well as with an actuarially favorable lottery $\tilde{x}$, i.e., with an actuarial value of $E(\tilde{x}) > 0$ (*see Insurance Applications of Life Tables; Insurance Pricing/Nonlife; Insurability Conditions*). Arrow and Pratt characterized the risk attitude of an individual in relation to the amount, priced p_a , at which an individual would be willing to *sell* the lottery or, in other words, the smallest *nonrandom amount* he would be willing to *receive* to remove the risk faced (an amount often loosely called the *cash equivalent* of the lottery to the given individual). Using the EU rule

$$u(C + p_a) = E[u(C + \tilde{x})] \quad (1)$$

with

$$p_a = E(\tilde{x}) - \pi(C, \tilde{x}) \quad (2)$$

From equation (2), we note that, if the individual wants to be insured against the risk (in case of an

actuarially unfavorable risk), he (she) will accept to *pay* the insurance premium by Arrow–Pratt’s definition. In this case, the individual will be said to display risk aversion. Note that to be accepted, the insurance premium will be *at most*

$$p_I(C, \tilde{x}) = -p_a = \pi(C, \tilde{x}) - E(\tilde{x}) \quad (3)$$

Finding a convenient analytical expression of $\pi(C, \tilde{x})$ seems thus an essential question to solve. Unfortunately, this is a very complicated function for discretionarily large values of $\tilde{x}$. However, Pratt showed that if $\tilde{x}$ displays finite variance and if its range is not too large with respect to C , then π can be satisfactorily approximated by the following expression:

$$\begin{aligned} \pi &\cong \frac{-u''[C + E(\tilde{x})]}{u'[C + E(\tilde{x})]} \cdot \frac{\sigma^2(\tilde{x})}{2} \\ &= \frac{-u''(W)}{u'(W)} \cdot \frac{\sigma^2(\tilde{x})}{2} = A(W) \cdot \frac{\sigma^2(\tilde{x})}{2} \end{aligned} \quad (4)$$

where

$$\begin{aligned} W &= C + E(\tilde{x}), \\ A(W) &= -\frac{u''(W)}{u'(W)} = \frac{d}{dW} \log u'(W) \end{aligned}$$

W is the expected wealth, which reduces obviously to C when $\tilde{x}$ is actuarially fair (i.e., has zero mean). From equation (4), it can be immediately seen that any individual endowed with a concave utility function will have positive $A(W)$ and π , as $u'(W) > 0$ for all possible values of W (by definition of a utility function; see **Subjective Expected Utility**) and as variances are always positive, the sign of π is opposite to that of $u''(W)$, i.e., π will be positive if and only if $u''(W) < 0$, which characterizes a concave utility function. We see below that there exists a simple link between the complicated function in equation (3) and the local approximation in equation (4).

The function $A(W)$ is known as the *Arrow–Pratt coefficient of absolute risk aversion* (hereunder *ARA*), and as equations (2) and (4) show, the larger the function $A(W)$, the smaller the *cash equivalent* (or *selling price*) of any given lottery $\tilde{x}$. By definition, therefore, an individual will be said to be *locally risk averse* if and only if he (she) displays positive π (and hence positive $A(W)$), *locally risk prone* (or risk lover) if she (he) displays negative π (and hence negative $A(W)$), and locally risk neutral if he

(she) displays null π (and hence null $A(W)$). This corresponds respectively to a concave, convex, and affine (or linear) utility function.

Risk can also appear as a multiplicative factor, for example, when an amount a of the assets is invested with risky actuarially favorable return $\tilde{g}$, the rest being held as cash C . Expected wealth is then equal to $W = E(C + a \cdot \tilde{g})$. Maximizing EU then leads to a *relative* risk premium $\pi^* \approx A^*(W) \cdot \sigma^2(g)/2$ by a similar computation. The appropriate coefficient of *relative* local risk aversion (RRA) is dimensionless (a percentage ratio), which is denoted here by $A^*(W)$ with

$$A^*(W) = W \cdot \frac{-u''(W)}{u'(W)} = W \cdot A(W) \quad (5)$$

Yet, these are only *local* properties. In what sense do they transfer to the *finite* case, as described by equations (2–4)? The major contribution of John Pratt has been to show precisely that these local properties convey their message to *the large*. In particular, he showed [5] that three propositions were equivalent when comparing two individuals A and B, or the same individual facing two different situations in terms of wealth:

$$1. \quad \forall W, A_1(W) \geq A_2(W) \quad (6)$$

(The ARA coefficient of individual 1 is never smaller than that of individual 2)

$$2. \quad \begin{aligned} &\exists G : \Re \mapsto \Re, s.t. \\ &G' \geq 0, G'' \leq 0, u_1 = G(u_2) \end{aligned} \quad (7)$$

(The utility function of individual 1 is a concave transform of that of 2) and finally

$$3. \quad \forall C, \tilde{x}, \pi_1 \geq \pi_2 \quad (8)$$

(For all possible wealth, the risk premium of 1 is no smaller than that of 2).

The equivalence of propositions 1 and 3 answers the question raised above, establishing some link between equations (1–3) and equation (4) above. Proposition 2 conveys the meaning of a risk attitude to the concave or convex shape of the Neumannian utility function.

How does the local ARA function $A(W)$ evolve in terms of the (final expected) wealth W of some given individual? If $A'(W) > 0$, ARA increases with wealth and the utility function is said to be endowed with IARA (increasing absolute risk aversion). An

example is provided by any quadratic utility function, say $u(W) = W - bW^2$, with $b > 0$ and $W < \frac{1}{2b}$ (to maintain $u'(W) > 0$, as should be the case for any utility function). If $A'(W) < 0$, ARA decreases with W and the utility function is said to be endowed with DARA (decreasing absolute risk aversion). For example, this is shown by any logarithmic utility function, $u(W) = \ln(W)$. For the special case of CARA (constant absolute risk aversion) utility functions, the exponential utility function $u(W) = -e^{-gW}$ (with $g > 0$) serves an example. In this latter case, $A'(W) = 0$, but it is easy to show that $A^{*'}(W) = g > 0$: we say that this utility function is endowed with increasing *relative* risk aversion (IRRA). Definitions of DRRA (Decreasing Relative Risk Aversion) and CRRA (Constant Relative risk Aversion) functions are then straightforward. The reader will check that the logarithmic utility function provides an example of CRRA, while an example for DRRA functions is given by the power function $u(W) = W^k$, where $0 < k < 1$.

Empirical estimates conducted on stock exchange data as in [8] or [9] indicate that $A'(W) < 0$ and most generally $R'(W) \leq 0$. Experimental studies as in [10] confirm these findings. This demonstrates that quadratic utility functions have been used for convenience in textbooks, but hardly fit observable behavior.

Example and Standard Applications

Consider the following situation: an individual owns $C = \$100\,000$ in a nonrisky asset (e.g., 3-month treasury bonds, or monetary open-end funds' shares), as well as a risky asset $\tilde{x}$ yielding $x_1 = \$22\,500$ with probability 0.7 and $x_2 = -\$10\,000$ with probability 0.3. The utility function is assumed to be $u(W) = W^{1/2}$ or $u(W) = \sqrt{W}$.

Computation of the Risk Premium “in the Large”. The EU of the individual amounts to

$$0.7 \cdot u(122\,500) + 0.3 \cdot u(90\,000) = 335$$

Using equation (1), $E[u(C + \tilde{x})] = 335 = u(C + p_a)$; hence $C + p_a = u^{-1}(335) = (335)^2 = 112\,225$ and thus $p_a = 12\,225$. It is straightforward to compute the *actuarial value* (or expected value) of the risky asset: $\$(22\,500 \times 0.7 - 10\,000 \times 0.3) = \$12\,750$. Equation (2) yields finally $\pi = E(\tilde{x}) - p_a = 12\,750 - 12\,225 = 525$.

Approximate Computation of the Risk Premium “in the Small”. What is the risk premium using the local approximation formula of equation (4)? First, note that from the actuarial value of the risky asset, i.e., $\$12\,750$, the *expected wealth* W is $\$112\,750$. The variance is $\sigma^2(\tilde{x}) = 2\,218\,125 \cdot 10^6$. Thus, from the utility function

$$\begin{aligned} u'(W) &= \frac{1}{2W^{1/2}}, \quad u''(W) = -\frac{1}{4(W)^{3/2}}, \\ r(W) &= \frac{1}{4(W)^{3/2}} \cdot \frac{2W^{1/2}}{1} = \frac{1}{2W} \end{aligned} \quad (9)$$

and from equation (4)

$$\begin{aligned} \pi &\approx r(W) \cdot \frac{\sigma^2(W)}{2} = \frac{1}{2W} \cdot \frac{\sigma^2}{2} \\ &= \frac{1}{225\,500} \cdot \frac{221\,812\,500}{2} = 491.82 \end{aligned}$$

which is the risk premium “in the small”. The relative error committed in using the Arrow–Pratt approximation formula, in this example, amounts to $\frac{525-492}{525} \approx 0.063$, i.e., approximately 6.3% of the risk premium, a relatively small error, though not negligible. However, the reader should convince himself (herself) that the less concave the utility function, the better the approximation in equation (4) and the smaller the error rate on the risk premium (for $u(W) = (W)^{3/4}$, for example, the error rate reduces to 5.3%).

Application: The Standard Simple Portfolio Problem and its Many Interpretations. We have assumed in equations (1) and (2) that the individual owned a given lottery $\tilde{x}$ in addition to his certain amount of cash C_0 , or, more appropriately, in addition to his riskless asset. However, we may often be interested in determining what amount $(1 - a^*)C$ of a (nonrandom) wealth C should be held in the riskless asset and what amount, a^* , should be invested in a single investment yielding a random return $\tilde{\rho}$ (we assume here $\tilde{\rho} > -1$ i.e., no realization of $\tilde{\rho}$ can ever be less than 100%). This standard simple portfolio problem can also be interpreted differently: suppose the agent has to decide what proportion a^* of a risky asset to bear by himself and what part $(1 - a^*)$ to sell to an insurance company for a price defined proportionate to expected compensations for the company. This is a simple insurance portfolio

interpretation found in almost every insurance and economics textbook.

Yet another interpretation can be found in agricultural economics: farmers have to decide ahead of time how much to produce, anticipating (using some probability distribution) the price at which they will be able to sell their produce. If their production function is linear in terms of their joined inputs (equipment, land, and work), $\tilde{\rho}$ can be seen as their unitary profit rate and a^* as the proportion of the production capacity they activate. This somewhat generic problem can be (as a first approximation) formalized as follows:

$$\begin{aligned} \text{Max!} & Eu[(1 - a^*)C(1 + i) + a^*C(1 + \rho)] \\ &= Eu[C(1 + i) + a^*C(\rho - i)] \\ &= Eu(C + a^*C\tilde{x}) \end{aligned} \quad (10)$$

(subject to the constraint $0 < a^* < 1$ iff short sales on markets are forbidden or impossible, and with $\tilde{x} = \tilde{\rho} - i$). Second-order conditions are met as $V(C, a^*) = Eu(C + a^*C\tilde{x})$ is a concave function of the decision variable a^* under risk aversion. The solution a° is then found by equating the first-order derivative to zero:

$$V'(a^\circ) = E[\tilde{x} \cdot u'(C + a^\circ C\tilde{x})] = 0 \quad (11)$$

which determines the proportion to be invested in the risky asset.

Intuitively, if $E(\tilde{x}) = 0$, the proportion a^* will be null (why take the burden of risk, if it does not provide a greater return than the riskless asset?). However, the most interesting property of the solution is that the higher the $A^*(W)$, the lower the $a^*(W)$, which follows the intuition, while a change in the distribution of $\tilde{x}$ has no obvious effect on the demand for the risky asset, as argued by Rothschild and Stiglitz [11]. Some conditions on risk attitude have been determined [12] under which the reduction would necessarily happen, but these conditions are somehow contrived ones in terms of absolute and relative risk attitudes and bear little relevance to empirical observations (which are most likely to require considering not only exogenous but also endogenous risks borne by farmers).

Strong and Weak Risk Aversion: Definition, Relationship

For the reasons just recalled, Rothschild and Stiglitz sought to determine some valid rule based on the

comparative static analysis of risk distributions, instead of remaining in the confines of utility or individual welfare. They were concerned with the issue of defining what “more” risk means [13], without referring exclusively to individual preferences. Denote by F and F^* respectively, the two cumulative probability distributions *with identical actuarial value* defined on some interval, assumed here for simplicity’s sake, to be $[0, M]$, i.e., with

$$\int_0^M F(x) dx = \int_0^M F^*(x) dx \quad (12)$$

They showed that F^* could be defined as *more risky* than F according to three apparently different concepts of “risk increase”, which turn out to be indeed equivalent:

- (a) Any “increase” in risk is what risk averters in the Arrow–Pratt sense hate. This translates to

$$\int_0^M u(x) dF^*(x) dx \leq \int_0^M u(x) dF(x) dx \quad (13)$$

for all individuals endowed with a *concave utility function*. This demonstrates that the two other concepts are equivalent *for distributions having identical mean* to the Arrow–Pratt concept.

- (b) An increase in risk obtains each time we add a “white noise”, i.e., a random variable $\tilde{e}$ with $E(\tilde{e}/x) \equiv 0$, for all x , to some distribution, an operation which does not change the mean. For the two cumulative distributions F and F^* , this is expressed by

$$F^*(\tilde{x}) = F(\tilde{x}) + \tilde{e} \quad (14)$$

This definition follows the common intuition that “when, for each scenario, instead of having one given consequence, we have a distribution around that same consequence, we will face more risk”.

- (c) An increase in risk is characterized by a mean-preserving spread (MPS hereunder). An MPS is defined as some *operator* that takes some probability mass from the center of a given distribution and puts it in the tails of it *without altering the expected value* of the distribution. This translates into

$$F^*(\tilde{x}) = \text{MPS}[\text{MPS}[\dots[\text{MPS}(F(\tilde{x}))]]] \quad (15)$$

i.e., the (more risky) distribution $F^*(\tilde{x})$ is the result of a *finite* sequence of MPS applied to $F(\tilde{x})$. It can be shown [13] that, if the three partial orders that obtain from the three definitions (a), (b) and (c) above, denoted here by $\leq_a$, $\leq_b$, and $\leq_c$, respectively, are mutually compared and compared with the partial order derived from the mean–variance (MV) criterion (denoted here by $\leq_v$), the results can be written as

$$\leq_a \iff \leq_b \iff \leq_c \iff \leq_v \quad (16)$$

In other words, any one of the definitions (a), (b), and (c) of what an “increase” in risk means, matches precisely the other two and is implied by anyone of these two other definitions. This however does not hold for the MV definition, which presumes “too much” with respect to the three other definitions as to what an increase in risk represents.

On the basis of the comparative static analyses of the probability distributions for individual attitudes, a new typology of risk attitude “in the strong sense” can then be generated from Rothschild and Stiglitz’s MPS definition. *Risk aversion* implies that MPS are disliked by the individual; *Risk neutrality* implies that any MPS or finite sequence of MPS in any distribution does not change an individual’s satisfaction; *Risk proneness* finally implies that the individual likes any MPS or finite sequence of it. These definitions are called *in the strong sense* because they imply the Arrow–Pratt definition of risk aversion, but not conversely. The proof is straightforward and left to the reader.

Multiple Dimensions – Background Risk

The analysis presented above relates to any single dimensional risk. A natural question to ask is whether it can be maintained in a multiple dimensional case (for example, defined over many commodities). The answer is negative in general, but three different aspects have to be examined:

1. Can any single index of risk aversion still be meaningful and retain the equivalence of Pratt’s three propositions under multiple dimensional risk?
2. Does the existence of an uninsurable risk (“background risk”) make the individual more or less averse to some given endogenous risk?

3. What about the special case of multiperiod risk and of shifting income from one period to the other?

We take up these three issues successively here.

Index of Risk Aversion under Multiple Dimensions. This first question was originally raised by Kihlstrom and Mirman. Their basic finding was that preference *orderings* defined on commodities play a role in comparing risk attitudes indexes, the latter being linked only to *cardinal* (more accurately *von Neumann–Morgenstern*) *one-dimensional* utility [14]. For example, comparing the risk attitudes of the same individual at different levels of wealth requires that the underlying preferences meet some prerequisite for comparability. For similar reasons, risk attitudes can be compared using the Arrow–Pratt coefficient *only* on the subsets of individuals having the same *ordinal* utility. For example, consider two individuals represented by utility functions u_1 and u_2 respectively, both functions defined on the space of commodities x_1 and x_2 , and two commodities bundles $x^1 = (x_1^1, x_2^1)$ and $x^2 = (x_1^2, x_2^2)$, such that $u_1(x^1) > u_1(x^2)$ and, from the point of view of the second individual, $u_2(x^1) < u_2(x^2)$. Now, consider the lottery $[x^1, p; x^2, (1 - p)]$, with any nonzero value of the probability p , and the sure outcome x^1 . Individual 1 prefers the sure outcome, that individual 2 prefers the lottery. But there would be no meaning in saying that individual 1 is more risk averse than individual 2, for it would be easy to find another lottery and another sure outcome for which the difference in observable choices of the two individuals would be exactly the opposite. The choices suggested here reflect differences between preference orderings rather than differences in risk attitudes.

Karni [15] argued that multivariate risks do not typically pertain to bundle of goods, but rather to the level of income, the uncertainty regarding the ultimate consumption bundle being only an indirect implication. Therefore, the indirect utility function should be the appropriate tool to quantify and measure risk aversion in those cases. One then obtains a *matrix* of local risk aversion indexes, *each* local risk premium being defined in terms of the minimal income necessary to reach the utility level of the commodity bundle (the “expenditure function” of microeconomic theory) at mean prices. This is a simple way of defining the *direction* in which

this balance is to be evaluated in the commodity space: quantifying the risk premium requires indeed specifying this direction.

For interpersonal comparisons on *local* risk aversions, the restriction on *ordinal* preferences does apply, as already mentioned, as well as some additional constraints. The major result in [15], however, is that if these restrictions hold everywhere in the commodity space, it makes sense to determine a *global* interpersonal comparison of risk attitude *independent of differences in ordinal preferences*, a unique result in the literature on multivariate risk. Other developments always rely on similar ordinal preferences [16].

Background Risk. As early as 1981 [17], it was discovered that the Arrow–Pratt definition was not “strong” enough in another sense: in the presence of another risk that the individual could not get rid of (“background” or “uninsurable” risk), an increase in risk aversion does not imply that the demand for any other independently distributed risky asset will decrease. In other words, $a^*(W, x)$ is not necessarily inversely related to $A^*(W)$ in the presence of background risk. For example, not taking into account the risk on human capital when estimating the optimal portfolio of assets can lead to overestimation of the demand for – and hence the price of – the risky asset. Following the equivalence recalled above from Rothschild and Stiglitz’s work, adding a white noise to some undesirable lottery might make that lottery desirable, if some background risk exists, which is quite a paradoxical result.

To avoid such results, Ross suggested a “strong” risk aversion concept [17] calling, in the interpersonal comparison problem, for:

$$\inf \frac{u_1''(w)}{u_2''(w)} \geq \sup \frac{u_1'(w)}{u_2'(w)} \quad (17)$$

or, equivalently:

$$\exists \lambda \in \mathbb{R}, \forall (w_1, w_2) s.t. \frac{u_1''(w_1)}{u_2''(w_1)} \geq \lambda \geq \frac{u_1'(w_2)}{u_2'(w_2)} \quad (18)$$

He showed [17] that this definition implies Arrow and Pratt’s definition immediate result, but not conversely (take some exponential utility function and give a counterexample). This result has been acknowledged

to attempt an explanation of the equity premium (expected net excess return $E(\tilde{r})$ of holding the market portfolio of risky assets over the risk-free rate of interest r_f), mostly “underpredicted” by standard EU models. Indeed, if one uses the values of the Arrow–Pratt RRA index recovered from empirical investigations, which mostly fall on or around the interval [2, 4], one can find the equity premium as being within the interval [.10, .20]. In contrast to that figure, known empirical observations [18] display an interval of [5, 7]. The equity premium would thus be undervalued by simple traditional EU reasoning by about 40 times! But it has been argued [19] that, if labor income risk is neither insurable nor diversifiable, this “background noise of life” [17] might explain, if not all, at least part of that puzzling gap between direct observations of the equity premium and computations from observations of RRA.

Yet, Ross had called for a research program consisting in defining a canonical class of utility functions, which would parametrically fit the requirement that Arrow–Pratt risk averse individuals would be made more risk averse when adding some background risk. This program has been fulfilled in the late 1980s and 1990s by Pratt and Zeckhauser, who defined *proper risk aversion* [20], then by Kimball [21], who defined *standard risk aversion*, and finally by Gollier and Pratt [22], who defined *risk vulnerability*. The latter concept is a generalization of the two preceding ones. It implies that the two first derivatives of the utility function are concave transforms of the original utility function. Vulnerability is both necessary and sufficient for the introduction of an unfair background risk to lower the optimal investment in any other independent risk, provided the risk to be borne is *independent* of the background risk. This concept of vulnerability has made harmonic absolute risk aversion (HARA) utility functions popular to the extent that they all represent individuals sensitive to locally proper and proper risks (in the sense of Pratt and Zeckhauser [20]), who, *a fortiori*, are risk vulnerable individuals.

Correlated Investment Risk and Background Risk.

This issue calls for a more detailed characterization, which is dealt with by Eeckhoudt *et al.* [23] and Ross [24]. The latter author draws in particular attention to the fact that two separately undesirable risks can be jointly desirable. Looking backward, this reminds

one of the Allais [25] *versus* the *main stream* debate: there can be, indeed, preferential positive synergies between two risks.

Income Timing and Risk. The preceding extensions and qualifications on risk aversion have mostly dealt with empirical issues making the portfolio or risky investment problem less simple than it first appeared to be. But one should not lose sight of the savings problem mentioned above. This problem has a seemingly different flavor than that of the portfolio, to the extent that it has also to deal with a specific intertemporal trade-off question: how does saving change when risk affecting future income rises? In some sense, this old idea of Keynes (the “precautionary motive” for saving) calls for a special multidimensional extension of the risk attitude question, utility being in this question defined on two dimensions, today’s consumption and tomorrow’s income. Trying to recover an expression for risk aversion similar to that of Arrow and Pratt has long been a concern to authors like Drèze and Modigliani [26], Caperaa and Eeckhoudt [27], Machina [28], Bryis *et al.* [29], Menezes and Hanson [30], and finally Kimball [21]. The story is told in part by Munier [31]. It led to the formal definition of prudence given by Kimball [21].

By definition, an agent can be said to be prudent if and only if the marginal utility of future consumption is convex. In such a case, it has been known since the late 1960s that the willingness to save would increase in response to an increase in a noninsurable risk bearing on the future wealth of some individual. Kimball [21] thus raises the question in terms that are quite different from those of Arrow and Pratt: the issue here is not to determine how much of a *compensating* balance should be given to some individual to *get rid* of some risk, but how much of an *equivalent* balance should be given to that individual in order for him to *take the same decisions* as without any (increase in) risk bearing on his future income and thus to *restore his utility* level. Prudence – or the intensity of the precautionary saving motive – will thus be measured by the *precautionary equivalent premium* ψ^e . Kimball shows that, assuming time separability of utility,

$$\psi^e[C_0, \tilde{x}, u(C_0 + \tilde{x})] = -\frac{u'''(w)}{u''(w)} \cdot \frac{\sigma^2(\tilde{x})}{2},$$

with $w = E(C_0 + \tilde{x})$ (19)

where the coefficient of prudence appears as

$$A_P = \frac{-u'''(w)}{u''(w)} \quad (20)$$

The empirical studies that have been conducted show that people more exposed to future income risk save more. *Prudence* is therefore a widely accepted concept. Moreover, prudence is a necessary condition for DARA, which is also a widely accepted view.

Relation between Mean–Variance and other Risk Models

Prior to Rothschild and Stiglitz [11], pointing out that the MV analysis was not entirely consistent with EU analysis, Borch’s counterexample [32] had been received as a blow to it, particularly in the financial community [33]. Given any two probability distributions, assume some individual sees them as equivalent in view of their respective means and variances. Borch then shows that one can construct, using only the means and variances of these two initial distributions, a new distribution with the same mean and the same variance as the earlier two, but which is nevertheless first order stochastically dominating (or FSD), one of the two initial distributions, providing thereby a clear contradiction to MV analysis. In general, indeed, one needs many more than two moments to entirely characterize a distribution. In addition, any convex combination of two assets with distributions of returns characterized by the same two moments would not necessarily yield the same type of distribution. For a while, standard portfolio theory seemed to be bogged down! Empirical research showed however that, at least over some period of time, differences in portfolios performances – under MV *versus* EU – would not be large, and the financial community continued using standard MV analysis at least as an approximation tool (see also [34–37]).

Meanwhile, the question as to whether there were some conditions under which EU and MV analysis could be considered as analogous or at least close to each other had been seriously raised. Meyer [38] provided a case paper in this direction.

A quantitative treatment of measures of risk attitudes and their correspondence between MV and EUs approaches can be found in Eichner and Wagener [39], building on results by Owen and Rabinovitch [40] and Meyer [38] showing that, if all distributions $\tilde{W} \in \Omega$ differ only by location along their real

line support and by scale, any of them is equal in distribution to $\mu_W + \sigma_W X$ if X is the normalized random variable of any element of Ω . Any *expected utility* defined on that class of distributions can then be written as a function of the couple (μ_W, σ_W) . We denote such EU by $U(\mu_W, \sigma_W)$. Denoting the wealth state of the investor by W , and the k th derivative with respect to wealth by $u^k(W)$, the ratio

$$\forall W \in \mathfrak{R}, A_n(W) = -\frac{u^{n+1}(W)}{u^n(W)} \quad (21)$$

obviously yields the Arrow–Pratt particular measurement of ARA for A_1 , while A_2 and A_3 are used to denote quantitative estimates of absolute *prudence* (see the section titled “Correlated Investment Risk and Background Risk”) and absolute *temperance* respectively. The case of $n = 4$ yields a ratio, which plays a role in Caballe and Pomansky’s concept of “mixed risk aversion” [41].

A number of relationships of that sort allow extending, for the given class Ω of random variables, some of the results recalled here above to MV analysis. Meyer [38] claimed that the class Ω of random variables covers a wide range of economic problems.

Generalized Risk Aversion: The Rank-dependent Model

From the late 1970s onward, one of the alternative models to EU has gained popularity and has been studied and used extensively: the RDU model. Essentially, RDU claims, on experimental [25, 42] and empirical grounds, that probabilities in the vector p are, under risky conditions, being *transformed* into a vector $h(p)$ (in which each element is denoted by $h_i(p)$) while being applied in evaluating risk, or, under uncertainty, that *decision weights* $h_i(p)$ [43] or nonadditive probabilities [44] are being used. Note that the model’s basic feature is that one should take into account the *whole* distribution – the vector p – should be taken into account to evaluate a prospect, expressing thus risk attitude as a result of the *whole* distribution, an old idea first expressed by Allais.

Under such a scheme, risk attitude is generalized and consists of two parts:

1. Probabilistic attitude toward risk, represented by the impact of the transformation function

$\theta(p)$ under risky conditions [43, 45]. The transformation function is usually defined on the decumulative distribution. It then yields decision weights

$$h_i(p) = \theta\left(\sum_{j=i}^n p_j\right) - \theta\left(\sum_{j=i+1}^n p_j\right) \quad (22)$$

where i is the index of the *ranked* (in increasing order) outcomes. Hence RDU (Rank Dependent Utility model).

2. Transformed Arrow–Pratt risk attitude, as illustrated by decreasing marginal utility and described above, *but computed using transformed probabilities*.

Hilton [46] was then able to decompose the individual overall risk premium in such an environment into two parts, given a prospect $L = (x_1, p_1, \dots, x_n, p_n)$ and transformed weights $h_i(p)$ of the probabilities p_i , as follows:

- *Transformed Arrow–Pratt risk premium* $\pi_h(L)$ (note that decision weights are used in the formula):

$$u\left[\sum_{i=1}^{i=n} [h_i(p)x_i - \pi_h(L)]\right] = \sum_{i=1}^{i=n} h_i(p)u(x_i) \quad (23)$$

- *Decision weights risk premium* π_w , representing “probabilistic” risk attitude:

$$\pi_w(L) = \sum_{i=1}^{i=n} p_i x_i - \sum_{i=1}^{i=n} h_i(p) x_i \quad (24)$$

- *Overall transformed risk premium* $\pi^0(L)$

$$u\left[\sum_{i=1}^{i=n} p_i x_i - \pi^0\right] = \sum_{i=1}^{i=n} h_i(p)u(x_i) \quad (25)$$

It is straightforward to show that $\pi^0(L) = \pi_h(L) + \pi_w(L)$. The most common interpretation consists in viewing $u(\cdot)$ as a *von Neumann–Morgenstern* utility function, in which case there is a real split between two different parts of risk attitude. Allais [25, 47] and others interpret $u(x)$ as a Bernoullian utility function whose shape is *explained* by the evaluation of marginal wealth but has *straightforward implications* by Jensen’s inequality for choosing prospects under risk, in which case the “probabilistic”

part is the *only* measure of attitude toward risk properly defined (for related issues, see [48, 49]). In any case, the generalization achieved by RDU theory with respect to EU theory is that risk aversion can emerge without implying any concavity of the utility function.

Risk Attitudes and Financial Markets

Recently, a broad number of techniques, statistical and otherwise, are applied to estimate and reveal the implied risk attitudes of decision makers. Some of these techniques include models we construct to reflect risk attitudes (embedded in explicit utility functions or derived by models that presume a broad range of potential risk attitudes as seen above), tested and confronted with statistical evidence. In other cases, attitudes are determined through the implied preferences in the utility function of a person, an investor, trader, or risk manager. In finance, risk attitudes or persons combine through an exchange between agents with different risk attitudes to determine the price of risk (implied in a “risk neutral distribution” (RND), based on the presumption that markets are “always” reflecting a rational behavior and a risk attitude, which we seek to reveal through the measurement of asset prices). In this sense, the finance, money-based, approach to risk attitudes is similar to the above developments. These models are of course just models, in the sense that they are limited, reflecting the model builder’s understanding of behavioral and economic attitudes, limitations, and bounded rationality, and thereby attitudes to specific events.

Finance uses instruments such as options, contracts, swaps, insurance, investment portfolios, etc. so that risks can be exchanged among traders and investors with broadly varying risk attitudes. When risks are priced by financial markets, they can be traded and distributed in a manner that benefits “all”, i.e., in a manner that those investors with different attitudes and willing to assume more risks will do so and those who seek to assume less risk will do so as well. Of course, this is possible if an exchange between such investors and traders can be realized. The market price is then the mechanism that allows such an exchange and reflects traders’ attitude to risk. When this is not the case, and risks are valued individually, implicitly due to individuals’ attitudes

to risk, a broad number of approaches are used to mitigate their effects. As noted above, measures of risk attitudes commonly used and based on EU are based on utility functions. Recently, attempts have been made to empirically characterize the risk aversion by extracting the RND, expressing the market (consisting of a large number of investors, speculators, buyers, and sellers) attitude toward future potential events and their probabilities. In this sense, market data can provide an implied risk aversion that the market is presumed to have. Investors are not indifferent to risk and their corresponding subjective probability is thus different than the RND. In fact, the RND is adjusted upward (or downward) for all states in which the dollars are more (or less) highly valued. Hence the higher the risk aversion, the more different the RND and the subjective probability. The risk aversion can thus be estimated from the joint observation of the two densities (see [50–54]). In other words, the index of ARA $A(\cdot)$, can be written as a functional relationship (assuming away the generalization of the section titled “Generalized Risk Aversion: The Rank-dependent Model”):

$$A(\cdot) = f(\text{RND, subjective probability}) \quad (26)$$

This was exploited by Jackwerth [51] and Ait-Sahalia and Lo [53] while extracting a measure of risk aversion in a standard dynamic exchange economy as in [55]. Such a relation can be defined simply by maximizing the EU of a portfolio at a future date whose current price is \$1 and using the investor’s subjective probability distribution in valuing the utility of money. In this formulation, risk attitude is a function of the risk neutral and the investor’s subjective probability distributions. Given any two of these elements, the third can then be inferred. This has led to empirical applications where the implied risk attitude can be determined on the basis of market option prices (where the implied RND can be determined using option prices for example). This important observation can be determined simply as we shall see below.

Assume that markets are complete. Then, a portfolio price is necessarily equal to the discounted future portfolio price with expectation taken with respect to the RND and discounted at the risk-free rate. A simple mathematical formulation that simultaneously captures an investor’s risk attitude, his subjective probability, and the market RND consists in maximizing the investor future EU subject to the current

market pricing of the portfolio (using a RND). Letting, for convenience, the current market price of the portfolio be \$1, we have the following problem:

$$\begin{aligned} & \text{Max}_W E_S u(W_T) \\ & \text{Subject to : } 1 = \frac{1}{(1+r_f)^T} E_{RN}(W_T) \end{aligned} \quad (27)$$

where W_T is the terminal time T portfolio price whose utility we seek to maximize and r_f is the risk-free market rate. Let λ be the constraint Lagrange multiplier. Then, optimization leads to the first-order condition:

$$u'(W_T) = \frac{\lambda}{(1+r_f)^T} \frac{f_{RN}}{f_S} \quad (28)$$

where $f_S(\cdot)$ and $f_{RN}(\cdot)$ are the subjective and the risk neutral distributions. An additional derivative with respect to the portfolio price W_T at time T yields after some elementary manipulations an index of ARA, which is given by the derivative with respect to wealth of a function for “discriminating” between the subjective and the risk neutral distribution:

$$A(W_T) = \frac{d}{dW_T} g(W_T), \quad g(W_T) = \ln \frac{f_S(W_T)}{f_{RN}(W_T)} \quad (29)$$

A broad number of techniques can then be used to estimate empirically the RND (for example, by using market data on derivatives and calculating the implied RND). This establishes a specific relationship between risk attitude and the subjective probability distribution.

For example, if the utility of an investor is assumed to be HARA (hyperbolic absolute risk aversion) with an index of ARA, $A(W_T)$, given by

$$\begin{aligned} u(W_T) &= \frac{1-\gamma}{\gamma} \left\{ \frac{aW_T}{1-\gamma} + b \right\}^\gamma, \\ A(W_T) &= \frac{a}{\left\{ \frac{aW_T}{1-\gamma} + b \right\}} \end{aligned} \quad (30)$$

Then, additional elementary manipulations indicate that

$$\begin{aligned} A(W_T) &= \frac{a(1-\gamma)}{aW_T + b(1-\gamma)} = \frac{d}{dW_T} g(W_T) \\ \text{and } (aW_T + b(1-\gamma))^{(1-\gamma)} &= \frac{f_S(W_T)}{f_{RN}(W_T)} \end{aligned} \quad (31)$$

Similarly, assume that both the subjective and the risk neutral distributions are normally distributed with means and variances given by (μ_S, σ_S^2) and $(\mu_{RN}, \sigma_{RN}^2)$. Then,

$$\begin{aligned} \frac{f_S(W_T)}{f_{RN}(W_T)} &= \frac{\sigma_{RN}}{\sigma_S} \\ &\times \exp \left[-\frac{(W_T - \mu_S)^2}{2\sigma_S^2} + \frac{(W_T - \mu_{RN})^2}{2\sigma_{RN}^2} \right] \end{aligned} \quad (32)$$

Therefore,

$$\begin{aligned} A(W_T) &= \frac{d}{dW_T} \left(\frac{\ln(\sigma_{RN} - \sigma_S) - \frac{(W_T - \mu_S)^2}{2\sigma_S^2} + \frac{(W_T - \mu_{RN})^2}{2\sigma_{RN}^2}}{\frac{(W_T - \mu_S)^2}{2\sigma_S^2} + \frac{(W_T - \mu_{RN})^2}{2\sigma_{RN}^2}} \right) \\ &= W_T \left(\frac{\mu_S}{\sigma_S^2} - \frac{\mu_{RN}}{\sigma_{RN}^2} \right) \end{aligned} \quad (33)$$

Thus, if $\frac{\mu_S}{\sigma_S^2} < \frac{\mu_{RN}}{\sigma_{RN}^2}$, risk aversion decreases linearly in wealth and vice versa. Note that a risk aversion is then defined by the sign of $A(W_T)$ as defined in this artificial example. In this sense, a risk attitude can be revealed by the market. Evidently, for a risk neutral investor, we have $A(W_T) = 0$ and

$$\ln \left(\frac{f_S(W_T)}{f_{RN}(W_T)} \right) = 0, \text{ or } f_S(W_T) = f_{RN}(W_T) \quad (34)$$

Higher-order risk attitudes can be similarly defined. Explicitly, the index of prudence can be defined as well in terms of the subjective and risk neutral distributions logarithmic function $g(W_T)$ and given by

$$A_P = \frac{g''(W_T)}{g'(W_T)} - g'(W_T), \quad g''(W_T) = (A_P + A)A \quad (35)$$

If an investor is risk averse, $A > 0$ and therefore, $g''(W_T) > 0$ (an increasing rate of discrimination rate with respect to wealth) implies prudence since $A_P > 0$. However, if empirical analysis indicates $g''(W_T) < 0$, then the prudence index is both negative and larger than the index of risk aversion. Extensive research, both theoretical and empirical, has extended this approach and indicated a number of important results. For example, risk attitude is not only state varying but is time varying as well. Further, it clearly sets out the concept of risk attitude in terms of a

distance between the subjective and the risk neutral (market) distributions. Again, in the case of the artificial normal example treated earlier, we have an index of prudence given by

$$A_P = \frac{1}{A} \left(\frac{1}{\sigma_{RN}^2} - \frac{1}{\sigma_S^2} \right) - A \quad (36)$$

In this equation, we clearly see the relationship between the risk neutral, the subjective, and the risk attitude of the investor, both with respect to the index of ARA and the investor's prudence.

References

- [1] Tapiero, C.S. (2004). Risk management, in *Encyclopedia on Actuarial Science and Risk Management*, J. Teugels & B. Sundt, eds, John Wiley & Sons, New York, London.
- [2] French, S. (1993). *Decision Theory*, Ellis Horwood, New York, London.
- [3] Machina, M.J. & Munier, B. (eds) (1999). *Beliefs, Interactions and Preferences in Decision Making*, Kluwer Academic Publishers, Boston, Dordrecht.
- [4] Abdellaoui, M., Luce, R.D., Machina, M.J. & Munier, B. (eds) (2007). *Uncertainty and Risk: Mental, Formal and Experimental Representations*, Springer, Berlin, New York.
- [5] Pratt, J.W. (1964). Risk aversion in the small and in the large, *Econometrica* **32**, 122–136.
- [6] Arrow, K.J. (1963). Aspects of the theory of risk bearing, YRJO Jahnsson lectures, also in 1971 *Essays in the Theory of Risk Bearing*, Markham Publishing Company, Chicago.
- [7] Arrow, K.J. (1982). Risk perception in psychology and in economics, *Economics Inquiry* **20**, 1–9.
- [8] Blume, M.E. & Friend, I. (1975). The asset structure of individual portfolios and some implications for utility functions, *Journal of Finance* **30**, 585–603.
- [9] Cohn, R.A., Lewellen, W.G., Lease, R. & Schlarbaum, G.G. (1975). Individual investor risk aversion and investment portfolio composition, *Journal of Finance* **30**, 605–620.
- [10] Levy, H. (1994). Absolute and relative risk aversion: an experimental study, *Journal of Risk and Uncertainty* **3**, 289–307.
- [11] Rothschild, M. & Stiglitz, J. (1971). Increasing risk II: its economic consequences, *Journal of Economic Theory* **3**, 66–84.
- [12] Hadar, J. & Seo, T.K. (1990). The effects of shifts in a return distribution on optimal portfolios, *International Economic Review* **31**, 721–736.
- [13] Rothschild, M. & Stiglitz, J. (1970). Increasing risk I: a definition, *Journal of Economic Theory* **2**, 225–243.
- [14] Kihlstrom, R.E. & Mirman, L.J. (1974). Risk aversion with many commodities, *Journal of Economic Theory* **8**, 337–360.
- [15] Karni, E. (1979). On multivariate risk aversion, *Econometrica* **47**, 1391–1401.
- [16] Levy, H. & Levy, L. (1991). Arrow-Pratt measures of risk aversion: the multivariate case, *International Economic Review* **32**, 891–898.
- [17] Ross, S.A. (1981). Some stronger measures of risk aversion in the small and in the large with applications, *Econometrica* **49**, 621–638.
- [18] Shiller, R. (1989). *Market Volatility*, MIT Press, Cambridge.
- [19] Weil, P. (1992). Equilibrium asset prices with undiversifiable labor income risk, *Journal of Economic Dynamics and Control* **16**, 769–790.
- [20] Pratt, J.W. & Zeckhauser, R. (1987). Proper risk aversion, *Econometrica* **55**, 143–154.
- [21] Kimball, M. (1990). Precautionary saving in the small and in the large, *Econometrica* **58**, 53–78.
- [22] Gollier, C. & Pratt, J.W. (1996). Risk vulnerability and the tempering effect of background risk, *Econometrica* **64**, 1109–1123.
- [23] Eeckhoudt, L., Gollier, C. & Schlesinger, H. (1996). Changes in background risk and risk taking behavior, *Econometrica* **64**, 683–689.
- [24] Ross, S.A. (1999). Adding risks: Samuelson's fallacy of large numbers revisited, *Journal of Financial and Quantitative Analysis* **34**, 323–340.
- [25] Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine, *Econometrica* **21**, 503–546.
- [26] Dreze, J. & Modigliani, A. (1972). Consumption decisions under uncertainty, *Journal of Economic Theory* **5**, 308–385.
- [27] Caperaa, P. & Eeckhoudt, L. (1975). Delayed risks and risk premiums, *Journal of Financial Economics* **2**, 309–320.
- [28] Machina, M.J. (1984). Temporal risk and the nature of induced preferences, *Journal of Economic Theory* **33**, 192–231.
- [29] Bryis, E., Eeckhoudt, L. & Loubergé, H. (1989). Endogenous risk and the risk premium, *Theory and Decision* **26**, 37–46.
- [30] Menezes, C.F. & Hanson, D.L. (1970). On the theory of risk aversion, *International Economic Review* **11**, 481–487.
- [31] Munier, B. (1995). From instrumental to cognitive rationality: contributions of the last decade to risk modeling, *Revue d'Economie Politique* **105**, 5–70. (French, with abstract in English).
- [32] Borch, K. (1969). A note on uncertainty and indifference curves, *Review of Economic Studies* **36**, 1–4.
- [33] Baron, D.P. (1977). On the utility theoretic foundations of mean variance analysis, *Journal of Finance* **32**, 1683–1697.

- [34] Epstein, L.G. & Zin, S.E. (1989). Substitution, risk aversion and the temporal behavior of consumption and asset returns: a theoretical framework, *Econometrica* **57**, 937–969.
- [35] Epstein, L.G. & Zin, S.E. (1991). Substitution, risk aversion and the temporal behavior of consumption and asset returns: an empirical analysis, *Journal of Political Economy* **99**, 263–286.
- [36] Gollier, C. (1995). The comparative statics of changes in risk revisited, *Journal of Economic Theory* **66**, 522–536.
- [37] Gollier, C. (2000). *The Economics of Risk and Time*, MIT Press, Cambridge.
- [38] Meyer, J. (1987). Two moments decision models and expected utility maximization, *American Economic Review* **77**, 421–430.
- [39] Eichner, T. & Wagener, A. (2005). Measures of risk attitude and correspondence between mean variance and expected utility, *Decisions in Economics and Finance* **28**, 53–67.
- [40] Owen, J. & Rabinovitch, R. (1983). On the class of elliptical distributions and their applications to the theory of portfolio choice, *The Journal of Finance* **38**, 745–752.
- [41] Caballe, J. & Pomanski, A. (1996). Mixed risk aversion, *Journal of Economic Theory* **71**, 485–513.
- [42] Abdellaoui, M. & Munier, B. (1997). Experimental determination of preferences under risk: the case of very low probability radiation, *Ciencia and Tecnologia dos Materiais* **9**, 25–31.
- [43] Wakker, P.P. (2001). On the composition of risk preference and belief, *Psychological Review* **111**, 236–241.
- [44] Schmeidler, D. (1989). Subjective probability and expected utility without additivity, *Econometrica* **57**, 571–587.
- [45] Wakker, P.P. (1994). Separating marginal utility and probabilistic risk aversion, *Theory and Decision* **36**, 1–44.
- [46] Hilton, R.W. (1988). Risk attitude under two alternative theories of choice under risk, *Journal of Economic Behavior and Organization* **9**, 119–136.
- [47] Allais, M. (1979). The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School, in *Expected Utility Hypothesis and the Allais Paradox*, M. Allais & O. Hagen, eds, D. Reidel Publishing Company, Dordrecht; The original publication by Allais can be found in: *Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine*, *Econometrica* **21**, 503–546.
- [48] Serrao, A. & Coelho, L. (2004). Cumulative prospect theory applied to farmers' decision behavior in the Alentejo dryland region of Portugal, Discussion paper presented at the *FUR XI Conference*, GRID, Cachan.
- [49] Bell, D.E. (1983). Risk premiums for decision regrets, *Management Science* **29**, 1156–1166.
- [50] Jackwerth, J.C. (2000). Recovering risk aversion from option prices and realized returns, *The Review of Financial Studies* **13**(2), 433–451.
- [51] Jackwerth, J.C. (1999). Option implied risk neutral distributions and implied binomial trees: a literature review, *Journal of Derivatives* **7**, 66–82.
- [52] Ait-Sahalia, Y. & Lo, A.W. (1998). Nonparametric estimation of state price densities implicit in financial asset prices, *Journal of Finance* **53**, 499–548.
- [53] Ait-Sahalia, Y. & Lo, A.W. (2000). Nonparametric risk management and implied risk aversion, *Journal of Econometrics* **94**, 9–51.
- [54] Perignon, C. & Villa, C. (2002). Extracting information from option markets: smiles, state price densities and risk aversion, *European Financial Management* **8**, 495–513.
- [55] Lucas, R. (1978). Asset prices in an exchange economy, *Econometrica* **46**, 1429–1446.

Related Articles

Subjective Expected Utility

Utility Function

BERTRAND R. MUNIER AND CHARLES TAPIERO

Risk Characterization

During the past decade, there has been a trend of using stochastic techniques in risk assessment to characterize risk more completely [1–4]. The results of these analyses can be summarized in the form of a distribution of possible risks within the exposed population, taking into account as many sources of uncertainty and variability as possible. This distribution indicates the maximal and minimal risks that might be experienced by different individuals and the relative likelihood of intermediate risks between these two extremes. Finley and Paustenbach [5] discuss the benefits and disadvantages of using probabilistic exposure assessment compared with using point estimates. Point estimates are simple and user-friendly but tend to overestimate exposure and provide no indication of confidence. Generally, risk and exposure are computed as a function of the number of risk factors. These factors may be thought of as input variables for a model in which the output is a quantitative measure of risk or exposure. Often, there is prior information about the admissible ranges or shape of the distributions of the input variables. An integrated environmental health risk assessment model (*see* **Environmental Health Risk**) gives a complete picture of the environmental risk assessment, including source characterization, fate and transport of the substance, exposure media, biological modeling, and estimation of risk using dose–response modeling (*see* **Dose–Response Analysis**). To the extent possible, distributions characterizing the uncertainty associated with each adjustment factor are developed using scientific knowledge. Comprehensive realism allows for characterizing the impact of professional judgments and assumptions on risk estimation by presenting a decision tree for systematically displaying all plausible scenarios [6]. Thus, the information upon which health risk assessment is based is subject to numerous factors called *risk factors*. Whether the assessment relies on data obtained from human epidemiological studies or on surrogate data such as animal bioassays (*see* **Cancer Risk Evaluation from Animal Studies**), the distribution of risk factors affects the reliability of risk estimates and the interpretation of the assessment results.

In the characterization of risk, risk factors can be categorized into two main groups, potency and exposure. Potency can be further classified by risk

factors. When human risks are predicted on the basis of toxicological data, there is a factor related to extrapolating the effects of the high doses used in laboratory studies to those of the lower levels of exposure typically experienced by humans under natural conditions. In addition, risk factors can include translating findings from laboratory animals data to humans, extrapolating toxicological results obtained *via* one route of exposure (oral, inhalation, or dermal) to another route, or using results of subchronic toxicity studies to predict chronic effects [7].

The exposure factor is equally important in characterizing risk. Building a proper distribution of the exposure factor is the key in estimating exposure. Epidemiological data on exposed human populations are subject to variability [8, 9]. Not only is exposure difficult to assess, both retrospectively and prospectively, but disease status could be in error. Retrospective exposure profiles can be difficult to construct, particularly in terms of chronic diseases such as cancer, for which exposure data from many years prior to disease ascertainment are needed. For example, radon measurements in homes taken today may not reflect past exposures because of changes in dwelling address, building renovations, or lifestyle (e.g., sleeping with the bedroom window open or closed), or inherent variability in radon measurements (*see* **Radon**). Exposure in prospective studies may also be quite variable. In dietary studies, for example, both dietary recall and food diaries can be subject to error. In epidemiological studies using computerized systems to link exposure data from one database with health status data in another database, even vital status (whether a subject is living or dead) can be in error [8]. Furthermore, because many epidemiological investigations are based on occupational groups with higher levels of exposure than the general population, uncertainty arises when data from occupational studies (*see* **Occupational Cohort Studies**) are extrapolated to environmental exposure conditions [10]. For example, lung cancer risks experienced by underground miners exposed to high levels of radon gas in the past may be used to predict the potential risks associated with lower levels of radon present in homes today [10].

For the purposes of this article, we will consider risk as a function of input (or risk) factors that are subject to uncertainty and variability. When the input

variable is fixed (deterministic in a broad sense) but unknown, the distribution of values obtained from repeated observations represents uncertainty. The repeated observations can be used to construct a confidence interval for which there is a given percent chance of bounding the true value. Note that large amounts of uncertainty indicate that there may be opportunities to improve the information base for decision making by conducting additional research in areas where information is most doubtful. When the input variable is stochastic and represented by a distribution with known parameters, variability is present in the input variable. Note that, other things being equal, a great deal of real variability indicates that societal resources can be more efficiently targeted to reduce risk where it is the most intense. Some risk factors have both uncertainty and variability, while some have only one. Details about uncertainty and variability and their impact on risk characterization are discussed in another entry of this encyclopedia (see **Uncertainty and Variability Characterization and Measures in Risk Assessment**).

In the remainder of this chapter, we will consider a general framework for risk characterization in a multiplicative model and in a general risk model in which the input variables are stochastic. In the section titled “General Risk Model Incorporating Stochastic Risk Factors”, we examine a general risk model using the example of characterizing the risk of lung cancer due to radon exposure. In the section titled “Multiplicative Risk Model: Ingestion of Contaminated Fish”, we characterize risks in a multiplicative model using the example of risk assessment for individuals who ingested contaminated fish [11]. In the section titled “Multiplicative Risk Model: Ingestion of Contaminated Fish”, we conclude with a discussion about the assumptions.

General Risk Model Incorporating Stochastic Risk Factors

Rai *et al.* [12] have developed a general framework for characterizing and analyzing uncertainty and variability for arbitrary risk models. Suppose that the risk R depends on p risk factors $X_1, \dots, X_p$ according to the function

$$R = H(X_1, X_2, \dots, X_p) \quad (1)$$

Each risk factor X_i may vary within the population of interest according to some distribution with

a probability density function $f_i(X_i|\theta_i)$, which is conditional upon the parameter θ_i , which may or may not be known. When a parameter is unknown, the uncertainty in X_i is characterized by the distribution $g_i(\theta_i|\theta_i^0)$ for θ_i , where θ_i^0 is a known constant. The stochastic risk factor (with total uncertainty/variability) X_i is then described by the density function

$$c_i(X_i|\theta_i^0) = \int f_i(X_i|\theta_i)g_i(\theta_i|\theta_i^0) d\theta_i \quad (2)$$

If θ_i is a valued vector, then g_i is a multivariate distribution. Here, we assume that the forms of the distributions f and g are known. Usually, a risk factor is considered a lognormal distribution, i.e., $f_i(X_i|\theta_i)$ has a lognormal distribution, with either known or unknown parameters. The most common distributions for unknown parameters are triangular distributions. The $g_i(\theta_i|\theta_i^0)$ has a triangular distribution for the mean and standard deviation parameters of $\log X$.

For any correlated pair (i, j) , let X_i and X_j have a joint distribution f_{ij} , conditional upon the parameter θ_{ij} . Here, f_{ij} represents the joint distribution of variability in X_i and X_j . The parameter vector θ_{ij} , in turn, possibly has a multivariate distribution g_{ij} with the known parameter θ_{ij}^0 . Here, g_{ij} represents the joint distribution of uncertainty in X_i and X_j .

Risk Distributions

To characterize risk, we will consider the distribution of $R^* = \log H(X_1, X_2, \dots, X_p)$. Allowing for both uncertainty and variability in risk, the distribution of R^* is given by

$$C^*(R^* \leq r^*) = \int_{\log H(x_1, \dots, x_p) \leq r^*} \int c(X_1, \dots, X_p|\theta^0) \times dX_p \dots dX_1 \quad (3)$$

where $c(\cdot)$ is a joint density function of all the risk factors; the parameter of this joint distribution is $\theta^0 = \{\theta_i^0, \theta_{ij}^0; i, j = 1, \dots, p\}$. The joint density function $c(\cdot|\theta^0)$ can be partitioned into two parts: $f(\cdot|\theta)$ represents the joint density due to variability in the risk factors and $g(\cdot|\theta^0)$ represents the joint density function due to uncertainty in the risk factors. The probability function in equation (3) can be further

expanded as follows:

$$C^*(R^* \leq r^*) = \int_{\log H(x_1, \dots, x_p) \leq r^*} \int \{f(X_1, \dots, X_p | \theta) \times dX_p \dots dX_1\} \{g(\theta | \theta^0) d\theta\} \quad (4)$$

The mean and variance of the distribution of R^* can be approximated by $\log H(\mu_1^0, \mu_2^0, \dots, \mu_p^0)$ and $W(R^*)$.

If all of the risk factors in equation (1) are log-normally distributed and the risk model is of the multiplicative form, then the distribution of R^* is normal with a mean $\log H(\mu_1^0, \mu_2^0, \dots, \mu_p^0)$ and variance $W(R^*)$. Sometimes, risk factors are assumed to be a beta, triangular, or log-triangular distribution, if not lognormally distributed. In that case, they can be approximated very well by a lognormal distribution. Thus, if all of the risk factors are approximately log-normally distributed, the distribution of R^* can be approximated by a normal distribution. The distribution of R^* can also be approximated by Monte Carlo simulation. Although straightforward, Monte Carlo simulation can become computationally intensive with a moderate number of risk factors.

To apply the Monte Carlo method, we simply draw a random sample of the values of the risk factors from the distribution $c(\cdot)$ and calculate the value of R^* based on that sample. Repetition of this procedure for a sufficiently large number of times yields the Monte Carlo distribution of R^* . Note that when $\{X_i\}$ are independent, values of the risk factors can be generated from their marginal distributions $\{c_i\}$.

Multiplicative Risk Model: Ingestion of Contaminated Fish

Consider the example analyzed by Hoffman and Hammonds [11] and reanalyzed by Rai *et al.* [12]. In this example, a multiplicative model is used to estimate the hazard quotient for a group of individuals potentially exposed to risk through the ingestion of contaminated fish. Here, risk is defined by the hazard quotient, which is equal to exposure divided by the reference dose. Exposure to the contaminant is estimated by multiplying the concentration of the contaminant in the fish X_1 by the ingestion rate of fish X_2 and then normalized by dividing by the body mass of a human X_3 and the reference dose X_4 . The

risk R can then be expressed as

$$R = X_1 \times X_2 \times (X_3)^{-1} \times (X_4)^{-1} \quad (5)$$

The risk factor X_4 is assumed to be subject to uncertainty only. For a specific human population of interest (with known mean and standard deviation), X_3 is subject to variability only. However, when extrapolating to a general human population, X_3 is also subject to uncertainty. The remaining variables X_1 and X_2 are assumed to be subject to both variability and uncertainty. (Characterizations of distributions of the risk factors are given in Table 1 of Rai *et al.* [12], and the resulting distributions are depicted in Figures 2 and 3 of the same reference.)

Complex Risk Model: Residential Radon Exposure

In this section, we apply the methods described in the section titled “General Risk Model Incorporating Stochastic Risk Factors” to characterize a general risk model for estimating lung cancer risk due to residential exposure to radon [13, 14]. A number of factors can influence radon risk estimates. The primary concern is the excess risk of lung cancer demonstrated in 11 cohort studies of underground miners conducted in a number of countries around the world. α Particles emitted by radon daughters appear to exert radon’s carcinogenic effects by causing DNA damage to lung tissue. Thus, even low levels of exposure to radon can confer some increase in risk. After considering different possible approaches to risk estimation, the Biological Effects of Ionizing Radiation (BEIR) VI Committee elected to base risk models on epidemiological data derived from studies of the mortality of miners. The committee conducted a combined analysis of updated data from the 11 miner cohorts and developed two models for projecting lung cancer risks in the general population. These two models are referred to as the *exposure-age-concentration* and *exposure-age-duration* models. We will demonstrate using the exposure-age-duration model.

If e_t denotes the excess relative risk (ERR) at age t , then

$$e_t = \beta \times \omega(t) \times \phi(t) \times \gamma_{\text{dur}}(t) \times K \quad (6)$$

The factor β ($(\text{Bq}\cdot\text{m}^3)^{-1}$) reflects the carcinogenic potency of radon, as modified by the other risk factors in model (6). The last term in this model is the dosimetric K factor (dimensionless), which is used to extrapolate from the conditions of occupational exposure to those of environmental exposure. The factor $\omega(t)$ represents a time-weighted average exposure to radon, expressed in $\text{Bq}\cdot\text{m}^3$, within exposure-time windows 5–14, 15–25, and 25+ years prior to disease diagnosis. The factor $\omega(t)$ can be expressed as follows:

$$\begin{aligned}\omega(t) &= \omega \times [\nabla_{[5,14]}(t) + \theta_2 \nabla_{[15,24]}(t) + \theta_3 \nabla_{[25,\infty]}(t)] \\ &= \omega \times \eta_t\end{aligned}\quad (7)$$

where

$$\nabla_{[a,b]}(t) = \begin{cases} 10 & \text{for } t > b \\ (t-a) & \text{for } a \leq t \leq b \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

The factor $\phi(t)(y^{-1})$ indicates the effects of attained age, categorized into four broad age groups:

$$\phi(t) = \begin{cases} \phi_1 & \text{for } t \leq 54 \\ \phi_2 & \text{for } 55 \leq t \leq 64 \\ \phi_3 & \text{for } 65 \leq t \leq 74 \\ \phi_4 & \text{for } t \geq 75 \end{cases} \quad (9)$$

The factor $\gamma_{\text{dur}}(t)(y^{-1})$ reflects the duration of exposure to radon (in years), where

$$\gamma_{\text{dur}}(t) = \begin{cases} \gamma_{d1} & \text{for } t < 10 \\ \gamma_{d2} & \text{for } 10 \leq t < 20 \\ \gamma_{d3} & \text{for } 20 \leq t < 30 \\ \gamma_{d4} & \text{for } 30 \leq t < 40 \\ \gamma_{d5} & \text{for } t \geq 40 \end{cases} \quad (10)$$

Our interest lies in risks associated with low (residential) radon exposures in middle-aged to elderly population. We demonstrate the risk characterization for the exposure-age-duration model for any person older than 40 years, for which $\gamma_{d5} = 10.18$. In this analysis, the risk function (equation (7)) represents the age-specific *ERR* model, which is represented as a product of four factors: the carcinogenic potency of radon $\beta = X_1$, the exposure to radon $\omega\eta_t = X_2$, the effect of age $\phi = X_3$, and $K = X_4$. Thus,

$$ERR = X_1 \times X_2 \times X_3 \times X_4 \times \gamma \quad (11)$$

where $\gamma = 10.18$. For simplicity, we assume that γ is a constant, i.e., it has neither uncertainty nor

variability. Note that the risk factors X_1 , X_2 , and X_3 are correlated. In this application, X_1 and X_3 are assumed to be subject to uncertainty only, and X_2 and X_4 are assumed to be subject to both variability and uncertainty.

The basic distribution assumptions are given in Tables 1a and 1b in [13]. The carcinogenic potency of radon X_1 is assumed not to vary among individuals, reflecting equal individual sensitivities within the average population. Uncertainty in X_1 is characterized by a lognormal distribution, with known geometric mean and geometric standard deviation. The level of radon $\omega = X_2/\eta_t$ is assumed to vary among homes in accordance with a lognormal distribution. Uncertainty in radon levels is described by means of lognormal distributions for both the geometric mean and geometric standard deviation of the distribution of radon levels. Although the modifying effect of age X_3 is assumed not to vary among individuals within a given age group, uncertainty is described by means of a lognormal distribution. Variability in the K factor X_4 is characterized by a log-uniform distribution for the geometric standard deviation. This description of uncertainty and variability was guided by both empirical data and expert judgment on the part of the BEIR VI Committee. The covariance structure used to define the correlation matrix is given in Table 2b in [13]. This correlation structure is based on the covariance matrix obtained when estimating the model parameters. Other than the statistical correlation between X_1 and X_3 , which is induced by the estimation procedure, the risk factors are assumed to be independent. Using these interlinked distributions, the distribution of risk can be estimated for different sections of the population.

Characterization of Lifetime Relative Risk of Lung Cancer due to Radon Exposure

To determine the lifetime relative risk (LRR) of radon-induced lung cancer, we require the hazard function (see **Hazard and Hazard Ratio**) for the death rate in the general population. Deaths can be classified into two categories: those due to lung cancer and those due to other competing causes. Assume that the age at death, T , is a continuous random variable. Let $h(t)$ be the rate of lung cancer

deaths and $h^*(t)$ be the rate of overall deaths at age $T = t$, in an unexposed population. Furthermore, let $e(t)$ be the ERR in an exposed population. The death rate due to lung cancer in the exposed population at age t is $h(t) + h(t)e(t)$, and the overall death rate is $h^*(t) + h(t)e(t)$.

Let $R(t)$ be the probability of death due to lung cancer for a person of age t , and $S(t)$ be the survival probability up to age t in the exposed population. Following Kalbfleisch and Prentice [15], the survival function can be expressed as

$$S(t) = \Pr\{T \geq t\} \\ = \exp \left\{ - \int_0^t [h^*(u) + h(u)e(u)] du \right\} \quad (12)$$

and the probability of death due to lung cancer, $R(t)$, can be written as

$$R(t) = [h(t) + h(t)e(t)] \\ \times \exp \left\{ \int_0^t [h^*(u) + h(u)e(u)] du \right\} \quad (13)$$

For simplicity, we assume that death times are recorded in years and identify the range of T as $1, 2, \dots, 110, \dots$. Let h_t be the lung cancer mortality rate, h_t^* be the overall mortality rate for age-group t in an unexposed population, and e_t be the ERR for lung cancer mortality in an exposed population for age-group t . Then, the death rates in the exposed population are given by $h_t + h_t e_t$ for lung cancer and $h_t^* + h_t e_t$ for all causes including lung cancer. The discrete-time version of equations (12) and (13) are given as follows:

$$S_t = \Pr\{T \geq t\} = \prod_{l=1}^{t-1} \{1 - (h_l^* + h_l e_l)\} \\ \approx \exp \left\{ - \sum_{l=1}^{t-1} (h_l^* + h_l e_l) \right\} \quad (14)$$

and

$$R_t = \Pr\{T = t\} = (h_t + h_t e_t) S_t \quad (15)$$

The approximation in equation (14) is highly accurate and will be treated as exact in what follows. Substituting equation (14) into equation (15),

the age-specific death rate due to lung cancer in the exposed population is given by

$$R_t = (h_t + h_t e_t) \exp \left\{ - \sum_{l=1}^{t-1} (h_l^* + h_l e_l) \right\} \quad (16)$$

Assuming a maximum lifespan of 110 years, the lifetime lung cancer risk is given by the sum of the annual risks:

$$R = \sum_{t=1}^{110} R_t = \sum_{t=1}^{110} (h_t + h_t e_t) \\ \times \exp \left\{ - \sum_{l=1}^{t-1} (h_l^* + h_l e_l) \right\} \quad (17)$$

In the unexposed population, the $\{e_i\}$ are assumed to be zero, so that

$$R_0 = \sum_{t=1}^{110} h_t \exp \left\{ - \sum_{l=1}^{t-1} h_l^* \right\} \quad (18)$$

Note that the evaluation of the lifetime risk R in equation (17) depends on the excess relative lung cancer risks $\{e_i\}$ within each of the age-groups. The LRR is the ratio of the lifetime risk in the exposed population in relation to that in the unexposed population:

$$LRR = \frac{R}{R_0} \quad (19)$$

Other risks such as population-attributable risks can be characterized using the functional relationship between the risk factors. Using full characterization, the mean LRR is found to be 1.001 and the 95% CI is 1.008–1.578.

Discussion

Although expert opinion is still an important and integral part of risk assessment and risk management, risk assessment can no longer rely solely on unsubstantiated expert opinion as a form of risk assessment. Here we have summarized a general framework for the characterization of risk as a function of risk factors. In this probabilistic assessment, risk factors are considered as random variables with one or two layers of distributional assumptions. Using these distributions, we can not only estimate the risk for specific subset

of the population, but we can also provide confidence intervals.

As we have demonstrated in two examples, risk can be estimated for any population subset. The accuracy of the results largely depends on the functional relationship between risk factors and risk. Thus the distributional form and values of the parameters are also subject to criticism. Given these limitations, we still believe that a probabilistic risk assessment is required to move forward in this field.

Acknowledgments

This research was supported in part by the Cancer Center Support CA21765 from the National Institutes of Health and by the American Lebanese Syrian Associated Charities (ALSAC). The author thanks Dr. A Tyagi for expert suggestions and Dr. Angela McArthur for critical scientific editing. The author is also grateful to the Editor and a referee for providing many helpful suggestions that significantly improved the chapter.

References

- [1] Helton, J.C. (1994). Treatment of uncertainty in performance assessments for complex systems, *Risk Analysis* **14**, 483–511.
- [2] Bartlett, S., Richardson, G.M., Krewski, D., Rai, S.N. & Fyfe, M. (1996). Characterizing uncertainty in risk assessment – conclusions drawn from a workshop, *Human and Ecological Risk Assessment* **2**, 217–227.
- [3] National Research Council (1994). *Science and Judgment in Risk Assessment*, National Academy Press, Washington, DC.
- [4] National Research Council (1999). *Health Effects of Exposure to Low Levels of Ionizing Radiation Biological Effects of Ionizing Radiation (BEIR) VI*, National Academy Press, Washington, DC.
- [5] Finley, B. & Paustenbach, D. (1994). The benefits of probabilistic exposure assessment: 3 case studies involving contaminated air, water, and soil, *Risk Analysis* **14**, 53–74.
- [6] Sielken, R. & Valdez-Flores, C. (1996). Comprehensive realism's weight-of-evidence based distributional dose-response characterization, *Human and Ecological Risk Assessment* **2**, 175–193.
- [7] Allen, B.C., Crump, K.S. & Shipp, A. (1988). Correlation between carcinogenic potency of chemicals in animal and humans, *Risk Analysis* **8**, 531–544.
- [8] Bartlett, S., Krewski, D., Wang, Y. & Zielinski, J.M. (1993). Evaluation of error rates in large scale computerized record linkage studies, *Survey Methodology* **19**, 3–12.
- [9] French, S. (1995). Uncertainty and impression: modeling and analysis, *Journal of the Operational Research Society* **46**, 70–79.
- [10] Lubin, J.H. (1994). Invited commentary: lung cancer and exposure to residential radon, *American Journal of Epidemiology* **140**, 223–232.
- [11] Hoffman, F.O. & Hammonds, J.S. (1994). Propagation of uncertainty in risk assessments: the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability, *Risk Analysis* **14**, 707–712.
- [12] Rai, S.N., Krewski, D. & Bartlett, S. (1996). A general framework for the analysis of uncertainty and variability in risk assessment, *Human and Ecological Risk Assessment* **2**, 972–989.
- [13] Krewski, D., Rai, S.N., Zielinski, J. & Hopke, P.K. (1999). Characterization of uncertainty and variability characterization in residential radon cancer risks, *The Annals of the New York Academy of Sciences* **895**, 245–272.
- [14] Rai, S.N., Zielinski, J.M. & Krewski, D. (1999). Uncertainties in estimates of radon lung cancer risks, *Proceedings of Statistics Canada Symposium 99: Combining Data from Different Sources*, Ottawa, Ontario, Canada, pp. 99–107.
- [15] Kalbfleisch, J.D. & Prentice R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley & Sons, New York.

Related Articles

Role of Risk Communication in a Comprehensive Risk Management Approach

SHESH N. RAI

Risk Classification in Nonlife Insurance

Within the actuarial profession, a major challenge can be found in the construction of a fair tariff structure. In light of the heterogeneity within, for instance, a car insurance portfolio, an insurance company should not apply the same premium for all insured risks. Otherwise the so-called concept of adverse selection will undermine the solvability of the company. The idea behind risk classification is to split an insurance portfolio into classes that consist of risks with a similar profile and to design a fair tariff for each of them. Classification variables, typically used in motor third-party liability insurance, are the age and gender of the policyholder and the type and use of their car.

Being able to identify important risk factors is an important skill for the nonlife actuary. When these explanatory variables contain *a priori* correctly measurable information about the policyholder (or for instance, the vehicle or the insured building), the system is called an *a priori classification scheme*. However, an *a priori* system will not be able to identify all important factors because some of them cannot be measured or observed. Think, for instance, of aggressiveness behind the wheel or the swiftness of reflexes. Thus, despite the *a priori* rating system, tariffication cells will not be completely homogeneous. For that reason, an *a posteriori* rating system will reevaluate the premium by taking the history of claims of the insured into account.

Owing to the quantitative nature of both *a priori* and *a posteriori* rating, one of the primary attributes of an actuary (see **Actuary**) should be the successful application of up-to-date statistical techniques in the analysis of insurance data. Therefore, this article highlights current techniques involved in this area of actuarial statistics. The article introduces basic concepts, illustrates them with real-life actuarial data and summarizes references to complementary literature. Examples of likelihood-based as well as Bayesian estimation are included where the latter has the advantage that it provides the analyst with the full predictive distribution of quantities of interest.

Remark 1: Link with statistical techniques for loss reserving The statistical techniques discussed here in the context of risk classification (see **Risk**

Classification/Life Risk Measures and Economic Capital for (Re)insurers), also provide a useful framework for a stochastic approach of the loss reserving problem (see **Nonlife Loss Reserving**) in actuarial science. In this context, the data are displayed in a traditional run-off triangle or variations of it. See [1] and [2] for connections with reserving techniques.

Regression Models for *a priori* Risk Classification

In order to build a tariff that reflects the various risk profiles in a portfolio in a reasonable way, actuaries will rely on regression techniques. Typical response variables involved in this process are the number of claims (or the claims frequency) on the one hand and its corresponding severity (i.e., the amount the insurer will have to pay, given that a claim occurred) on the other hand.

Generalized Linear Models

The history of generalized linear models (GLMs) in actuarial statistics goes back to the actuarial illustrations in the standard text by McCullagh and Nelder [3]. See [4] for an overview in actuarial science. GLMs extend the framework of general (normal) linear models to the class of distributions from the exponential family. A whole variety of possible outcome measures (like counts, binary, and skewed data) can be modeled within this framework. This paper uses the canonical form specification of densities from the exponential family, namely,

$$f(y) = \exp\left(\frac{y\theta - \psi(\theta)}{\phi} + c(y, \phi)\right) \quad (1)$$

where $\psi(\cdot)$ and $c(\cdot)$ are known functions, θ is the natural and ϕ the scale parameter. Members of this family, often used in actuarial science, are the normal, the Poisson, the binomial, and the gamma distribution. Instead of a transformed data vector, GLMs model a transformation of the mean as a linear function of explanatory variables. In this way,

$$g(\mu_i) = \eta_i = (\mathbf{X}\boldsymbol{\beta})_i \quad (2)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ contains the model parameters and $\mathbf{X}$ ($n \times p$) is the design matrix, g is

the link function, and η_i is the i th element of the so-called linear predictor. In a likelihood-based approach, the unknown but fixed regression parameters in β are estimated by solving the maximum-likelihood equations with an iterative numerical technique (such as Newton–Raphson). In a Bayesian approach, priors are assigned to every parameter in the model specification and inference is based on samples generated from the corresponding full posterior distributions.

Illustration 1: Poisson regression for claims frequencies. For detailed case studies on Poisson regression for claim counts, [5] and [6] contain nice examples.

Flexible, Parametric Families of Distributions and Regression

Modeling the severity of claims as a function of their risk characteristics (given as covariate information) might require statistical distributions outside the exponential family, distributions with a heavy tail, for instance. Principles of regression within such a family of distributions are illustrated here with the Burr XII and the GB2 (“generalized beta of the second kind”) distribution. More details on Burr regression are in [7] and for GB2 regression [8] and [9] are useful.

Illustration 2: Fire insurance portfolio. The cumulative distribution function for the Burr type XII and the GB2 distribution are given by

$$F_{\text{Burr},Y}(y) = 1 - \left(\frac{\beta}{\beta + y^\tau} \right)^\lambda, \quad y > 0, \beta, \lambda, \tau > 0 \quad (3)$$

$$F_{\text{GB2},Y}(y) = B \left(\frac{(y/b)^a}{1 + (y/b)^a}; \gamma_1, \gamma_2 \right), \quad y > 0, a \neq 0, b, \gamma_1, \gamma_2 > 0 \quad (4)$$

where $B(\cdot, \cdot)$ is the incomplete beta function. Say, the available covariate information is in $\mathbf{x}$ ($1 \times p$). By allowing one or more of the parameters in equation (3) or equation (4) to vary with $\mathbf{x}$, a Burr or GB2 regression model is built. To illustrate this approach, consider a fire insurance portfolio (see [7]) that consists of 1823 observations. We want to assess how the loss distribution changes with the sum insured and the type of building. Claims expressed as a fraction of the sum insured are used as the response. Explanatory variables are the type of building and the sum insured. Parameters are estimated with maximum likelihood. Residual QQplots like those in Figure 1 can be used

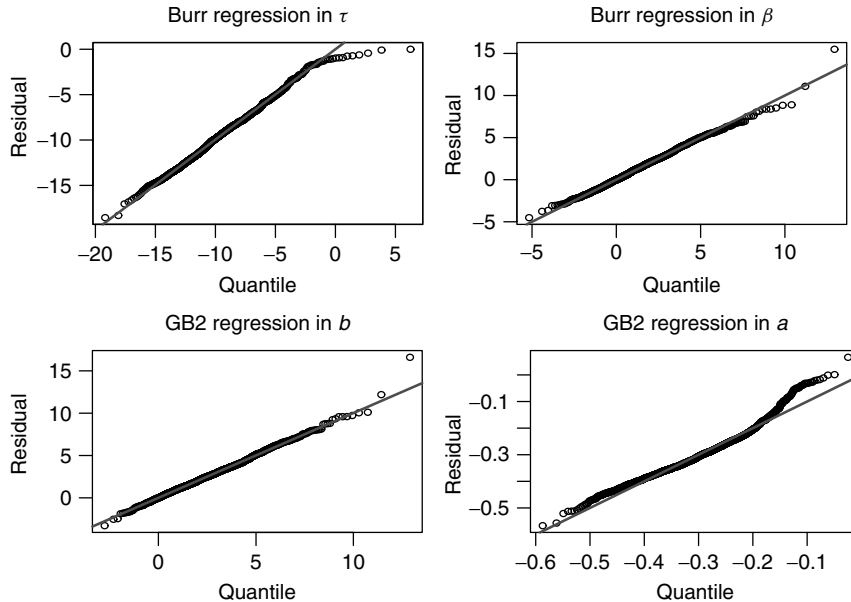


Figure 1 Fire insurance portfolio: residual QQplots for Burr and GB2 regression

to judge the goodness of fit of the proposed regression models. Beirlant *et al.* [7] explains their construction.

A *posteriori* Ratemaking

To update the *a priori* tariff or to predict future claims when historical claims are available over a period of insurance, actuaries can use statistical models for longitudinal or panel data (i.e., data observed on a group of subjects, over time). These statistical models (also known as *mixed models* or *random-effects models*) generalize the so-called credibility models from actuarial science. Bonus–malus systems (see **Bonus–Malus Systems**) for car insurance policies are another example of a *posteriori* corrections to *a priori* tariffs: insured drivers reporting a claim to the company will get a malus, causing an increase of their insurance premium in the next year. More details are, for instance, in [6].

The credibility ratemaking problem is concerned with the determination of a risk premium that combines the observed, individual claims experience of a risk and the experience regarding related risks. The framework of our discussion of this *a posteriori* rating scheme is the concept of generalized linear mixed models (GLMMs). For a historical, analytical discussion of credibility, [10] is a good reference. References [11–15] will provide additional background. References [16–18] discuss explicitly the connection between actuarial credibility schemes and (generalized) linear mixed models and contain more detailed examples. Reference [19] provides yet another statistical approach using copulas instead of random effects.

GLMMs extend GLMs by allowing for random, or subject-specific, effects in the linear predictor. Say we have a data set at hand consisting of N policyholders. For each subject i ($1 \leq i \leq N$) n_i observations are available. These are the claims histories. Given the vector $\mathbf{b}_i$ with the random effects for subject (or cluster) i , the repeated measurements $Y_{i1}, \dots, Y_{in_i}$ are assumed to be independent with a density from the exponential family

$$f(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) = \exp\left(\frac{y_{ij}\theta_{ij} - \psi(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right),$$

$$j = 1, \dots, n_i \quad (5)$$

The following (conditional) relations then hold

$$\mu_{ij} = E[Y_{ij}|\mathbf{b}_i] = \psi'(\theta_{ij}) \quad (6)$$

and

$$\text{Var}[Y_{ij}|\mathbf{b}_i] = \phi\psi''(\theta_{ij}) = \phi V(\mu_{ij}) \quad (7)$$

where $g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i$, $g(\cdot)$ is called the *link*, and $V(\cdot)$ the *variance function*. $\boldsymbol{\beta}$ ($p \times 1$) denotes the fixed effects parameter vector and $\mathbf{b}_i$ ($q \times 1$) the random effects vector. $\mathbf{x}_{ij}$ ($p \times 1$) and $\mathbf{z}_{ij}$ ($q \times 1$) contain subject i 's covariate information for the fixed and random effects, respectively. The specification of the GLMM is completed by assuming that the random effects, $\mathbf{b}_i$ ($i = 1, \dots, N$), are mutually independent and identically distributed with density function $f(\mathbf{b}_i|\boldsymbol{\alpha})$. Hereby $\boldsymbol{\alpha}$ denotes the unknown parameters in the density. Traditionally, one works under the assumption of (multivariate) normally distributed random effects with zero mean and covariance matrix determined by $\boldsymbol{\alpha}$. The random effects $\mathbf{b}_i$ represent unobservable, individual characteristics of the policyholder. Correlation between observations on the same subject arises because they share the same random effects. Reference [18] provides a discussion of likelihood-based and Bayesian estimation (see **Mathematics of Risk and Reliability: A Select History**) for GLMMs; with references to the literature and worked-out examples.

Illustration 3: Workers' compensation insurance. The data are taken from [20]. A total of 133 occupation or risk classes are followed over a period of 7 years. Frequency counts in workers' compensation insurance are observed on a yearly basis. Possible explanatory variables are **Year** and **Payroll**, a measure of exposure denoting scaled payroll totals adjusted for inflation. The following models are considered

$$Y_{ij}|\mathbf{b}_i \sim \text{Poisson}(\mu_{ij}) \quad (8)$$

where

$$\log(\mu_{ij}) = \log(\text{Payroll}_{ij}) + \beta_0 + \beta_1 \text{Year}_{ij} + b_{i,0} \quad (9)$$

versus

$$\log(\mu_{ij}) = \log(\text{Payroll}_{ij}) + \beta_0$$

$$+ \beta_1 \text{Year}_{ij} + b_{i,0} + b_{i,1} \text{Year}_{ij} \quad (10)$$

Table 1 Workers' compensation data (frequencies): results of maximum likelihood and Bayesian analysis. $\delta_0 = \text{Var}(b_{i,0})$, $\delta_1 = \text{Var}(b_{i,1})$, and $\delta_{0,1} = \delta_{1,0} = \text{Cov}(b_{i,0}, b_{i,1})$

	PQL		Adaptive G–H		Bayesian	
	Estimation	SE	Estimation	SE	Mean	90% Credibility interval
Model (9)						
β_0	−3.529	0.083	−3.557	0.084	−3.565	(−3.704, −3.428)
β_1	0.01	0.005	0.01	0.005	0.01	(0.001, 0.018)
δ_0	0.790	0.110	0.807	0.114	0.825	(0.648, 1.034)
Model (10)						
β_0	−3.532	0.083	−3.565	0.084	−3.585	(−3.726, −3.445)
β_1	0.009	0.011	0.009	0.011	0.008	(−0.02, 0.04)
δ_0	0.790	0.111	0.810	0.115	0.834	(0.658, 1.047)
δ_1	0.006	0.002	0.006	0.002	0.024	(0.018, 0.032)
$\delta_{0,1}$	—	—	0.001	0.01	0.006	(−0.021, 0.034)

Table 2 Workers' compensation data (frequencies): predictions for selected risk classes

Class	Actual Values		Expected number of claims					
			PQL		Adaptive G–H		Bayesian	
	Payroll _{<i>i</i>7}	Count _{<i>i</i>7}	Mean	SE	Mean	SE	Mean	90% Credibility interval
11	230	8	11.294	2.726	11.296	2.728	12.18	(5, 21)
20	1315	22	33.386	4.109	33.396	4.121	32.63	(22, 45)
70	54.81	0	0.373	0.23	0.361	0.23	0.416	(0, 2)
89	79.63	40	47.558	5.903	47.628	6.023	50.18	(35, 67)
112	18 810	45	33.278	4.842	33.191	4.931	32.66	(21, 46)

Hereby, Y_{ij} represents the j th measurement on the i th subject of the response Count. β_0 and β_1 are fixed effects and $b_{i,0}$, versus $b_{i,1}$, is a risk class specific intercept, versus slope. It is assumed that $\mathbf{b}_i = (b_{i,0}, b_{i,1})' \sim N(\mathbf{0}, \mathbf{D})$ and that, across subjects, random effects are independent. The results of both a maximum likelihood (penalized quasi-likelihood and adaptive Gauss–Hermite quadrature) and a Bayesian analysis are given in Table 1. The models were fitted to the data set without the observed Count_{*i*7}, to enable out-of-sample prediction later on. To illustrate prediction with model (10), Table 2 compares the predictions for some selected risk classes with the observed values. Predictive distributions obtained with a Bayesian analysis are illustrated in Figure 2.

Zero-Inflated, Additive, and Spatial Regression Models for *a priori* and *a posteriori* Ratemaking

This section contains references to some more advanced regression techniques for both cross-

sectional data (i.e., one observation per subject, as in section titled “Regression Models for *a priori* Risk Classification”) and panel data.

Dealing with regression models for claim counts, the huge number of zeros (i.e., no claim events) is often apparent. For data sets where the inflated number of zeros causes a bad fit of the regular Poisson or negative binomial distribution, zero-inflated regression models [21] provide an alternative. See [22] for a discussion of zero-inflated regression models for *a priori* classification schemes for count data.

In the modeling of severity data that consist of exact zeros and strictly positive payments, so-called two-part regression models are often used. They specify separate regression models for $Y = 0$ and $Y > 0$. See [2] for an illustration.

So far, only regression models with a linear structure for the mean or a transformation of the mean have been discussed. To allow for more flexible relationships between a response and a covariate, generalized additive models (GAMs) are available, see e.g. [2] for several actuarial examples. When the

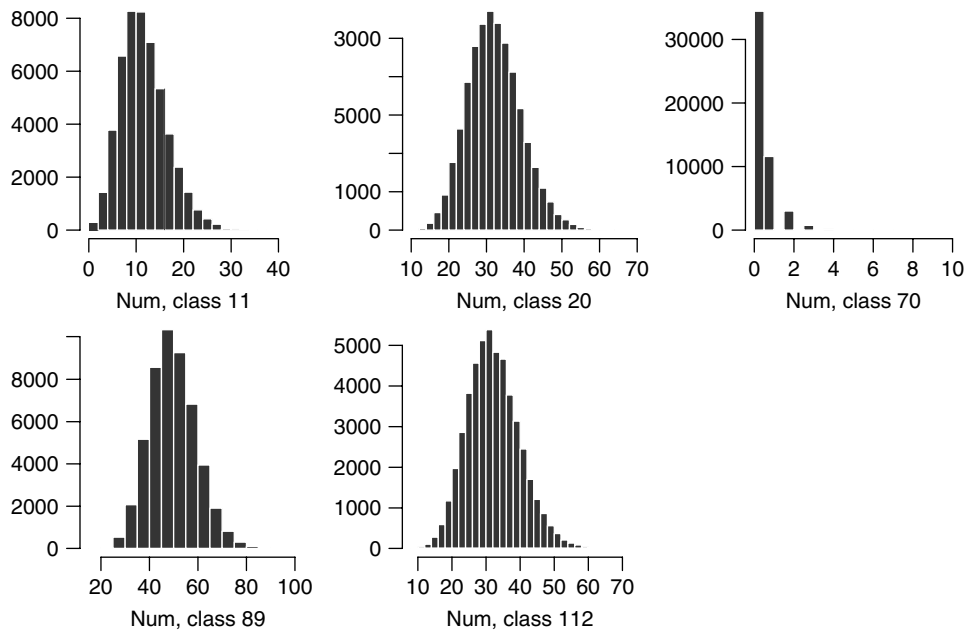


Figure 2 Workers' compensation insurance (counts): predictive distributions for selection of risk classes

place of residence of a policyholder is available as covariate information, spatial regression models (*see Statistics for Environmental Toxicity Geographic Disease Risk*) allow to bring this into account. See [23] for an illustration with spatial GAMs for cross-sectional observations on claim counts and severities.

Evidently, the mixed models for *a posteriori* ratemaking discussed above can be extended to zero-inflated, generalized additive mixed models (GAMMs) and spatial models to allow for more flexibility in model building. See e.g. [2] for an example with GAMMs.

References

- [1] Antonio, K., Beirlant, J., Hoedemakers, T. & Verlaak, R. (2006). Lognormal mixed models for reported claims reserving, *North American Actuarial Journal* **10**(1), 30–48.
- [2] Antonio, K. & Beirlant, J. (2006). *Issues in Claims Reserving and Credibility: A Semiparametric Approach with Mixed Models*, Working Paper, available online at <http://www.econ.kuleuven.be/insurance>.
- [3] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, *Monographs on Statistics and Applied Probability*, Chapman & Hall, New York.
- [4] Haberman, S. & Renshaw, A.E. (1996). Generalized linear models and actuarial science, *The Statistician* **45**(4), 407–436.
- [5] Denuit, M. & Charpentier, A. (2005). *Mathématiques de L'assurance Non-Vie: Tarification et Provisionnement (Tome 2)*, Economica, Paris.
- [6] Denuit, M., Marechal, X., Pitrebois, S. & Walhin, J.-F. (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Scales*, John Wiley & Sons.
- [7] Beirlant, J., Goegebeur, Y., Verlaak, R. & Vynckier, P. (1998). Burr regression and portfolio segmentation, *Insurance: Mathematics and Economics* **23**, 231–250.
- [8] Cummins, D.J., Dionnes, G., McDonald, J.B. & Pritchett, M.B. (1990). Applications of the GB2 family of distributions in modelling insurance loss processes, *Insurance: Mathematics and Economics* **9**, 257–272.
- [9] Sun, J., Frees, E.W. & Rosenberg, M.A. (2007). *Heavy-tailed longitudinal data modelling using copulas*, *Insurance: Mathematics and Economics*, in press.
- [10] Dannenburg, D.R., Kaas, R. & Goovaerts, M.J. (1996). *Practical Actuarial Credibility Models*, Institute of Actuarial Science and Econometrics, University of Amsterdam, Amsterdam.
- [11] Dionne, G. & Vanasse, C. (1989). A generalization of actuarial automobile insurance rating models: the negative binomial distribution with a regression component, *ASTIN Bulletin* **19**, 199–212.

- [12] Pinquet, J. (1997). Allowance for cost of claims in bonus-malus systems, *ASTIN Bulletin* **27**(1), 33–57.
- [13] Pinquet, J. (1998). Designing optimal bonus-malus systems from different types of claims, *ASTIN Bulletin* **28**(2), 205–229.
- [14] Pinquet, J., Guillén, M. & Bolancé, C. (2001). Allowance for age of claims in bonus-malus systems, *ASTIN Bulletin* **31**(2), 337–348.
- [15] Bolancé, C., Guillén, M. & Pinquet, J. (2003). Time-varying credibility for frequency risk models: estimation and tests for autoregressive specifications on random effects, *Insurance: Mathematics and Economics* **33**, 273–282.
- [16] Frees, E.W., Young, V.R. & Luo, Y. (1999). A longitudinal data analysis interpretation of credibility models, *Insurance: Mathematics and Economics* **24**(3), 229–247.
- [17] Frees, E.W., Young, V.R. & Luo, Y. (2001). Case studies using panel data models, *North American Actuarial Journal* **5**(4), 24–42.
- [18] Antonio, K. & Beirlant, J. (2007). Applications of generalized linear mixed models in actuarial statistics, *Insurance: Mathematics and Economics* **40**, 58–76.
- [19] Frees, E.W. & Wang, P. (2005). Credibility using copulas, *North American Actuarial Journal* **9**(2), 31–48.
- [20] Klugman, S. (1992). *Bayesian Statistics in Actuarial Science with Emphasis on Credibility*, Kluwer, Boston.
- [21] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics* **34**, 1–14.
- [22] Boucher, J.-P., Denuit, M. & Guillén, M. (2005). *Risk Classification for Claim Counts: Mixed Poisson, Zero-Inflated Mixed Poisson and Hurdle Models*, Working Paper, available online at <http://www.actu.ucl.ac.be>.
- [23] Denuit, M. & Lang, S. (2004). Nonlife ratemaking with Bayesian GAMs, *Insurance: Mathematics and Economics* **35**(3), 627–647.

KATRIEN ANTONIO AND JAN BEIRLANT

Risk Classification/Life

Risk Pooling, Heterogeneity, Risk Factors

The traditional principle of insurance management, the pooling of risks (*see* **Dependent Insurance Risks; Correlated Risk**), relies on the availability of a large number of similar insured lives, so that there is a high probability that the actual outcome in the portfolio is close to what was expected. In a properly managed insurance business, individual policies are important not by themselves but in terms of the risk sharing opportunities, which they provide to the portfolio.

Insurance companies use the underwriting process to select the lives to be insured and to detect differences among the insureds, given that unidentified heterogeneity introduces biases, which will undermine the realization of the pooling effect.

As a result of the underwriting process, policies are usually grouped according to the different levels of the important aspects affecting mortality, the so-called risk factors. A specific premium rate is then calculated for each group, which is in particular based on a sound set of mortality rates (*see* **Pricing of Life Insurance Liabilities; Estimation of Mortality Rates from Insurance Data**).

When referring to life insurance, among the major risk factors, we can list [1] the following:

1. biological and physiological factors, such as age, gender, genotype;
2. features of the living environment, in particular with regard to climate and pollution, nutrition standards (in particular, excesses and deficiencies in diet), population density, and hygienic and sanitary conditions;
3. occupation, in particular in relation to exposure to injury or adverse conditions, and educational attainment;
4. individual lifestyle, in particular with regard to nutrition, alcohol and drugs consumption, smoke, physical activities, and pastimes;
5. current health conditions, personal and/or family medical history, civil status, and so on.

Item 2 above affects the overall mortality of the population living in a given region. As such, it is not an explicit risk rating variable; it is only implicitly

considered given that the mortality table referred to for insurance rating in a market usually reflects the general features of mortality in the relevant region (*see* **Individual Risk Models**). The remaining items concern the individual and, when dealing with life insurance, they can be observed at policy issue. Their assessment is performed through questions in the application form and, as to personal health conditions, possibly through a medical examination.

The specific items considered for rating depend on the types of benefits provided by the insurance contract. Age is always considered, and can be traced back to the introduction of scientific life insurance by Dodson in 1756 [2]. Thus the probability of dying within 1 year for a person aged x (also called the *mortality rate*) is usually denoted as q_x . Gender is usually accounted for, especially when living benefits are involved, given that females on average live longer than males. In this regard, distinct mortality tables for males and females are normally adopted. As far as genetic aspects are concerned, the evolving knowledge in this area has raised a lively debate (which is still open) on whether it is legitimate for insurance companies to resort to genetic tests for underwriting purposes. So far, this aspect has not entered the rating process.

A distinctive phenomenon in insurance is adverse selection (or self-selection). Adverse selection is the “tendency of high risks to be more likely to buy insurance or to buy larger amounts of insurance than low risks” [3]. An individual has a level of insurance risk and a particular level of demand for insurance, which may be correlated with the risk level, and which itself may be influenced by a range of factors. The presence of adverse selection provides the conventional economic argument in favor of underwriting and risk classification. Thus, those individuals applying for living benefits are usually in good health conditions, while those buying a death benefit may be in poor health conditions. In both cases, the mortality experienced by a life insurer may be different from the mortality observed in the general population, in particular lower in the former case of living benefits, higher in the latter case of death benefits. For this reason, a thorough investigation is conducted to check individual health conditions when death benefits are involved. Clearly, the level of accuracy of such an investigation depends on the amounts of insurance involved (*see* **Dynamic Financial Analysis**).

When death benefits are dealt with, the main risk factors considered are health conditions, occupation, and smoking status. According to the level of these factors, the risks are classified into standard and substandard risks. For the latter (also referred to as impaired lives or extra risks), a higher premium rate is adopted, given that they have a higher probability of becoming eligible for the benefit. In some markets, the standard risks are further split into regular and preferred risks, the latter having a better profile than the former (for example, because they have never smoked); as such, the latter are allowed to pay a reduced premium rate.

As far as annuities and living benefits are concerned, adverse selection leads to reduced mortality rates (relative to the general population) and these are adopted for rating annuities. There is also evidence that mortality rates are lower for annuity products (see **Longevity Risk and Life Annuities**) where the purchase is voluntary rather than compulsory. In recent years, market innovations have led to the introduction of special (i.e., favorable) annuity prices for those lives assessed to have reduced life expectancies (as introduced into the UK market in 1995).

A similar approach to assessment of risk factors and underwriting takes place in the growing areas of disability, health, critical illness, and long-term care insurance [4].

Differential Mortality

The mortality experienced by people in poorer or better conditions than the average is usually expressed in relation to average (or standard) mortality. This allows us to deal only with one mortality table (or law), from which both normal and adjusted premium rates for substandard or preferred risks may be obtained. In the case of annuities and pensions, usually specific tables are constructed, which allow for both adverse selection and the secular downward trend in mortality rates at adult and old ages. This latter feature is incorporated in order to avoid the underestimation of future costs.

Let us index with (S) a quantity expressing standard mortality and with (D) a quantity expressing a different (higher or lower) mortality. According to the usual notation, $q_x^{(S)}$ denotes the mortality rate at age x (either for a male or a female), while $\mu_x^{(S)}$ represents the force of mortality at age x (also in this

case gender-specific); further, we understand that x is the current age and t is the time elapsed since policy issue ($t \geq 0$). The following are examples of models expressing differential mortality, i.e., mortality for either substandard risks, preferred lives or annuitants (details are provided below).

$$q_x^{(D)} = a q_x^{(S)} + b \quad (1)$$

$$\mu_x^{(D)} = a \mu_x^{(S)} + b \quad (2)$$

$$q_x^{(D)} = q_{x+k}^{(S)} \quad (3)$$

$$\mu_x^{(D)} = \mu_{x+k}^{(S)} \quad (4)$$

$$q_x^{(D)} = \psi(x, q_x^{(S)}) \quad (5)$$

$$q_x^{(D)} = q_x^{(S)} \cdot \phi(x) \quad (6)$$

$$q_{[x-t]+t}^{(D)} = \rho(x-t, t) \cdot q_x^{(S)} \quad (7)$$

$$q_{[x-t]+t}^{(D)} = q_{[x-t]}^{(S)} \cdot \nu(t) \quad (8)$$

$$q_{[x-t]+t}^{(D)} = q_{[x-t]+t}^{(S)} \cdot \eta(x, t) \quad (9)$$

Models (1) and (2) are usually adopted for substandard risks; adoption of equation (1) or equation (2) depends on the information available for standard mortality, either a mortality table or the parameters of a given law for the force of mortality. Letting $a = 1$ and $b = \delta q_{x-t}^{(S)}$, $\delta > 0$, in equation (1) ($b = \delta \mu_{x-t}^{(S)}$ in equation (2)) the so-called additive model is obtained, where the increase applied to mortality depends on the initial age and, for a given individual, is therefore constant. An alternative model is obtained by choosing $b = \theta$, $\theta > 0$, in equation (2), i.e., a mortality increase that is constant and independent of the initial age; such a model is consistent with extra mortality due to accidents (related, for example, either to occupation or the extreme sports). Letting $a = 1 + \gamma$, $\gamma > 0$, and $b = 0$ the so-called multiplicative model is derived, where the mortality increase depends on current age and therefore changes (typically, increases) with time; in the case of equation (2), this model is widely known as the “proportional hazards” model. When risk factors are only temporarily effective (for example, extra mortality due to some diseases, which either lead to an early death or recovery in a short time), parameters a , b may be positive up to some proper time τ ; and for $t > \tau$, standard mortality is assumed, so that $a = b = 0$.

Models (3) and (4) are very common in practice, both for substandard and preferred risks, due to their

simplicity; they are called *age rating* or *age shifting models*. Assuming, as is evidenced by mortality experience, that after some (young) age mortality rates (and the force of mortality, as well) are increasing with respect to age, the idea is to assign, for rating purposes, an older initial age to substandard risks and a younger initial age to preferred lives. This adjustment process results in a higher (lower) premium rate. Model (4), in particular, can be formally justified if the standard force of mortality is driven by the Gompertz law. Actually, let $\mu_x^{(S)} = \beta \cdot c^x$. Assume the multiplicative model for differential mortality. Then we have $\mu_x^{(D)} = (1 + \gamma) \cdot \beta \cdot c^x$. Let k be the real number (unambiguously defined) such that $1 + \gamma = c^k$. We obtain

$$\mu_x^{(D)} = \beta \cdot c^{x+k} = \mu_{x+k}^{(S)} \quad (10)$$

Adoption of equation (4) under laws other than the Gompertz or the adoption of model (3) represent rules that are not formally justified, albeit supported by equation (10). In actuarial practice, sometimes the age shifting is also applied directly to the premium rate.

In equations (5) and (6), mortality is adjusted in relation to age. Such a choice is common when annuities are dealt with. For example, $\phi(x)$ may be a step-wise linear function, with $\phi(x) < 1$. In such an application, $q_x^{(S)}$ describes mortality in the general population (usually projected, i.e., embedding a forecast of future trends), while $\phi(x)$ is a reduction factor accounting for self-selection.

The notation $q_{[x-t]+t}^{(\cdot)}$ is usually adopted when mortality depends not only on current age, x , but also on the time elapsed since policy issue, t . Such dependence is, in particular, observed in portfolios of death benefits; because of the medical selection, insured risks may experience better mortality than the average in initial years. Such dependence is also occasionally observed in annuity portfolios. The effect tends to reduce as time passes on. So we expect $q_{[x]}^{(\cdot)} < q_{[x-1]+1}^{(\cdot)} < \dots$. Mortality rates $q_{[x-t]+t}^{(\cdot)}$ are called *issue-select*, while the $q_x^{(\cdot)}$'s (dependent on current age only) are the so-called *aggregate mortality rates*. Model (7) expresses issue-select mortality in terms of aggregate mortality; typically, $\rho(x - t, t) < 1$, and so, it represents a reduction factor. Sometimes, for $t \geq \tau$ (where τ represents the length of the period in which initial selection affects mortality) the reduction factor is simply adopted dependent on current age,

i.e., $\rho(x - t, t) = \tilde{\rho}(x)$. Models (8) and (9) express issue-select differential mortality through a transform of the issue-select standard mortality rates. In particular, $v(x)$ and $\eta(x, t)$ may be chosen to be linear.

For details on the models listed above, as well as on more specific models, the reader may refer for example, to [1].

A particular implementation of model (1) (with $b = 0$) is given by the so-called “numerical rating system”, introduced in 1919 by New York Life Insurance and still adopted by many insurers [5]. If we assume that there is a set of m risk factors, then the mortality rate specific for a given individual is

$$q_x^{(\text{spec})} = q_x^{(S)} \cdot \left(1 + \sum_{h=1}^m \rho_h \right) \quad (11)$$

where the rates ρ_h lead to a higher or lower mortality rate for the individual in relation to the values assumed by the chosen risk factors (clearly, $\sum_{h=1}^m \rho_h > -1$). Note that an additive effect of the risk factors is assumed. For further details, see [3].

The Frailty

Once policies have been grouped according to the different levels of the relevant risk factors, there is still heterogeneity within the groups (*see Life Insurance Markets*). Such heterogeneity may be due both to risk factors which, for commercial reasons, have not been considered for rating purposes and to features attributable to unobservable risk factors (such as, for example, the individual's attitude towards health, the congenital propensity to survive or die or to become ill). In respect of the latter set Vaupel *et al.* [6] have introduced the idea of frailty, which is defined as a nonnegative quantity whose level expresses the unobservable risk factors affecting individual mortality. The underlying idea is that those with higher frailty die on average earlier than others. The mortality model (*see Hazard and Hazard Ratio; Logistic Regression*) is specified in terms of the force of mortality.

Let us refer to a cohort, which is homogeneous in respect of observable risk factors, but heterogeneous because of the different individual levels of frailty. The specific value of the frailty of the individual does not change in time, while remaining unknown. However, because of the deaths that

occur, the distribution of people with regard to frailty changes with age, given that people with low frailty are expected to live longer. We denote with Z_x the relevant random variable at age x , for which a continuous density function $g_x(z)$ is assumed. It must be mentioned that the hypothesis of constant individual frailty, which is reasonable when considering genetic aspects, seems weak when referring to environmental factors, which may change in time affecting the risk of death. However, empirical evidence does validate this assumption quite satisfactorily.

For a person of current age x and frailty level z , the force of mortality is defined as

$$\mu_x(z) = \lim_{t \rightarrow 0} \frac{\Pr(T_x \leq t | Z_x = z)}{t} \quad (12)$$

where, according to standard notation, T_x represents the random lifetime at age x .

A very simple link between $\mu_x(z)$ and a standard force of mortality is a multiplicative relation, as suggested by Vaupel *et al.* [6]. If μ_x represents the standard force of mortality, we can assume

$$\mu_x(z) = z \cdot \mu_x \quad (13)$$

Note that if $z > 1$ ($z < 1$) then $\mu_x(z) > \mu_x$ ($\mu_x(z) < \mu_x$). An individual with $z = 1$ has a standard mortality profile, while those with higher (lower) frailty bear a higher (lower) probability of early death.

The complete definition of the mortality model requires (a) the density function of the frailty at age 0 and (b) the law of the force of mortality. For full details, the reader should refer to [6]; here we just comment on some of the main results relevant for insurance applications.

The distribution of the frailty changes with time (i.e., with increasing age) because of the varying composition of the cohort due to deaths. Thanks to the multiplicative assumption in equation (13), the density function of Z_x can be derived from that of Z_0 by noting that $Z_x = Z_0 | T_0 > x$. Let $S(x|z)$ denote the survival function for a person with frailty z ; similarly to the standard valuation, we have

$$S(x|z) = e^{-\int_0^x \mu_t(z) dt} = e^{-z \int_0^x \mu_t dt} \quad (14)$$

It is easy to derive the following result for the density function of Z_x

$$g_x(z) = \frac{S(x|z) \cdot g_0(z)}{\int_0^\infty S(x|z) \cdot g_0(z) dz} \quad (15)$$

The expected frailty at age x is

$$\bar{z}_x = \int_0^\infty z g_x(z) dz \quad (16)$$

If we define the average force of mortality in the population at age x as

$$\bar{\mu}_x = \int_0^\infty \mu_x(z) \cdot g_x(z) dz \quad (17)$$

then it can be shown that, under the assumptions introduced above,

$$\bar{\mu}_x = \mu_x \cdot \bar{z}_x \quad (18)$$

Let us comment on the above-mentioned results in relation to possible insurance applications. First of all, it should be stressed that the frailty model does not provide us with a specific premium rate for each individual (due to the unknown individual frailty level); rather, some information concerning the aggregate mortality in a heterogeneous cohort (*viz* $\bar{\mu}_x$) can be obtained.

We note that the average frailty $\bar{z}_x$ is decreasing with age (this can be shown formally, but it is an intuitive result, due to the fact that those with a high frailty die earlier). This implies that, from a given (young) age onwards $\bar{\mu}_x$ increases with age but more slowly than the standard force of mortality μ_x . So, in an insured portfolio, where risks are selected through the underwriting process or are self-selected (especially in the case of life annuities), one should expect mortality experience to be on average lower than that observed in the general population. This is why, when living benefits are dealt with, reduction factors should be applied to the mortality rates relating to the general population, as was mentioned earlier. In the case of death benefits, adoption of mortality tables relating to the general population leads to the inclusion of a safety loading in premium rates for standard risks.

A further issue, relevant when life annuities are dealt with, concerns the shape of mortality rates in a given population at the oldest ages. When referring to

a time-continuous setting, the usual assumption in life insurance is that the (standard) force of mortality is exponential with respect to age (typically, a Gompertz or Makeham model is adopted). If one disregards the changing distribution of frailty in a group of people, then the exponential assumption relates to any age. When allowing for frailty, different conclusions may be obtained. For example, if one assumes a gamma distribution for the frailty and the Gompertz model for the standard force of mortality, then the following result holds for the average force of mortality:

$$\bar{\mu}_x = \frac{\beta c^x}{1 + \delta c^x} \quad (19)$$

where parameters β , δ , c are properly defined in terms of the parameters adopted for the gamma distribution and the Gompertz law. It must be stressed that according to equation (19) the average force of mortality in the group has a logistic shape. In particular, this affects the shape of mortality at the oldest ages, which is flatter than otherwise assumed.

Several generalizations of the original frailty model described by Vaupel *et al.* [6] have been discussed in demographic literature. For some investigations concerning mortality in insured portfolios, see in particular [7], where possible extensions to the basic frailty model are recalled.

We finally mention an alternative procedure to obtain equation (19), which is interesting within the insurance area, due to its similarity to the Poisson gamma model used for detecting heterogeneity in nonlife insurance portfolios. We refer to a population where the force of mortality for an individual, $\mu_x^{(\cdot)}$, is Gompertz-like; more specifically, we assume

$$\mu_x^{(\cdot)} = A \cdot c^x \quad (20)$$

where the parameter c , expressing the rate of mortality increase in respect of age, is known and common to all individuals, while the parameter A , expressing

base mortality, is unknown and specific to the life considered. If $f(\alpha)$ is the density function of A , then the average force of mortality in the population can be calculated as

$$\bar{\mu}_x = \int_0^\infty \alpha \cdot c^x \cdot f(\alpha) d\alpha = c^x \cdot E(A) \quad (21)$$

If a gamma distribution is assumed for A , with parameters such that $E(A) = \beta/(1 + \delta c^x)$, then equation (19) follows. This procedure was first described by Beard [8]; see also [3] for further details.

References

- [1] Benjamin, B. & Pollard, J.H. (1993). *The Analysis of Mortality and Other Actuarial Statistics*, The Institute of Actuaries, Oxford.
- [2] Haberman, S. & Sibbett, T. (1995). *History of Actuarial Science*, William Pickering, London.
- [3] Cummins, J.D., Smith, B.D., Vance, R.N. & Van Derhei, J.L. (1983). *Risk Classification in Life Insurance*, Kluwer-Nijhoff Publishing, Boston/The Hague/London.
- [4] Booth, P., Chadburn, R., Haberman, S., James, D., Khorasane, Z., Plumb, R.H. & Rickayzen, B. (2005). *Modern Actuarial Theory and Practice*, 2nd Edition, Chapman & Hall/CRC, Boca Raton.
- [5] Rogers, O. & Hunter, A. (1919). Numerical rating system. The numerical method of determining the value of risks for insurance, *Transactions of the Actuarial Society of North America* **20**, 273–300.
- [6] Vaupel, J.W., Manton, K.G. & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography* **16**, 439–454.
- [7] Butt, Z. & Haberman, S. (2004). Application of frailty-based mortality models using generalized linear models, *ASTIN Bulletin* **34**, 175–197.
- [8] Beard, R.E. (1971). Some aspects of theories of mortality, cause of death analysis, forecasting and stochastic processes, in *Biological Aspects of Demography*, W. Brass, ed, Taylor & Francis, London, pp. 57–68.

STEVEN HABERMAN AND ANNAMARIA
OLIVIERI

Risk from Ionizing Radiation

Radiation risk commonly refers to *ionizing* radiations: energetic photons such as γ and X rays and various accelerated subatomic particles such as alphas, betas, and neutrons, in contrast to other, non-ionizing forms of directed energy such as electromagnetic radiations with quantum energies less than about 1000 eV (i.e., wavelength greater than about 1 nm) [1], sonic energy, etc. Humans are exposed to ionizing radiation from a number of sources: natural background radiation (including *radon* (see **Radon**)), medical diagnosis and therapy, nuclear power generation, occupations involving radioactive materials (mining, nuclear workers, hospital, laboratory technologists and radiologists, etc.), nuclear weapons testing, and occasional accidents. Because radon is covered separately in this encyclopedia, we do not address that area of research here.

The ionizing radiations share fundamentally important biophysical properties from the nature of the damage to living organisms that can originate with ionization of molecules in cells. The nonionizing radiations in contrast are typically treated separately according to category: i.e., microwave *versus* ultraviolet radiation, etc. and involve different possible biophysical mechanisms. Since the early decades of the twentieth century, ionizing radiations have been associated, through animal experiments and human experience with accidental or (originally) poorly understood excessive doses, with damage to tissue, genetic mutations, and neoplastic lesions including cancer. Studies of radiation effects are based on *radiation dose* received by tissues of affected individuals: the amount of ionizing energy deposited per unit mass of tissue; a basic unit is $1 \text{ Gy} = 1 \text{ J kg}^{-1}$. Among ionizing radiations, some create much more closely spaced ionizations at a microscopic level than others: heavier particulate radiations such as neutrons and alphas are much more densely ionizing than electrons and γ rays. A radiation's density of ionization, or "quality," as quantified by linear energy transfer (LET), characteristic of its type and energy, is strongly associated with its biological effect per unit dose (radiobiological effectiveness, RBE). The estimation of radiation dose (radiation dosimetry) for

major health effects studies was recently reviewed in a special issue of *Radiation Research* [2].

Effects of ionizing radiation have typically been divided broadly into acute effects and late effects. Acute effects are nonstochastic outcomes that occur with certainty but have a *severity* that depends on the radiation dose, although there is typically a practical dose threshold, below which the severity is not sufficient to be distinguished from no effect, and there is variation in radiosensitivity among individuals. In contrast, most late effects are seen as stochastic events that occur with a *probability* that depends on radiation dose, among many other covariates. Acute effects are forms of morbidity due to cell killing and are mainly of interest at higher doses. As "risk assessment" generally denotes evaluating the probability of stochastic late effects rather than the likely severity of acute effects, we focus here on late effects. These do not have any unique attributes associated with the involvement of radiation in their etiology, and can only be discerned statistically as increased rates associated with exposure in studies on groups of people with varying exposure. The effect of a covariate such as attained age or age at exposure, on the risk of a stochastic effect per unit radiation dose, is termed *risk modification*. Most work has focused on cancer and other neoplastic diseases such as leukemia, along with genetic effects, although some has been aimed at other effects such as nonspecific life span shortening. More recently, the occurrence of chronic diseases other than cancer has been associated as a late effect with ionizing radiation [3]. A few outcomes, such as cataract, have aspects of both acute and late effects.

Radiation Risk Estimation

Estimates of radiation risk are based upon epidemiologic studies of populations exposed to ionizing radiation: the atomic bomb (A-bomb) survivors, persons exposed medically or occupationally (particularly nuclear industry workers and radiologists/radiological technologists), and persons exposed environmentally (e.g., high-background areas, areas near the Chernobyl nuclear reactor, and areas affected by large releases of radioactive material from nuclear weapons production and testing). See biological effects of ionizing radiation BEIR VII [3] and United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2000 [4] for complete listings. These

studies involve a wide range of radiations, doses, and configurations (whole body *versus* localized exposure; acute *versus* chronic or protracted exposure). In all of these studies, the usual epidemiologic considerations of comparison population and counterfactual hypotheses apply [5]. The A-bomb survivor follow-up study of approximately 120 000 individuals in Hiroshima and Nagasaki, Japan, who were exposed acutely to primarily γ rays, constitutes the longest follow-up study on late effects of radiation and arguably the most informative analysis of radiation risk in terms of assessing risk modification and investigating causal mechanisms. Even in that study, inference on risk is sensitive to choice of an unexposed comparison group from a possible source population that resided over a wide geographic range (urban *versus* rural environment, etc.), within the cities but too distant from the sites of bombs to have an appreciable dose, including 32 000 control persons who were temporarily away from the cities at the time of the bombings [6, 7].

Current analyses of radiation risk rely heavily on ideas originally proposed by Thomas [8], which allow flexibility in modeling the shape of the dose–response. Risk estimation involves not only estimation of a risk coefficient (e.g., relative risk (RR)), but also requires accounting for spontaneous (background) rates of disease or death and assessment of effect modification by age, time, gender, and other factors. Perhaps the most common form of risk model is the *excess relative risk* (ERR) model:

$$r_{\beta}(d, \mathbf{z}) = 1 + \rho_{\beta}(d) \exp(\boldsymbol{\gamma}\mathbf{z}) \quad (1)$$

where $r_{\beta}(d, \mathbf{z})$ is the relative risk, $\rho_{\beta}(d)$ is the ERR at dose d depending on parameter β , and $\exp(\boldsymbol{\gamma}\mathbf{z})$ represents effect modification by factor $\mathbf{z}$ (typically $\boldsymbol{\gamma}\mathbf{z} = \gamma_1 \times \text{gender} + \gamma_2 \times (\text{exposure age}) + \gamma_3 \times (\text{age at risk})$, although time since exposure could be used; note that there are problems in the interpretation of age–time risk modification owing to collinearity among the three age–time scales, see [9]).

The simplest ERR model is the linear ERR model: $\rho_{\beta}(d) = \beta d$ (β is the ERR per unit dose). The RR – the ratio of the hazard of disease or death at nonzero dose to the baseline hazard – is estimated in the context of a *multiplicative model*: $\lambda_{\beta}^{\text{RR}}(\mathbf{t}, \mathbf{s}, d, \mathbf{z}) = \lambda^0(\mathbf{t}, \mathbf{s}) r_{\beta}(d, \mathbf{z})$, with $\mathbf{t}$ being a generic timescale variable denoting both attained age (*age at risk*) and calendar time, $\mathbf{s}$ being a collection of factors impacting the background rate (e.g., gender, geographic location, lifestyle and genetic factors, etc.;

the factors $\mathbf{s}$ and $\mathbf{z}$ can have elements in common), and λ^0 being the baseline hazard.

Another measure of risk often of interest is the excess absolute rate (EAR) of disease or death defined by the *additive model*

$$\lambda_{\varphi}^{\text{EAR}}(\mathbf{t}, \mathbf{s}, d, \mathbf{z}) = \lambda^0(\mathbf{t}, \mathbf{s}) + \rho_{\varphi}(d) \delta(\mathbf{t}, \mathbf{s}, \mathbf{z}) \quad (2)$$

where $\rho_{\varphi}(d) \delta(\mathbf{t}, \mathbf{s}, \mathbf{z})$ is the excess rate and $\rho_{\varphi}(d)$ can be linear [$\rho_{\varphi}(d) = \varphi d$] or otherwise [10]. Because the RR $r_{\beta}(d, \mathbf{z}) \equiv \lambda_{\varphi}^{\text{EAR}}(\mathbf{t}, \mathbf{s}, d, \mathbf{z}) / \lambda^0(\mathbf{t}, \mathbf{s})$, the additive risk model equates to the multiplicative model if the added risk due to radiation has the same functional form with respect to covariates $(\mathbf{t}, \mathbf{s})$ as the baseline risk, i.e., if $\delta(\mathbf{t}, \mathbf{s}, \mathbf{z}) = C \lambda^0(\mathbf{t}, \mathbf{s})$, where C is some constant. Although this is often not the case, as excess rates can depend on age and time in ways that are not proportional to the background rates, given sufficient flexibility in their respective risk-modifying terms the two models are equivalent in terms of goodness of fit to the data, so they merely provide alternative descriptions of the same underlying process. Choice of the model to use in describing the rates is therefore a matter of preference as to how one wishes to describe the radiation effects, and not a matter of statistical model selection. With either model, appropriate modeling of background rates is important for making inference about risk modification (the term $\exp(\boldsymbol{\gamma}\mathbf{z})$ in equation (1) or $\delta(\mathbf{t}, \mathbf{s}, \mathbf{z})$ in equation (2)), because improperly controlled covariate effects on background rates can bias estimates of effect modification [11].

For large cohorts, it is common to estimate risk using the grouped survival approach with Poisson regression [12]. Briefly, if a hazard $\lambda(t)$ can be assumed to be sufficiently small over suitable intervals of age and time conditional on factors $\mathbf{s}$, so that only a fraction of the cohort members are expected to evidence outcome in the interval, it is a reasonable approximation to treat the number of cases of disease or death in age/time/factor stratum i as a Poisson distributed variable with mean $P_i \lambda_i^*(\mathbf{t}_i, \mathbf{s}_i, d_i, \mathbf{z}_i)$, where P_i is the total person time in the stratum, $\mathbf{t}_i$ denotes the person year weighted average age and time values in the stratum, $\mathbf{s}_i$ and $\mathbf{z}_i$ denote the stratum-specific person year weighted average values of background and effect-modifier variables defined above, and the $*$ denotes the piecewise constant rate. Regression analysis of Poisson rate data can be handled using the generalized linear model (GLM) framework with maximum likelihood estimation using iteratively

reweighted least squares (IRLS) [13]. It produces results close to those from continuous-data methods when the piecewise constant hazard assumption is approximately true (i.e., when the strata are sufficiently small the rates do not change appreciably within strata [14]). To accommodate the ERR model nonlinear fitting procedures are required; these are easily accommodated with the *Epicure* software package [15]. The *Epicure* software also allows general modeling of the joint effects of multiple risk factors, including risk modification by factors other than age and time (e.g., smoking; see the next section). Statistical significance of parameter estimates is typically assessed using likelihood ratio tests and likelihood-ratio-based confidence intervals. Goodness of fit may be examined using deviance residuals.

In studies of radiation risk, as with epidemiologic cohort studies in general, there is nonnegligible measurement error (random error or uncertainty) in the exposure estimate. In the A-bomb survivor analyses based on Poisson rate regression, replacement methods assuming classical measurement error have been used [16]; these impute the expected value of true, unknown dose given the measured value and conditional on various assumptions about the error distribution relating true and measured dose as well as the unknown distribution of true dose. Assumptions may be relaxed by using nonparametric methods to estimate the true-dose distribution [17]. Other methods of dealing with measurement error in risk estimation are also available, such as models based on structural equations in which the true, unknown dose can be integrated out of a full likelihood, or the marginal relationship of the measured dose and the outcome can be estimated without explicit modeling of the true, unknown dose, based on certain assumptions (see, e.g. [18]).

Risk estimation may be based alternatively on individual follow-up data, using e.g., Cox regression. A disadvantage of Cox regression is the computing time required with large cohorts having extensive amounts of follow-up time and many failures; this is precisely the situation in which the Poisson approximation is especially good. For smaller cohorts with relatively small total follow-up time the person-year stratification may not be so fine and the assumption that rates do not change appreciably within strata may not hold; with such data, however, the expected number of events will be relatively low so that Cox regression is a tractable approach. Cox regression

also allows the researcher to leave the mathematical structure of background rates unspecified, but this is merely a matter of preference as simple descriptive models in age and time generally capture these effects well with sufficient data. The likelihood for individual data may be extended to include mechanistic models of carcinogenesis such as multistage models [19] and two-stage clonal expansion (TSCE) models [20].

Radioepidemiological Tables and Application of Risk Estimates to Radiation Protection

In addition to ERR or EAR, other quantities are of interest in radiation protection. These include the *excess lifetime risk* and *lifetime detriment* (loss of life expectancy [21]), as well as the *probability of causation* (PC assigned shares, or excess fraction; see “radioepidemiological tables” [22]). Excess lifetime risk – the increase in lifetime probability of death from a particular cause – is complicated by competing causes related to exposure [23]. Radiation protection guidelines are often formulated using measures such as excess lifetime risk. As this integrates over all possible attained ages at expression of the health effect in question, it involves assumptions about the composition of the population to which the standards are to be applied, with respect to the effect-modifier variables defined above; this is an issue more generally in transporting or generalizing risk estimates from a study population to other populations.

Estimation of causal probabilities is not generally possible in epidemiologic studies without strong and untestable assumptions [24]; nevertheless, the primary parameter used to approximate PC is the attributable fraction, $ERR/RR = (RR - 1)/RR = ERR/(1 + ERR)$. In the face of other risk factors, which may interact with radiation, PC should be calculated conditionally on the level of the other factor; this is implemented in the radioepidemiological tables in the case of smoking and lung cancer (see also [25]). With a multiplicative model for joint effect between dose d and another risk factor f , the ERR for exposure d is independent of f [$ERR(d|f) = ERR(d)$], so the PC for radiation does not depend on the other factor. With an additive model, $ERR(d|f) = ERR(d)/[1 + ERR(f)]$, so the PC for radiation depends on the level of the other factor f . As generally occurs in epidemiologic studies, lack of power to discriminate between additive

and multiplicative scales for joint effects (i.e., assess interaction between d and f) means that calculation of PC can be subject to large uncertainty when there is exposure to multiple risk factors [26], but ignoring other factors produces estimates of PC that are not specific to an individual case [27].

Major Results

Assessment of risk from ionizing radiation, perhaps the best studied of all *environmental hazards* (see **Environmental Hazard**), is a classical example of the *low-dose extrapolation* (see **Low-Dose Extrapolation**) problem common to environmental and toxicological risk assessment in general: the primary interest is in risk of exposure to low doses (in particular medical uses and occupations involving radiation exposure), but studies at low dose do not provide adequate statistical power, so results from exposures to high doses must be extrapolated to low doses with appropriate adjustment for other differences such as dose rate. The Japanese A-bomb survivor studies are often considered high-dose studies, but large portions of the survivor population actually received relatively low doses compared to the maximum survivable whole-body dose (the $LD_{50/60}$ – the dose producing 50% mortality in 60 days – is about 3.5–4.0 Gy air or surface dose [28]). The mean dose among A-bomb survivors with dose estimates is about 0.2 Gy but the median dose is only about 0.003 Gy. Current radiation protection standards recommended by the International Commission on Radiological Protection (ICRP) allow a maximum of 0.02 Gy yr^{-1} averaged over 5 years for occupationally exposed persons and a maximum of 0.001 Gy yr^{-1} for the public [29], and there is great interest in the range of total dose below about 0.1 Gy. The problem in the A-bomb cohort is that studies including only doses in this range or below do not have enough fitted excess cases to provide meaningful estimates, and they produce results that depend on the choice of unexposed control group, whereas studies including higher doses depend on both the choice of unexposed control group and the assumed shape of the dose–response [6]. The transport of risk to populations affected by radiation protection guidelines also involves problems of extrapolating to exposures that may be chronic, may involve high-LET radiations, and may involve populations with very different baseline rates of the health effect under consideration.

The major findings so far of the A-bomb survivor studies can be summarized as follows: an early elevated risk of leukemia mortality (albeit on top of a small background or spontaneous rate) that has subsided – though not disappeared altogether – to this day [30], an elevated risk of cancer incidence and mortality later in life that has persisted with slight decline over many decades following the exposure and is inversely related to the age at exposure [10], and an increased risk of mortality from certain diseases other than cancer – coronary heart disease, stroke, respiratory diseases (primarily pneumonia), and digestive system diseases (including liver cirrhosis) [10]. The cancer dose–response is usually assumed to be linear without a threshold (the so-called linear no-threshold or LNT assumption); this has recently received new consensus scientific support in BEIR VII and has been corroborated in another recent study [31]. It is important to note, however, that low-dose studies may not definitively prove linearity [32], and analyses of the A-bomb survivor cancer incidence dose–response using nonparametric methods suggest the possibility of nonlinearity at low doses [6].

Developmental (teratogenic) effects at higher doses have been noted in A-bomb survivors exposed *in utero* [33]; increased cancer rates have also been noted, but data are unavailable for the first 5 years after the bombings and the stronger evidence for solid cancer as well as leukemia after prenatal exposure is provided by studies of medical irradiation [3].

Genetic effects of radiation have been notorious since the *Drosophila* experiments of Müller in the 1920s and have been well characterized in animals, but the existence of discernible effects in exposed human populations has remained controversial (e.g., the issue of purported increases in rates of childhood cancer (particularly leukemia) in regions of the U.K. proximal to nuclear installations [34]). Estimates of genetic doubling dose have been promulgated based mostly on animal studies, but there is as yet no firm evidence in humans. BEIR VII uses a combination of animal-based doubling dose estimates and human data on spontaneous mutation rates to calculate estimates of radiation-induced excess rates of genetic diseases in humans, which are compatible with the lack of statistically significant excess rates observed to date in, e.g., the children of A-bomb survivors [3]. Studies in children of A-bomb survivors have progressed over several decades from studies of early,

Table 1 Sample of radiation risk estimates (Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation)^(a)

Outcome	Excess relative risk per gray (95% CI ^(b))	Study population
Thyroid cancer	0.53 (0.14, 2.0) (males) 1.05 (0.28, 3.9) (females) (risk primarily in persons exposed under the age of 20)	Pooled analyses
Female breast cancer	0.51 (0.28, 0.83) highest in women diagnosed under the age of 35	Pooled analysis
All solid cancer mortality except thyroid and nonmelanoma skin	0.35 (0.22, 0.55) sex-averaged, for exposure at ages between 30 and 45; declining with attained age	A-bomb survivors
Leukemia mortality	1.1 (0.1, 2.6) (males) 1.2 (0.1, 2.9) (females) (linear term, with significantly positive quadratic term), declining with age at exposure and time since exposure	A-bomb survivors
Mortality from causes other than cancer/neoplasia	0.14 (0.09, 0.19) (insufficient data to get good estimates of effect modification)	A-bomb survivors

^(a) Reproduced from [3]. © The National Academies Press, 2006

^(b) CI: confidence interval

easily observable effects such as perinatal deaths and malformations to studies of complex, multifactorial diseases expressed in middle age and older age; the latter are ongoing and results are expected to be forthcoming soon.

Studies based on doses received in medical practice, which almost always involve much larger dose to a limited part of the body than the remainder, have contributed information mostly from therapy, particularly on second primary cancers following radiation treatment of a first cancer. Diagnostic procedures, with a few notable exceptions, have resulted in cumulative doses too low to be informative [3]. Studies of occupationally exposed cohorts, particularly nuclear industry workers, have contributed important information, including a large new multinational study [31]. Owing to the lack of power of many studies, a number of analyses have been conducted on pooled data (e.g., [35, 36]). A sample of radiation risk estimates is shown in Table 1, which gives an indication of the difficulty in presenting simple summaries of risk due to complex modification by age, time, and gender.

References

- [1] Cember, H. (1996). *Introduction to Health Physics*, 3rd Edition, McGraw-Hill, New York.
- [2] Simon, S.L., Kleinerman, R.A., Ron, E. & Bouville, A. (2006). Uses of dosimetry in radiation epidemiology, *Radiation Research* **166**, 125–318.
- [3] Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation (2006). *Health Risks from Exposure to Low Levels of Ionizing Radiation*. BEIR VII Phase 2, The National Academies Press, Washington DC.
- [4] United Nations Scientific Committee on the Effects of Ionizing Radiation (2000). *Sources and Effects of Ionizing Radiation*, United Nations, New York.
- [5] Greenland, S., Robins, J.M. & Pearl, J. (1999). Confounding and collapsibility in causal inference, *Statistical Science* **14**, 29–49.
- [6] Pierce, D.A. & Preston, D.L. (2000). Radiation-related cancer risks at low doses among atomic bomb survivors, *Radiation Research* **154**, 178–186.
- [7] Cologne, J.B. & Preston, D.L. (2001). Impact of comparison group on cohort dose–response regression: an example using risk estimation in atomic-bomb survivors, *Health Physics* **80**, 491–496.
- [8] Thomas, D.C. (1981). General relative-risk models for survival time and matched case–control analysis, *Biometrics* **37**, 673–686.
- [9] Lagarde, F. (2006). Understanding estimation of time and age effect-modification of radiation-induced cancer risks among atomic bomb survivors, *Health Physics* **91**, 608–618.
- [10] Preston, D.L., Shimizu, Y., Pierce, D.A., Suyama, A. & Mabuchi, K. (2003). Studies of mortality of atomic bomb survivors. Report 13: solid cancer and noncancer disease mortality 1950–1997, *Radiation Research* **160**, 381–407.
- [11] Cologne, J.B., Izumi, S., Shimizu, Y. & Preston, D.L. (2002). Effect of comparison group on inference about effect modification by demographic factors in cohort risk regression, *Japanese Journal of Biometrics* **23**, 49–66.
- [12] Frome, E.L. (1983). The analysis of rates using Poisson regression models, *Biometrics* **39**, 665–674.
- [13] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition, Chapman & Hall, London.
- [14] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research. Volume II – The Design and Analysis of Cohort Studies*, International Agency for Research on Cancer, Lyon.

- [15] Preston, D.L., Lubin, J.H., Pierce, D.A. & McConney, M.E. (1993). *Epicure User's Guide*, Hirosoft International, Seattle.
- [16] Pierce, D.A., Stram, D.O., Vaeth, M. & Schafer, D.W. (1992). The errors-in-variables problem: considerations provided by radiation dose-response analyses of the A-bomb survivor data, *Journal of American Statistical Association* **87**, 351–359.
- [17] Pierce, D.A. & Kellerer, A.M. (2004). Adjusting for covariate errors with nonparametric assessment of the true covariate distribution, *Biometrika* **91**, 863–876.
- [18] Thomas, D., Stram, D. & Dwyer, J. (1993). Exposure measurement error: Influence on exposure-disease relationships and methods of correction, *Annual Review of Public Health* **14**, 69–93.
- [19] Pierce, D.A. & Vaeth, M. (2003). Age-time patterns of cancer to be anticipated from exposure to general mutagens, *Biostatistics* **4**, 231–248.
- [20] Luebeck, G.E., Heidenreich, W.F., Hazelton, W.D., Paretzke, H.G. & Moolgavkar, S.H. (1999). Biologically based analysis of the data for the Colorado uranium miners cohort: age, dose, and dose-rate effects, *Radiation Research* **152**, 339–351.
- [21] Thomas, D., Darby, S., Fagnani, F., Hubert, P., Vaeth, M. & Weiss, K. (1992). Definition and estimation of lifetime detriment from radiation exposures: principles and methods, *Health Physics* **63**, 259–272.
- [22] Land, C.A., Gilbert, E. & Smith, J.M. (2003). *Report of the NCI-CDC Working Group to Revise the 1985 NIH Radioepidemiological Tables*, NIH Report No. 03–5387, U.S. Department of Health and Human Services, Washington, DC.
- [23] Vaeth, M. & Pierce, D.A. (1990). Calculating excess lifetime risk in relative risk models, *Environmental Health Perspectives* **87**, 83–94.
- [24] Robins, J.M. & Greenland, S. (1989). Estimability and estimation of excess and etiologic fractions, *Statistics in Medicine* **8**, 845–859.
- [25] Pierce, D.A., Sharp, G.B. & Mabuchi, K. (2003). Joint effects of radiation and smoking on lung cancer risk among atomic bomb survivors, *Radiation Research* **159**, 511–520.
- [26] Cologne, J.B., Pawel, D.J., Sharp, G.B. & Fujiwara, S. (2004). Uncertainty in estimating probability of causation in a cross-sectional study: joint effects of radiation and hepatitis-C virus on chronic liver disease, *Journal of Radiological Protection* **24**, 131–145.
- [27] Beyea, J. & Greenland, S. (1999). The importance of specifying the underlying biologic model in estimating the probability of causation, *Health Physics* **76**, 269–274.
- [28] Mettler, F.A. & Upton, A.C. (1995). *Medical Effects of Ionizing Radiation*, 2nd Edition, W.B. Saunders, Philadelphia.
- [29] International Commission on Radiological Protection (ICRP) (1991). Recommendations of the International Commission on Radiological Protection, *Annals of the ICRP*, ICRP Pub. No. 60, Pergamon Press, New York.
- [30] Pierce, D.A., Shimizu, Y., Preston, D.L., Vaeth, M. & Mabuchi, K. (1996). Studies on the mortality of atomic bomb survivors. Report 12, part I. Cancer: 1950–1990, *Radiation Research* **146**, 1–27.
- [31] Cardis, E., Vrijheid, M., Blettner, M., Gilbert, E., Hakama, M., Hill, C., Howe, G., Kaldor, J., Muirhead, C.R., Schubauer-Berigan, M., Yoshimura, T., Bermann, F., Cowper, G., Fix, J., Hacker, C., Heinmiller, B., Marshal, I.M., Thierry-Chef, I., Utterback, D., Ahn, Y.-O., Amoros, E., Ashmore, P., Auvinen, A., Bae, J.-M., Solano, J.B., Biau, A., Combalot, E., Deboodt, P., Diez Sacristan, A., Eklof, M., Engels, H., Engholm, G., Gulis, G., Habib, R., Holan, K., Hyvonen, H., Kerekes, A., Kurtinaitis, J., Malke, H., Martuzzi, M., Mastauskas, A., Monnet, A., Moser, M., Pearce, M.S., Richardson, D.B., Rodriguez-Artalejo, F., Rogel, A., Tardy, H., Telle-Lamberton, M., Turai, I., Usel, M. & Veress, K. (2005). Risk of cancer after low doses of ionising radiation: retrospective cohort study in 15 countries, *British Medical Journal* **331**, 77–83.
- [32] Lagarde, F., Cullings, H.M., Shimizu, Y. & Cologne, J.B. (2005). Tiny excess relative risks hard to pin down. Rapid response, *British Medical Journal*, <http://www.bmj.com/cgi/eletters/331/7508/77> (accessed 9 Aug 2005).
- [33] Delongchamp, R.R., Mabuchi, K., Yoshimoto, Y. & Preston, D.L. (1997). Cancer mortality among atomic bomb survivors exposed in utero or as young children, October 1950 to May 1992, *Radiation Research* **147**, 385–395.
- [34] Gardner, M.J. (1989). Review of reported increases of childhood cancer rates in the vicinity of nuclear installations in the UK, *Journal of the Royal Statistical Society Series A* **152**, 307–325.
- [35] Ron, E., Lubin, J.H., Shore, R.E., Mabuchi, K., Modan, B., Pottern, L., Schneider, A.B., Tucker, M.A. & Boice, J.D. (1995). Thyroid cancer after exposure to external radiation: a pooled analysis of seven studies, *Radiation Research* **141**, 259–277.
- [36] Preston, D.L., Mattsson, A., Holmberg, E., Shore, R., Hildreth, N.G. & Boice, J.D. (2002). Radiation effects on breast cancer risk: a pooled analysis of eight cohorts, *Radiation Research* **158**, 220–235.

Related Articles

Environmental Carcinogenesis Risk

Extremely Low Frequency Electric and Magnetic Fields

Statistics for Environmental Mutagenesis

Statistics for Environmental Teratogenesis

HARRY M. CULLINGS AND JOHN B. COLOGNE

Risk in Credit Granting and Lending Decisions: Credit Scoring

When deciding whether to lend money to a friend, relative, or colleague at work, one is not usually concerned about making a precise estimate of the probability that the loan will be repaid. This is because you know the individuals and something about them. Therefore, the decision can be perceived as a simple one that is based on your prior knowledge of the person's character, intent, and ability to repay the debt. You use what you know about their personality and past behavior to make what may be termed a *qualitative decision* where you weigh the social damage that will be done to the relationship if you do not lend the money against the financial loss that you will experience should the loan not be repaid.

Now let us consider the case of a credit granting organization, which must decide whether or not to grant credit to an individual who has applied for a loan. The general principle is the same. The credit provider must infer something about the individual's ability and intent to repay the debt. In the credit scoring literature and jargon, a customer that meets his/her credit obligations is referred to as a *good* customer; otherwise he/she is a *bad* customer. Lenders must also consider the likely return that they will make from a good customer who repays the loan against the expected loss from a bad customer who defaults. However, in a modern setting, most credit granting institutions deal with hundreds of thousands or even millions of individuals seeking credit every year. Therefore, it is not practical or cost-effective for an organization to develop the same type of in-depth relationship with their customers as can exist between two individuals who know each other well. Consequently, credit granting institutions have to rely heavily on the use of predictive models of consumer behavior to produce quantitative estimates of the likelihood of individuals defaulting on any credit that they are granted.

The techniques and methodologies used to develop and apply models of consumer repayment behavior are generally referred to as *credit scoring*. A credit scoring model can, therefore, be defined as any

quantitative predictive model that generates an estimate of the likelihood that a credit obligation will be met. Within the credit industry, the term *credit score* is usually used to describe the estimate generated by a credit scoring model for a given individual. A credit score can, therefore, be interpreted as the probability that a given individual will repay the credit applied for, should request for credit be granted. As we shall see in later sections, the most popular mathematical and statistical techniques used to construct credit scoring models, such as logistic regression, are widely used within many different areas of applied statistics and data mining. However, what differentiates credit scoring from other applications of such techniques are the objectives of the financial institutions that employ it, the nature of the data employed, and the social, ethical, and legal issues that constrain its use within the decision making process.

In this paper, our objective is to present an overview of how credit scoring is applied within the context of credit risk assessment and credit management. We focus our attention on retail sector lending for consumers: credit cards, mortgages, and instalment lending. We will particularly look at how lenders use various types of information to build predictive models of the risk of default. The remainder of this paper is organized as follows. In the next section, we explain how lenders approach the problem of modeling the risk of defaulting. We provide the context of the problem of credit scoring and explore the main approaches to developing credit scoring models, which are used to estimate the risk of default. In the section titled "Research Issues" of the paper, we present some interesting research issues that the authors feel will continue to burgeon. In the last section of the paper we offer our concluding remarks.

Consumer Credit Modeling

Introduction and the Context

Credit scoring is usually divided into application scoring and behavioral scoring. Application scoring is concerned with the decision of whether to grant credit to an individual. Behavioral scoring is concerned with the "management" of the consumer, who has been granted credit. Typical actions that the lender will take in behavioral scoring include increasing or decreasing the credit limit, or to make

better offers to consumers to reduce the probability of attrition. In this paper, we will concentrate on the estimation of the risk of default in the context of application scoring, i.e., the likelihood of an individual displaying “good” or “bad” repayment behavior. In practice, every lender has its own definition of good/bad/indeterminate customers owing to differing heads, operational, and accounting procedures, as well as the subjective views of individual credit managers as to what constitutes a good or bad customer. However, as Lewis [1] describes, most definitions of good/bad/indeterminate customers are along the following lines:

- *Good*: A customer who has maintained repayments in line with the terms of the credit agreement.
- *Bad*: A customer with a history of serious arrears, where serious is usually defined as *3 months or more behind with the repayments*. Three months is usually chosen because customers who are this far in arrears nearly always go further into arrears, resulting in a loss for the lender. Even if the customer subsequently pays the arrears, the additional cost to the lender for dealing with such an account will often mean that it will not be a profitable relationship in the long run.
- *Indeterminate*: All other cases.

The indeterminate category defines cases where good/bad performance is, for whatever reason, ambiguous for – example, a personal loan that did not go to term because the borrower repaid the loan early, or where the customer has missed the odd payment, but has not been seriously in arrears. Indeterminate categories are not usually included within the model development process [2] and therefore the problem is normally reduced to one of binary classification.

Before the development and use of credit scoring, the majority of lending decisions were made judgmentally by trained underwriters. This would typically involve a face-to-face meeting where the lender would have to judge the creditworthiness of the applicant. Decisions were made on the basis of what is known as the 5Cs of lending decisions: Capital, Collateral, Capacity, Character, and Conditions [3, 4]. First is capital, which is the amount to be borrowed, second is collateral, which are any assets which can be used as security for the debt, third is capacity which is the ability of the borrower to maintain their repayments, fourth is character, which is

the applicant’s intent to repay the debt and finally we have the conditions such as the current economic climate and forecasts. There are three main problems with the judgmental approach. First, it is time consuming and therefore expensive. Second, it is subjective, i.e., two different underwriters may make different decisions about the same loan application. Third, it is subject to prejudice; individuals from some sections of the society may be declined credit for reasons other than their likelihood of repaying the loan.

In order to make credit scoring based decisions, lenders will need analytical predictive models (*see Credit Risk Models*) to estimate the risk of default. There is still a need to collect information to infer the risk of default. In the strict empirical philosophy of building good analytical models, if any variable is likely to provide some useful information on the creditworthiness of applicants it should be used to assess the risk. However, it is possible that there are some constraints on the type of information that can be used for credit scoring. In the United Kingdom and the United States, for example, lenders are prohibited from using gender or race in their analytical models to credit score consumers. Other countries may impose other country-specific rules and regulations on the use of information for credit scoring. If a consumer applies for credit, the applicant is usually asked to complete an application form providing details about themselves that are believed to be predictive of the risk default – for example, their age, residential status, marital status, employment status, nationality, etc. The application form data is not the only piece of information that is used to infer the risk of default. The application form information does not tell us anything about the applicant’s previous credit record; i.e., whether he or she was a “good or bad” customer with other credit providers in the past. Lenders will therefore obtain this type of information from a credit reference agency, which acts as a data exchange database resource for lenders allowing them to share data about the behavior of credit applicants.

Analytical Models for Credit Scoring

In the discussion above, we have briefly discussed how lenders use different types and levels of information to estimate the risk of default on a nonjudgmental basis. Lenders, like all empirical modelers, are always looking for better models to reduce

the cost of misclassifying bad customers as good and *vice versa*. It has to be noted that what may seem to be a very small improvement in the performance of the models in statistical terms may effectively represent a huge improvement in revenue generated from these decisions. It is, therefore, not surprising that literally dozens of modeling approaches have been studied in order to find ones that yield the best efficient coefficients for modeling credit risk. In the discussion that follows, we review some of the models that are commonly described in the literature and/or used by lenders to discriminate between good and bad applicants for credit. It is impossible for us to provide a full review of all the modeling approaches. The pedagogical objective of the discussion here is to give an insight into how the most popular models are used in practice to estimate the risk of default. The models described below are logistic regression model (*see Logistic Regression*), decision trees (*see Decision Trees*), and neural networks. Readers and those who wish to pursue research in this area will find the following references useful for a more comprehensive review of the models used in credit scoring: Thomas [3] and Thomas *et al.* [5, 6].

In credit scoring, the risk of default, PD (the probability of default thereafter) is assumed to be a function of an applicant's characteristics (e.g., age, employment status, and residential status), which we define by the vector $\mathbf{X}$, $\mathbf{X} = \{X_1, X_2, X_3, \dots, X_n\}$, and unobservable or unaccountable factors, which we denote by ε :

$$PD = f(\mathbf{X}, \varepsilon) \quad (1)$$

It is to be noted that one could also model the risk of nondefaulting (1-PD). The most popular model used in credit scoring, both in academia and by lenders, is the logistic regression model [6]. The logistic model and the probability of default is formulated as follows:

$$PD = \frac{\exp(X\beta_p)}{\exp(X\beta_{NPD}) + \exp(X\beta_p)} \quad (2)$$

$\exp(X\beta_p)$ is the exponential combined effect of defaulting where $\mathbf{X}$ is the vector of characteristics as defined previously and β_p is the parameter vector defining each of the parameters for each variable used in the model. $\exp(X\beta_{NPD})$ is the exponential combined effect of nondefaulting where $\mathbf{X}$ is the vector of characteristics and β_{NPD} is the parameter vector. The

formulation of the model ensures, unlike the linear regression model, that the probability of default lies between 0 and 1. One can also transform the above logistic formulation such that the Log of the odds ratio is now a linear function of the characteristics and this is represented in equation (3). The coefficient for any variable indicates whether the log odds ratio of defaulting increases or decreases for a unit increase in that variable. In order to predict whether a customer is likely to default in the future, the applicant's characteristics information is "fed into" the estimated logistic model and resulting predicted probabilities of defaulting and not defaulting indicate the risk that will be taken if the loan or credit is granted. On the basis of the predictions of the probability model alone, the loan or credit is likely to be granted if the probability of not defaulting is less than the probability of defaulting, i.e., a good risk to take on.

$$\text{Log} \left(\frac{PD}{NPD} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (3)$$

In decision trees (also commonly referred to as *classification trees*) recursive partitioning algorithms are applied where categorical splits are made recursively within the population to produce smaller and smaller homogeneous groups. Each subsidiary group (also referred to as the *child node*) contains a "purer" set of observations than its parent, containing an increasing proportion of one of the two classes of interest: that is "good" or "bad". The segmentation process is halted only when a given set of stopping criteria are met, such as the node is $>x\%$ pure (discrimination between good and bad risks) or the number of observations in the node is less than a specified minimum. The segmentation process is usually captured in the form of a *classification tree* as shown in Figure 1.

The rules for creating each branch of the tree are determined by examining the domain of all further potential splits and choosing the split that maximizes the discrimination between the two resulting groups. To illustrate, the results from decision tree above indicate that applicants who are 35 or older are four times less likely to default than others (ratio of good/bad in each node). Other rules on "what type" of applicants to grant credit, based on their characteristics, can be derived from classification trees and such rules will be employed by the decision

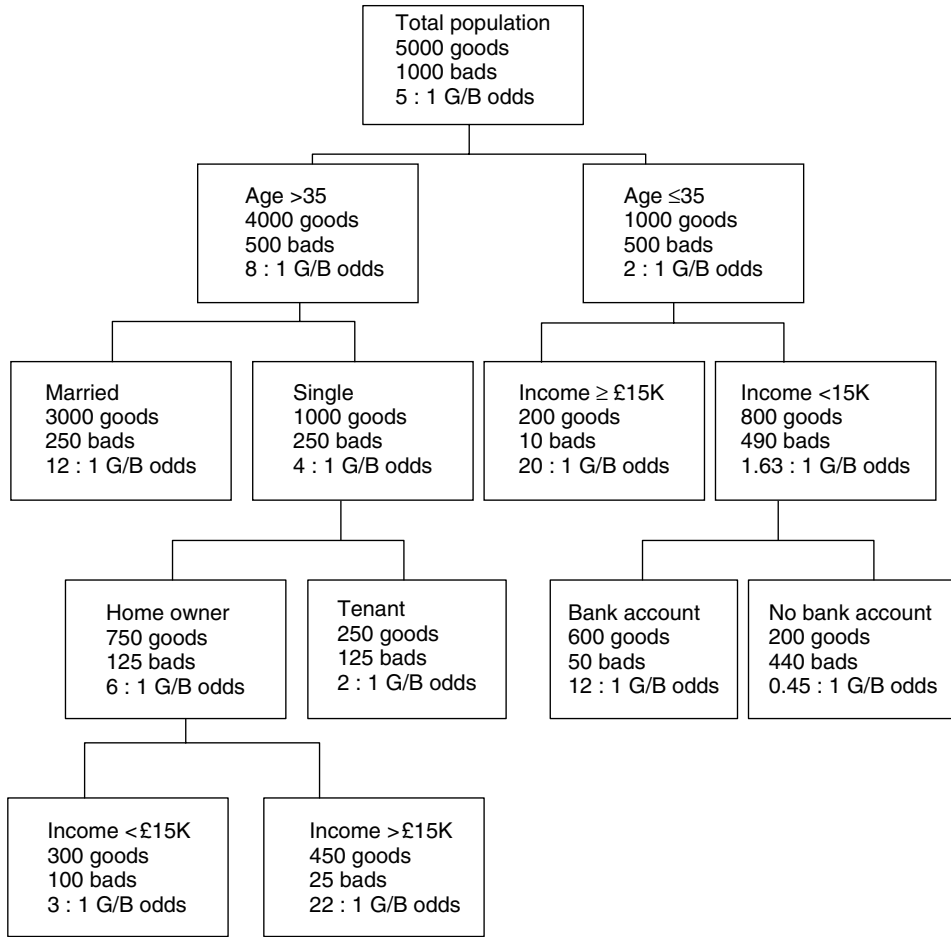


Figure 1 Example of a classification tree

maker to make a decision as to whether to grant or refuse credit. There are a number of possible algorithms available to find the different segments or nodes of a tree and some of them can be found in [5].

Neural network models are based on artificial intelligence methods for structuring and solving problems of classification and prediction and they are now being widely applied for data mining related problems such as customer relationship management. The process of modeling is supposed to reflect the activity process of the brain in the same manner that the “human brain” solves problems. The output results from neural network models are not as intuitive explaining why such type of modeling is less popular with lenders. While in regression terminology we

refer to variables and coefficients, in neural network models these are referred to as *inputs and weights*, respectively. The classification outcome from a neural network model is referred to as the *output*. Neural network models are less restrictive compared to the statistical underpinnings of the regression type modeling for classification. The simplest form of neural network is the single layer model illustrated in Figure 2 and the more complex but more flexible model is the multilayer formulation illustrated in Figure 3. In both figures, F stands for the transformation functions, which the analyst uses to transform the combined effect of the inputs. The multilayer formulation allows for the various interactions of independent variables or inputs described by the vector $\mathbf{X} = \{X_1, X_2, X_3, \dots, X_n\}$.

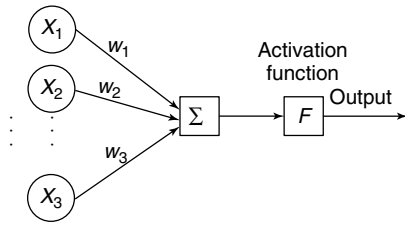


Figure 2 Single layer neural network [Reproduced from [5]. © Society for Industrial and Applied Mathematics, 2002.]

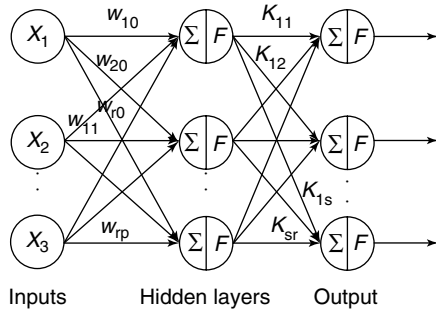


Figure 3 Multilayer neural network (Adapted from Thomas *et al.* [5].)

Evaluating Model Performance

The primary use of credit scoring models is as a predictive tool, to estimate the likelihood of individuals being good or bad prospects when they apply for credit. It is also true that users of credit scoring tend to deal with large populations and the point estimates for individual cases may not be of importance. Therefore, users are not overly interested in common measures of model fit, such as the R^2 coefficient for discriminant analysis (see **Credit Scoring via Altman Z-Score**) or the likelihood ratio in logistic regression; nor are lenders particularly interested in confidence intervals for the parameter estimates or individual point estimates. Instead, what is of most importance are the misclassification properties of the scoring model at specific points in the distribution of the model scores or the discriminatory power of the model, as measured across the range of model scores. It is generally recommended that the analyst use only a proportion of the data (the training sample) to estimate the model, and use the remaining proportion (the holdout sample) to assess model performance.

It is important to note here that the holdout sample is not used in the estimation of the model, and this, in a way, replicates the credit scoring process where one would feed information on new applicants into the model to assess whether they should be granted credit. The following measures of model performance tend to be of interest:

- The misclassification rate/cost within the highest scoring $x\%$ of the population, that is, the number/cost of bad scoring $\geq c$, and number/cost of good scoring $< c$ where c is the cutoff score at or above which $x\%$ of the population score.
- The misclassification rate/cost for the cutoff defined by the good/bad odds at score. For example, the number of bad risks accepted for a cutoff score where the good/bad odds are in the ratio of 10 : 1. This measure can only be calculated precisely if the output of the model can be directly interpreted as a probability of class membership.
- The overall discrimination of the model, that is, some measure of the separation between the groups of interest that the resulting scorecard provides. In this case, one can use a cross tabulation of the actual and predicted classifications of good and bad risks of the model. The error rate of the model will be defined by the number of misclassified cases (bad as good and *vice versa*) divided by the total number of cases.

A comprehensive discussion of the assessment methods that are used by credit modelers can be found in [2, 5].

A final point to note here is that predictive ability is not the only criteria by which credit scoring models are assessed. There is often a requirement for the model to be easily explicable and for the coefficients to display sensible patterns in line with the expert's opinion and univariate analysis. This is primarily owing to legislation found in many countries that requires lenders to explain why a model allocates certain scores to certain individuals. This, therefore, is perhaps most easily achieved where the model structure that is employed is in the form of either decision trees – comprising simple binary segmentation – or models that are represented by linear combination of parameter coefficients, such as with linear and logistic regression. Therefore, it is possible that when a practitioner is faced with two competing models and if the performance of each

model is very similar, the model with the simplest parameter structure will usually be chosen.

Research Issues

There are many interesting research issues and areas in credit scoring far more than can be discussed in such a short paper. Four main areas of research are summarized below. We believe that such areas of research will continue to flourish and may also represent further avenues of research.

The first is research into better classification methodologies. This primarily involves evaluating the performance classification and regression techniques that have not been previously applied to credit scoring. However, there is also growing interest into how better models can be developed by including new sources of data and choosing the most appropriate way to format the data before it is fed into the chosen model [7]. Classifier combination (ensemble modeling) has also demonstrated potential, where two or more credit scoring models are developed using different techniques and/or different data samples. The output of each model is then combined in some way to yield a single decision that may be more accurate than that produced by individual models. For a fuller discussion of classifier combination strategies refer to [8–10] *inter alia*.

A second avenue of research is into using information on consumer behavior. Information on consumer behavior such as spending patterns and lifestyle can be used into the modeling of credit risk and determining the amount and price of credit [5], for example, by overlaying the decision to lend on the basis of risk of default using measures of affordability, mailing response, attrition to a rival product, fraud, etc.

A third avenue of research is into reformulating the objectives of credit scoring. Identifying good and bad credit risks is, after all, a crude approximation to the “true” goal of segmenting customers on the basis of their contribution to profit for the lending organization. Therefore, there is increasing interest in ways in which customer profitability can be modeled. See [5, 11] for discussion on the issues surrounding profitability modeling and lending decisions.

A fourth growing area of research is on the application of predictive approaches used in consumer

credit scoring to evaluate the risk of default of companies [12, 13]. Banks and other financial institutions that grant loans or offer other credit facilities use information such as financial reports or other operations related information on organizations to assess the risk of default of companies. Moody’s rating (<http://www.moody.com>) is the best known example of the use of credit scoring predictive techniques to assess the risk of default and financial “health” of companies (*see Credit Scoring via Altman Z-Score; Decision Analysis; Utility Function*).

Conclusions

The modeling of risk will become increasingly important for lenders with the coming into effect of the Basel II accord as laid out by the Bank of International Settlements. Under this accord, lenders will need to demonstrate that the amount of capital they have is in line with what is commonly referred to as their *risk exposure*, that is, the overall portfolio risk of the various loans and credit they have granted. In this paper, we have introduced the readers to the basic principles of credit scoring. Credit scoring is the application of predictive approaches in support of lending decisions made on the estimated risk of loan default. We have discussed both the modeling approaches to credit scoring and the environmental conditions within which credit scoring is applied by practitioners. We hope that we have demonstrated the peculiarities of credit scoring within the wider disciplines of data mining, statistical modeling, and risk analysis, respectively.

References

- [1] Lewis, E.M. (1992). *An Introduction to Credit Scoring*, Athena Press.
- [2] Hand, D.J. & Henley, W.E. (1997). Statistical classification methods in consumer credit scoring: a review, *Journal of the Royal Statistical Society, Series A: Statistics in Society* **160**, 523–541.
- [3] Thomas, L.C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers, *International Journal of Forecasting* **16**(2), 149–172.
- [4] Finlay, S. (2005). *Consumer Credit Fundamentals*, Palgrave, Macmillan.
- [5] Thomas, L.C., Edelman, D.B. & Crook, J.N. (2002). *Credit Scoring and Its Applications*, SIAM, Philadelphia.

- [6] Thomas, L.C., Oliver, R.W. & Hand, D.J. (2005). A survey of the issues in consumer credit modelling research, *Journal of the Operational Research Society* **56**, 1006–1015.
- [7] Crone, S.F., Lessmann, S. & Stahlbock, R. (2006). The impact of pre-processing on data mining: an evaluation of classifier sensitivity in direct marketing, *European Journal of Operational Research* **173**(3), 781–800.
- [8] Kittler, J., Hatef, M., Duin, R.P. & Matas, J. (1998). On combining classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 226–239.
- [9] Zhu, H., Beling, P. & Overstreet, G. (2002). A Bayesian framework for the combination of classifier outputs, *Journal of the Operational Research Society* **53**(7), 719–727.
- [10] West, D., Dellana, S. & Qian, J. (2005). Neural network in credit scoring models, *Computers and Operations Research* **27**, 1131–1152.
- [11] Finlay, S.M. (2006). *Towards Profitability: A Utility Approach to the Credit Scoring Problem*, Lancaster University Management School Working Paper, 2006/021.
- [12] Back, P. (2005). Explaining financial difficulties based on previous payment behaviour, management background variables and financial ratios, *European Accounting Review* **14**(4), 839–868.
- [13] Minussi, J., Worthington, D. & Soopramanien, D. (2006). *Statistical Modelling Defaulting Companies for the Basel Accord: A Brazilian Case Study*, Lancaster University Management School Working Papers 2006/028.

Related Articles

Credit Scoring via Altman Z-Score

Decision Analysis

Utility Function

DIDIER G.R. SOOPRAMANIEN AND STEVE
FINLAY

Risk Management of Construction Defects

Disputes concerning alleged defects have been a common element in construction projects for many years. Attempts to manage the risks associated with such defects are properly focused on preventing the occurrence of defects and minimizing their costs when they do occur. Advice on prevention includes the use of quality materials and workmanship, as well as attending to proper maintenance after construction. The impact of a defect, if and when it occurs, may be minimized by ensuring that both developer and owner understand their respective rights and responsibilities, maintain open lines of communication, and take good faith actions to build trust with the other party.

In resolving construction disputes, including those involving alleged defects, the parties have various resolution options: negotiation, use of a project neutral, mediation, dispute review boards (DRBs), administrative board, mini-trial, binding arbitration, or litigation. As illustrated in Figure 1, the cost of resolving a dispute, the degree of control over the outcome retained by the parties involved, and the level of hostility between them will vary depending on the option used. Anecdotal evidence suggests that dispute resolution techniques through which the parties in disagreement retain control of the dispute incur fewer costs during the resolution process and keep hostilities to a minimum [1]. However, several factors, including the amount of damages being sought, influence the likelihood that a relatively inexpensive option (e.g., negotiation) will be used instead of a more expensive option (e.g., litigation).

Risk management has often been defined as a process involving three steps: hazard identification, risk assessment, and risk mitigation/communication. In the context of construction defects, as in other applications, the value of quantitative risk assessment – and, consequently, the likelihood that such an assessment will be undertaken – typically increases with the magnitude of the dispute. An assessment of construction defect risk involves estimating both the number of defects and the unit costs of addressing them. Statistical sampling and estimation are well suited to play critical roles in conducting such an assessment, particularly in estimating the prevalence of defects.

Estimation of Defect Prevalence

The use of statistical methods of inference in assessing risks associated with construction defects requires defining a population of interest and collecting information about that population by various means, including inspection and testing on a sample basis. In disputes involving a single residence or commercial building, the population of interest could refer to all work performed at the site by one or more specific contractors or subcontractors. For condominium projects or large developments of single-family homes, the population would typically comprise all housing units involved in the dispute.

If the existence of an alleged defect is in question, then an initial investigation may be undertaken for the purpose of discovery and confirmation of the defect. Often, visual inspections will be made, and some destructive testing may be performed. Attention will focus on the structures and locations on each structure showing the most evidence of distress compatible with the alleged defect. Inspection and test data collected in this way for the purpose of demonstrating the existence of an alleged defect do not provide a valid basis for extrapolating to unexamined units of the population. Instead, valid extrapolation requires scientific sampling designed and implemented according to statistical principles and best practices.

The sampling approach that is used to collect information about the population may vary depending on whether the alleged defect concerns the architectural design, building materials, methods of construction, or workmanship associated with the project in dispute. For disputes involving design, materials, or methods, the chief purpose of sampling may be to verify that the structure(s) has been built as designed, using the specified construction materials and methods. In cases where the alleged defect is expected to be present in all units, a relatively small sample can provide the confirmation. From a random sample of five units, all of which are found to possess an allegedly defective characteristic, one can conclude with greater than 95% statistical confidence that a majority of units in the population possesses that characteristic. Such a statement may be regarded as addressing the legal standard of a preponderance of evidence in that, at the prescribed level of confidence,

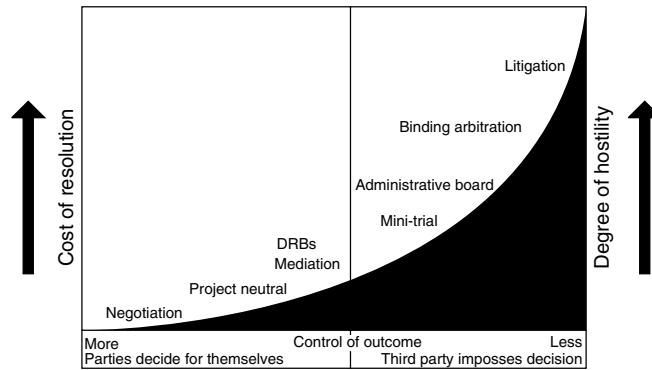


Figure 1 Characteristics of alternative methods for dispute resolution

a randomly selected unit more likely than not would possess the allegedly defective characteristic.

For disputes involving workmanship, the occurrence of a defect may be regarded and modeled as essentially a random event. Statistical sampling and estimation are then employed to address the task of estimating the prevalence of the defect in the population. In this setting, a defensible sampling plan must satisfy three key criteria: (a) suitability of the sampling design, (b) adequacy of the sample size, and (c) accounting for potential selection biases.

Sampling Designs

Any sample drawn for the purpose of estimating prevalence or any other characteristic of a population must be selected so that the sampled units may be regarded as being representative of the larger population. For that reason, sampling should always be done using a random mechanism to select units for inspection and/or testing. Strictly speaking, a random sample is one whose every unit of the population has a known probability of being included in the sample. Technical experts sometimes feel they can select a representative sample by exercising their professional judgment in choosing units from the population. Such an approach was discredited in statistical studies more than 70 years ago [2]. A judgment or “purposive” sample is vulnerable to the charge that a conscious or unconscious bias had influenced the selection of units. Also, statistically derived estimates from such a sample are technically invalid, because the theory on which such estimates are based includes an

assumption of random sampling from the population of interest.

The most basic form of random sampling is simple random sampling, in which every unit of the population has an equal chance of being selected for the sample. In the context of construction defects, one common objection to simple random sampling is that units of the population possessing key characteristics might be underrepresented – if not overlooked entirely – in a random probability sample. For example, in a large development of single-family homes, there may be variations in floor plans and in the time phases and parties involved in construction. Suppose that a population can be divided into subpopulations that are expected to be more homogeneous regarding the presence or absence of an alleged defect. Then, stratified random sampling can be implemented by which random samples are drawn from each subpopulation rather than from the population as a whole. Alternatively, if the sample size is sufficiently large, then one can reasonably expect that particular subsets of the population will be represented in the sample in approximate proportion to their numbers.

Sample Size

Determining the sample size – i.e., the number of units in the population to be inspected and/or tested – needed to estimate prevalence most commonly involves specifying the desired level of accuracy as well as the probability of achieving that level. The term *margin of error* is most frequently associated with a confidence level of 95%, but such values as 99, 90, and even 80% are sometimes used

Table 1 Required sample size by population size and margin of absolute error

Margin of error (%)	Number of units in population			
	100	250	500	1000
3	92	203	341	517
5	80	152	218	278
10	50	70	81	88
20	20	22	23	24

instead. An estimated prevalence of 25% with a three-percentage-point margin of error means that, with 95% confidence, the percentage of defective units in the population is between 22 and 28.

A common misconception in the construction industry, as well as in the general public, is that a fixed percentage of units in the population (e.g., 10%) must be sampled to obtain reliable results. Instead, the sample size will depend on the prescribed level of accuracy and the number of units in the population [3]. Table 1 lists the sample sizes required from populations of varying sizes to obtain estimates that achieve varying margins of error (assuming 95% confidence). Note that a higher percentage of units must be sampled from smaller populations to achieve the same margin of error as in larger populations. The absolute number, rather than the percentage, of sampled units determines the accuracy of resulting estimates of prevalence.

Selection Bias and Substitution

Although the number of units to be sampled is typically determined to control the magnitude of sampling error in resulting estimates, attention must also be given to sources of nonsampling error. In assessing the prevalence of defects, particularly in residential construction, sample-based estimates may be especially vulnerable to bias from nonresponding households. Suppose, for example, that determination of whether an alleged defect is present in a housing unit requires not only visual inspection but some degree of destructive testing also. If the owner of the housing unit sees no evidence of damage and believes the unit is unlikely to possess the alleged defect, he or she may be less willing to cooperate and grant access if the unit is selected for inclusion in the sample. Consequently, the data from responding households

may overestimate the prevalence of the defect in the population.

Selection bias can be exacerbated when alternates are substituted for originally selected housing units that are not available for measurement. However, substitution does provide gains in precision and, by restoring the original sample size, yields estimates that possess the prescribed margin of sampling error. There appears to be no clear theoretical or empirical evidence to accept or reject the practice of substitution [4]. If the percentage of nonresponding sampled units is small, the difference between estimates with and without substitution is likely to be negligible.

Regardless of whether substitution is used or not, the potential for nonresponse bias in sampling can be monitored by listing nonresponding units and documenting the reasons for their unavailability. In some cases, it may be possible to estimate and adjust for selection bias. If data on one or more characteristics associated with the occurrence of the alleged defect are available for both nonresponding and responding units, then a formal statistical test can be conducted to determine whether the two groups of units differ significantly with respect to the measured characteristic. If differences exist, then estimates of prevalence for nonresponding units can be obtained on the basis of the empirical link between values of the measured characteristic and occurrence of the alleged defect. Alternatively, prevalence limits based on bounding assumptions may be calculated to assess the sensitivity of estimates to possible sources of bias.

Other Statistical Considerations

In disputes concerning alleged construction defects, the involved parties will typically be interested in establishing whether there is a physical link between an alleged defect and observed damage. Because the data gathered from inspections and testing of sampled units are observational in nature – and not the result of controlled experimentation – proper interpretation of these data requires consideration of whether other so-called “confounding” factors might be present. For example, moisture intrusion might be attributed to improper window installation but could also be caused by poor roof sheathing. Inspections and tests should be designed and executed to generate information that enables evaluation of alternative explanations for any observed association between occurrence of an alleged defect and property damage.

Inspection and test data can thus provide a basis for judging the empirical evidence for damage theories advocated in a construction dispute. For example, consider the allegation that improper soil compaction has caused tilting and cracking of foundation slabs. This allegation can be examined by measuring the strength of association between foundation tilt (as determined from manometer surveys) and variables such as soil density and depth of fill. Standard statistical methods, including regression and analysis of variance, can be applied in drawing conclusions from formal comparisons.

The findings from inferential statistical methods depend on one or more assumptions (which may or may not be explicitly stated) concerning the process by which the data were generated. The validity of these assumptions should be checked using the data and all other relevant and available information. In circumstances when data are sparse, more than one statistical model may provide an acceptable fit. Attention should be given to determining the robustness of conclusions – that is, the extent to which final answers differ when plausible alternative statistical models are used.

Cost of Repair

Completion of a formal quantitative risk assessment requires integrating sample-based estimates of defect prevalence with expert judgments concerning the unit cost of addressing the observed defects. At this stage, particularly in litigious disputes, construction experts retained by the opposing parties may produce substantially different estimates of unit repair costs. These differences can usually be traced to divergent opinions concerning whether a minor repair will suffice or a major replacement is indicated.

The Construction Defect Risk Manager

Possessing information concerning both the prevalence of defects and their estimated unit costs, the risk manager is then equipped to weigh the options and pursue a course of action that is compatible with his or her client's risk/reward profile. Information about the costs of alternative methods of dispute resolution, as well as their likelihoods of success, can also be considered at this point in the evaluation. The risk manager may choose to apply Bayesian methods for subjective probability assessment (*see Bayesian Statistics in Quantitative Risk Assessment*) and *decision analysis* (*see Decision Analysis*) as formal tools to assist in this process.

References

- [1] Gebken, R.J. & Gibson, G.E. (2006). Quantification of costs for dispute resolution procedures in the construction industry, *Journal of Professional Issues in Engineering Education and Practice* **132**, 264–271.
- [2] Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society* **97**, 558–625.
- [3] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition, John Wiley & Sons, New York.
- [4] Vehovar, V. (1999). Field substitution and unit nonresponse, *Journal of Official Statistics* **15**, 335–350.

Related Articles

Managing Infrastructure Reliability, Safety, and Security

Product Risk Management: Testing and Warranties

DUANE L. STEFFEY

Risk Measures and Economic Capital for (Re)insurers

Risk Measurement and Capital

A *risk measure* is a function that assigns real numbers to random variables representing uncertain payoffs, e.g., insurance losses. The interpretation of the outcome of a risk measure depends on the context in which it is used. Historically, there have been three main areas of application of risk measures:

- as representations of risk aversion in asset pricing models, with a leading paradigm being the use of the variance as a risk measure in Markowitz portfolio theory [1];
- as tools for the calculation of the insurance price corresponding to a risk – under this interpretation, risk measures are called *premium calculation principles* (see **Premium Calculation and Insurance Pricing**) in the classic actuarial literature, e.g., [2];
- as quantifiers of the economic capital that the holder of a particular portfolio or risks should safely invest in, e.g., [3].

This article is mainly concerned with the last-mentioned interpretation of risk measures.

The *economic* or *risk capital* (see **Solvency**) held by a (re)insurer corresponds to the level of safely invested assets used to protect itself against unexpected volatility of its portfolio's outcome. One has to distinguish economic capital from regulatory capital, which is the minimum required economic capital level as set by the regulator. In fact, much of the impetus for the use of risk measures in the quantification of capital requirements comes from the area of regulating financial institutions. Banking supervision [4] and, increasingly, insurance regulation [5] have been promoting the development of companies' *internal models* for modeling risk exposures. In this context, the application of a risk measure (most prominently *value-at-risk* (see **Value at Risk (VaR) and Risk Measures**)) on the modeled aggregate risk profile of the insurance company is required.

Economic capital generally exceeds the minimum set by the regulator. Subject to this constraint, economic capital is determined so as to maximize performance metrics for the insurance company, such as total shareholder return [6]. Such maximization takes into account two conflicting effects of economic capital [7]:

- An insurance company's holding economic capital incurs costs for its shareholders that can be opportunity or frictional costs.
- Economic capital reduces the probability of default of the company as well as the severity of such default on its policyholders. This enables the insurance company to obtain a better rating of its financial strength, and thereby attract more insurance business at higher prices.

Calculation of the optimal level of economic capital using such arguments is quite complicated and depends on factors that are not always easy to quantify, such as frictional capital costs, and on further constraints, such as the ability of an insurance company to raise capital in a particular economic and regulatory environment.

We could, however, consider that there is a particular calibration of the (regulatory or other) risk measure, which gives for the insurance company's exposure, a level of economic capital that coincides with that actually held by the company. In that sense, risk measures can be used to interpret exogenously given economic capital amounts. Such interpretation can be in the context of capital being set to achieve a target rating, often associated with a particular probability of default. Discussion of economic capital in the context of risk measures should therefore be caveated as being *ex post*.

Finally, we note that the level of economic capital calculated by a risk measure may be a notional amount, as the company will generally not invest all its surplus in risk-free assets. This can be dealt with by absorbing the volatility of asset returns in the risk capital calculation itself.

Definition and Examples of Risk Measures

We consider a set of risks $\mathcal{X}$ that the insurance company can be exposed to. The elements $X \in \mathcal{X}$ are random variables, representing losses at a fixed time

horizon T (see **Risk Classification/Life; Risk Classification in Nonlife Insurance**). If under a particular state of the world ω , the variable $X(\omega) > 0$, then we consider this to be a loss, while negative outcomes will be considered as gains. For convenience, it is assumed throughout that the return from risk-free investment is 1 or, alternatively, that all losses in $\mathcal{X}$ are discounted at the risk-free rate. A risk measure ρ is then defined as a functional

$$\rho : \mathcal{X} \mapsto \mathbb{R} \quad (1)$$

If X corresponds to the aggregate net risk exposure of an insurance company (i.e., the difference between liabilities and assets, excluding economic capital), and the economic capital corresponds to $\rho(X)$, then we assume that the company defaults when $X > \rho(X)$.

In the terminology of [3] (and subject to some simplification), a risky position X is called *acceptable* if $\rho(X) < 0$, implying that some capital may be released without endangering the security of the holder of X , while $\rho(X) \geq 0$ means that X is a nonacceptable position and that some capital has to be added to it.

Some examples of simple risk measures proposed in the actuarial and financial literature (e.g., [8, 9]) are given below.

Example 1: Expected value principle

$$\rho(X) = \lambda E[X], \quad \lambda \geq 1 \quad (2)$$

Besides its application in insurance pricing, where it represents a proportional loading, this risk measure, in essence, underlies simple regulatory minimum requirements, such as the current EU Solvency rules, which determine capital as a proportion of an exposure measure, such as premium.

Example 2: Standard deviation principle

$$\rho(X) = E[X] + \kappa \sigma[X], \quad \kappa \geq 0 \quad (3)$$

In this case, the loading is risk sensitive, as it is a proportion of the standard deviation. This risk measure is encountered in reinsurance pricing (see **Reinsurance**), while also relating to Markowitz portfolio theory. In the context of economic capital, it is usually derived as an approximation to other risk

measures, with this approximation being accurate for the special case of multivariate normal (more generally elliptical) distributions [10].

Example 3: Exponential premium principle

$$\rho(X) = \frac{1}{a} \ln E[e^{aX}], \quad a > 0 \quad (4)$$

The exponential premium principle is a very popular risk measure in the actuarial literature, e.g., [11]. Part of the popularity stems from the fact that, in the classic ruin problem (see **Ruin Theory**), it gives the required level of premium associated with Kramer–Lundberg bounds for ruin probabilities. We note that this risk measure has been recently considered in the finance literature under the name *entropic risk measure* [12].

Example 4: Value-at-risk

$$\rho(X) = \text{VaR}_p(X) = F_X^{-1}(p), \quad p \in (0, 1) \quad (5)$$

where F_X is the cumulative probability distribution of X and F_X^{-1} is its (pseudo)inverse. $\text{VaR}_p(X)$ is easily interpreted as the amount of capital that, when added to the risk X , limits the probability of default to $1 - p$. Partly because of its intuitive attractiveness, value-at-risk has become the risk measure of choice for both banking and insurance regulators. For example, the UK regulatory regime for insurers uses $\text{VaR}_{0.995}(X)$ [13], while a similar risk measure has been proposed in the context of the new EU-wide Solvency II regime [5].

Example 5: Expected shortfall

$$\rho(X) = \text{ES}_p(X) = \int_p^1 F_X^{-1}(q) \, dq, \quad p \in (0, 1) \quad (6)$$

This risk measure, also known as *tail-(or conditional-)value-at-risk*, corresponds to the average of all VaR_p 's above the threshold p . Hence, it reflects both the probability and the severity of a potential default. Expected shortfall has been proposed in the literature as a risk measure correcting some of the theoretical weaknesses of value-at-risk [14]. Subject to continuity of F_X at the threshold VaR_p , expected shortfall coincides with the *tail conditional expectation*, defined by

$$\rho(X) = E[X | X > F_X^{-1}(p)] \quad (7)$$

Example 6: Distortion risk measure

$$\rho(X) = - \int_{-\infty}^0 (1 - g(1 - F_X(x))) dx + \int_0^{\infty} g(1 - F_X(x)) dx \quad (8)$$

where $g : [0, 1] \mapsto [0, 1]$ is increasing and concave [15]. This risk measure can be viewed as an expectation under a distortion of the probability distribution effected by the function g . It can be easily shown that expected shortfall is a special case obtained by a bilinear distortion [14]. Distortion risk measures can be viewed as Choquet integrals [16, 17], which are extensively used in the economics of uncertainty, e.g., [18]. An equivalent class of risk measures defined in the finance literature are known as *spectral risk measures* [19].

Properties of Risk Measures

The literature is rich in discussions of the properties of alternative risk measures, as well as the desirability of such properties (*see Individual Risk Models*), e.g., [2, 3, 9, 20]. In view of this, the current discussion is invariably selective.

An often required property of risk measures is that of *monotonicity*, stating

$$\text{If } X \leq Y, \text{ then } \rho(X) \leq \rho(Y) \quad (9)$$

This reflects the obvious requirement that losses that are always higher should also attract a higher capital requirement.

A further appealing property is that of *translation* or *cash invariance*,

$$\rho(X + a) = \rho(X) + a, \quad \text{for } a \in \mathbb{R} \quad (10)$$

This postulates that adding a constant loss amount to a portfolio increases the required risk capital by the same amount. We note that this has the implication that

$$\rho(X - \rho(X)) = \rho(X) - \rho(X) = 0 \quad (11)$$

which, in conjunction with monotonicity, facilitates the interpretation of $\rho(X)$ as the minimum capital amount that has to be added to X to make it acceptable.

Two conceptually linked properties are the ones of *positive homogeneity*,

$$\rho(bX) = b\rho(X), \quad \text{for } b \geq 0 \quad (12)$$

and *subadditivity*,

$$\rho(X + Y) \leq \rho(X) + \rho(Y), \quad \text{for all } X, Y \in \mathcal{X} \quad (13)$$

Positive homogeneity postulates that a linear increase in the risk exposure X also implies linear increase in risk. Subadditivity requires that the merging of risks should always yield a reduction in risk capital because of diversification.

Risk measures satisfying the four properties of monotonicity, translation invariance, positive homogeneity, and subadditivity have become widely known as *coherent* [3]. This particular axiomatization, also proposed in an actuarial context [16, 21], has achieved near-canonical status in the world of financial risk management. While value-at-risk generally fails the subadditivity property, because of its disregard for the extreme tails of distributions, part of its appeal to regulators and practitioners stems from its use as an approximation to a coherent risk measure.

Nonetheless, coherent risk measures have also attracted criticism because of their insensitivity to the aggregation of large positively dependent risks implied by the latter two properties, e.g., [20]. The weaker property of *convexity* has been proposed in the literature [22], a property already discussed in [23]. Convexity requires that:

$$\rho(\lambda X + (1 - \lambda)Y) \leq \lambda\rho(X) + (1 - \lambda)\rho(Y), \quad \text{for all } X, Y \in \mathcal{X} \text{ and } \lambda \in [0, 1] \quad (14)$$

Convexity, while retaining the diversification property, relaxes the requirement that a risk measure must be insensitive to aggregation of large risks. It is noted that subadditivity is obtained by combining convexity with positive homogeneity. Risk measures satisfying convexity, and applying increasing penalties for large risks have been proposed in [24].

Risk measures produce an ordering of risks, in the sense that $\rho(X) \leq \rho(Y)$ means that X is considered less risky than Y . One would wish that ordering should conform to standard economic theory, i.e., to be consistent with widely accepted notions of stochastic order such as first- and second-order stochastic dominance and convex order; see [9, 25].

It has been shown that, under some relatively mild technical conditions, risk measures that are monotonic and convex produce such a consistent ordering of risks [26].

A further key property relates to the dependence structure between risks under which the risk measure becomes *additive*

$$\rho(X + Y) = \rho(X) + \rho(Y) \quad (15)$$

as this implies a situation where neither diversification credits nor aggregation penalties are assigned. In the context of subadditive risk measures, *comonotonic additivity* is a sensible requirement, as it postulates that no diversification is applied in the case of comonotonicity (*see Comonotonicity*) (the maximal level of dependence between risks, e.g., [27]). On the other hand, one could require a risk measure to be *independent additive*. If such a risk measure is also consistent with the stop-loss or convex order, by the results of [28], it is guaranteed to penalize any positive dependence by being superadditive (i.e., $\rho(X + Y) \geq \rho(X) + \rho(Y)$) and reward any negative dependence by being subadditive.

The risk measures defined above satisfy the following properties:

Expected value principle: monotonic, positive homogenous, additive for all dependence structures;

Standard deviation principle: translation invariant, positive homogenous, subadditive;

Exponential premium principle: monotonic, translation invariant, convex, independent additive;

Value-at-risk: monotonic, translation invariant, positive homogenous, subadditive for joint elliptically distributed risks [10], comonotonic additive;

Expected shortfall: monotonic, translation invariant, positive homogenous, subadditive, comonotonic additive;

Distortion risk measure: monotonic, translation invariant, positive homogenous, subadditive, comonotonic additive.

Finally, we note that all risk measures discussed in this contribution are *law invariant*, meaning that $\rho(X)$ only depends on the distribution function of X [21, 29]. This implies that two risks characterized by the same probability distribution would be allocated the same amount of economic capital.

Constructions and Representations of Risk Measures

Indifference Arguments

Economic theories of choice under risk seek to model the preferences of economic agents with respect to uncertain payoffs. They generally have representations in terms of *preference functionals* $V : -\mathcal{X} \mapsto \mathbb{R}$, in the sense that

$$-X \text{ is preferred to } -Y \iff V(-X) \geq V(-Y) \quad (16)$$

(Note that the minus sign is applied because we have defined risk as losses, while preference functionals are typically applied on payoffs.)

Then a risk measure can be defined by assuming that the addition to initial wealth W of a liability X and the corresponding capital amount $\rho(X)$ does not affect preferences [8] (*see Risk-Neutral Pricing: Importance and Relevance*)

$$V(W_0 - X + \rho(X)) = V(W_0) \quad (17)$$

Often in this context, $W = 0$ is assumed for simplicity.

The leading paradigm of choice under risk is the von Neumann–Morgenstern *expected utility theory* (*see Utility Function*) [30], under which

$$V(W) = E[u(W)] \quad (18)$$

where u is an increasing and concave *utility function*. A popular choice of utility function is the *exponential utility*

$$u(w) = \frac{1}{a} (1 - e^{-aw}), \quad a > 0 \quad (19)$$

It can be easily seen that equations (17), (18), and (19) yield the exponential premium principle defined in the section “Definition and Examples of Risk Measures”.

An alternative theory is the *dual theory of choice under risk* [31], under which

$$V(W) = - \int_{-\infty}^0 (1 - h(1 - F_W(w))) dw + \int_0^{\infty} h(1 - F_W(w)) dw \quad (20)$$

where $h : [0, 1] \mapsto [0, 1]$ is increasing and convex. It can then be shown that the risk measure obtained

from equations (17) and (20) is a distortion risk measure with $g(s) = 1 - h(1 - s)$. For the function

$$h(s) = 1 - (1 - s)^{\frac{1}{\gamma}}, \quad \gamma > 1 \quad (21)$$

the well-known *proportional hazards transform*, with $g(s) = s^{1/\gamma}$ is obtained [15].

More detailed discussions of risk measures resulting from alternative theories of choice under risk (see **Axiomatic Models of Perceived Risk**) and references to the associated economics literature are given in [24, 32].

It should also be noted that the construction of risk measures from economic theories of choice must not necessarily be *via* indifference arguments. If a risk measure satisfies the convexity and monotonicity properties, then by setting $U(W) = -\rho(-W)$ we obtain a monotonic concave preference functional. The translation invariance property of the risk measure then makes U also translation invariant. Hence we could consider convex risk measures as the subset of concave preference functionals that satisfy the translation invariance property (subject to a minus sign). Such preference functionals are sometimes called *monetary utility functions*, as their output can be interpreted as being in units of money rather than of an abstract notion of satisfaction.

Axiomatic Characterizations

An alternative approach to deriving risk measures is by fixing a set of properties that risk measures should satisfy and then seeking an explicit functional representation (see **Axiomatic Measures of Risk and Risk-Value Models**).

For example, coherent (i.e., monotonic, translation invariant, positive homogenous, and subadditive) risk measures can be represented by Artzner *et al.* [3]

$$\rho(X) = \sup_{\mathbb{P} \in \mathcal{P}} E_{\mathbb{P}}[X] \quad (22)$$

where $\mathcal{P}$ is a set of probability measures. By including the comonotonic additivity property, one gets the more specific structure of $\mathcal{P} = \{\mathbb{P} : \mathbb{P}(A) \leq v(A) \text{ for all sets } A\}$, where v is a submodular set function known as a (*Choquet*) *capacity* [17]. The additional property of law of invariance enables writing $v(A) = g(\mathbb{P}_0(A))$, where $\mathbb{P}_0$ is the objective probability measure and g a concave distortion function [21]. This finally yields a representation of coherent, comonotonic additive, law-invariant risk measures as distortion risk measures. An alternative

route toward this representation is given by Kusuoka [29].

The probability measures in $\mathcal{P}$ have been termed *generalized scenarios* [3] with respect to which the worst case expected loss is considered. On the other hand, representations such as equation (22) have been derived in the context of robust statistics [33] and decision theory, known as the *multiple-priors model* [34].

A related representation result for convex risk measures is derived in [22], while results for independent additive risk measures are given in [35, 36].

Reweighting Probabilities

An intuitive construction of risk measures is by reweighting the probability distribution of the underlying risk

$$\rho(X) = E[X\zeta(X)] \quad (23)$$

where ζ is generally assumed to be an increasing function with $E[\zeta(X)] = 1$ and representation (23) could be viewed as an expectation under a change of measure. Representation (23) is particularly convenient when risk measures and related functionals have to be evaluated by Monte Carlo simulation.

Many well-known risk measures can be obtained in this way. For example, making appropriate assumptions on F_X and g , one can easily show that for distortion measures it is

$$\rho(X) = E[Xg'(1 - F_X(X))] \quad (24)$$

On the other hand, the exponential principle can be written as

$$\rho(X) = E \left[X \int_0^1 \frac{e^{\gamma a X}}{E[e^{\gamma a X}]} d\gamma \right] \quad (25)$$

The latter representation is sometimes called a *mixture of Esscher principles* and studied in more generality in [35, 36].

Capital Allocation

Problem Definition

Often the requirement arises that the risk capital calculated for an insurance portfolio has to be allocated to business units. There may be several reasons for such a capital allocation exercise, the main ones being performance measurement/management and insurance pricing.

Capital allocation is not a trivial exercise, given that, in general, the risk measure used to set the aggregate capital is not additive. In other words, if one has an aggregate risk Z for the insurance company, breaking down to subportfolios $X_1, \dots, X_n$, such that

$$Z = \sum_{j=1}^n X_j \quad (26)$$

it generally is

$$\rho(Z) \neq \sum_{j=1}^n \rho(X_j) \quad (27)$$

because of diversification/aggregation issues.

The capital allocation problem then consists of finding constants $d_1, \dots, d_n$ such that

$$\sum_{j=1}^n d_j = \rho(Z) \quad (28)$$

where the allocated capital amount d_i should, in some way, reflect the risk of subportfolio X_i . Early papers in the actuarial literature that deal with cost allocation problems in insurance are [37, 38], the former taking a risk theoretical view, and the later examining alternative allocation methods from the perspective of cooperative game theory. A specific application of cooperative game theory to risk capital allocation, including a survey of the relevant literature, is [39].

Marginal Cost Approaches

Marginal cost approaches associate allocated capital to the impact that changes in the exposure to subportfolios have on the aggregate capital. Denote for vector of weights $\mathbf{w} \in [0, 1]^n$,

$$Z^{\mathbf{w}} = \sum_{j=1}^n w_j X_j \quad (29)$$

Then the marginal cost of each subportfolio is given by

$$MC(X_i; Z) = \left. \frac{\partial \rho(Z^{\mathbf{w}})}{\partial w_i} \right|_{\mathbf{w} = \mathbf{1}} \quad (30)$$

subject to appropriate differentiability assumptions. If the risk measure is positive homogenous, then by Euler's theorem we have that

$$\sum_{j=1}^n MC(X_j; Z) = \rho(Z) \quad (31)$$

and we can hence use marginal costs $d_i = MC(X_i; Z)$ directly as the capital allocation.

If the risk measure is in addition subadditive, then we have that [40]

$$d_i = MC(X_i; Z) \leq \rho(X_i) \quad (32)$$

i.e., the allocated capital amount is always lower than the stand-alone risk capital of the subportfolio. This corresponds to the game theoretical concept of the *core*, in that the allocation does not provide an incentive for splitting the aggregate portfolio. This requirement is consistent with the subadditivity property, which postulates that there is always a benefit in pooling risks.

In the case that no such strong assumptions as positive homogeneity (and subadditivity) are made with respect to the risk measure, marginal costs will, in general, not yield an appropriate allocation, as they will not add up to the aggregate risk. Cooperative game theory then provides an alternative allocation method, based on the *Aumann–Shapley value* [41], which can be viewed as a generalization of marginal costs

$$AC(X_i, Z) = \int_0^1 MC(X_i; \gamma Z) d\gamma \quad (33)$$

It can easily be seen that if we set $d_i = AC(X_i, Z)$ then the d_i s add up to $\rho(Z)$ and that for positive homogenous risk measures the Aumann–Shapley allocation reduces to marginal costs. Early applications of the Aumann–Shapley value to cost allocation problems are [42, 43].

For the examples of risk measures that were introduced in the section “Definition and Examples of Risk Measures”, the following allocations are obtained from marginal costs/Aumann–Shapley.

Example 7: Expected value principle

$$d_i = \lambda E[X_i] \quad (34)$$

Example 8: Standard deviation principle

$$d_i = E[X_i] + \kappa \frac{\text{Cov}(X_i, Z)}{\sigma[Z]} \quad (35)$$

Example 9: Exponential premium principle

$$d_i = \int_0^1 \frac{E[X_i \exp(\gamma a Z)]}{E[\exp(\gamma a Z)]} d\gamma \quad (36)$$

Example 10: Value-at-risk [44]

$$d_i = E[X_i | Z = \text{VaR}_p(Z)] \quad (37)$$

under suitable assumptions on the joint probability distribution of (X_i, Z) .

Example 11: Expected shortfall [44]

$$d_i = E[X_i | Z > \text{VaR}_p(Z)] \quad (38)$$

under suitable assumptions on the joint probability distribution of (X_i, Z) .

Example 12: Distortion risk measure [45]

$$d_i = E[X_i g'(1 - F_Z(Z))] \quad (39)$$

assuming representation (24) is valid.

Alternative Approaches

While marginal cost-based approaches are well established in the literature, there are a number of alternative approaches to capital allocation. For example, we note that marginal costs generally depend on the joint distribution of the individual subportfolio and the aggregate risk. In some cases, this dependence may not be desirable, for example, when one tries to measure the performance of subportfolios to allocate bonuses. In that case, a simple *proportional repartition of costs* [38] may be appropriate:

$$d_i = \rho(X_i) \frac{\rho(Z)}{\sum_{j=1}^n \rho(X_j)} \quad (40)$$

Different issues emerge when the capital allocation is to be used for managing the performance of the aggregate portfolio, as measured by a particular metric, such as return on capital. Assume that $\hat{X}_i$, $i = 1, \dots, n$ correspond to the liabilities from subportfolio i minus reserves corresponding to those liabilities, such that $E[\hat{X}_i] = 0$. We then have the breakdown

$$X_i = \hat{X}_i - p_i \quad (41)$$

where p_i corresponds to the underwriting profit from the insurer's subportfolio (e.g., line of business) i , such that $\sum_{j=1}^n \hat{X}_j = \hat{Z}$ and $\sum_{j=1}^n p_j = p$. Then we define the return on capital for the whole insurance portfolio by

$$RoC = \frac{p}{\rho(\hat{Z})} \quad (42)$$

This is discussed in depth in [44] for the case that ρ is a coherent risk measure. It is then considered whether assessing the performance of subportfolios by

$$RoC_i = \frac{p_i}{d_i} \quad (43)$$

where d_i represents capital allocated to $\hat{X}_i$, provides the right incentives for optimizing performance. It is shown that marginal costs is the unique allocation mechanism that satisfies this requirement as set out

in that paper. A closely related argument is that under the marginal cost allocation, a portfolio balanced to optimize aggregate return on capital has the property that $RoC = RoC_i$, for all i . While this produces a useful performance yardstick that can be used throughout the company, some care has to be taken when applying marginal cost methodologies. In particular, if the marginal capital allocation to a subportfolio is small, e.g., for reasons of diversification, the insurer should be careful not to let that fact undermine underwriting standards. A proportional allocation method could also be used for reference, to avoid that danger.

Often, one may be interested in calculating capital allocations that are, in some sense, optimal. For example, in [46] capital allocations are calculated such that a suitably defined distance function between individual subportfolios and allocated capital levels is minimized. This methodology reproduces many capital allocation methods found in the literature, while also considering the case that aggregate economic capital is exogenously given rather than calculated *via* a risk measure. A different optimization approach to capital allocation is presented in [47].

An alternative strand of the literature on capital allocation relates to the pricing of the policyholder deficit due to the insurer's potential default [48]. This is discussed in slightly greater detail in the section titled "Frictional Capital Costs and the Cost of Default".

Economic Capital and Insurance Pricing

Cost of Capital

A way of associating risk measures and economic capital with insurance prices is *via cost-of-capital* arguments. It is considered that the shareholders of an insurance company incur an opportunity cost by providing economic capital. Therefore, they also require a return on that capital that is in excess of the risk-free rate. This is typically calculated *via* equilibrium arguments, with the methodology of *weighted average cost of capital* (WACC) being the prime example [49]. It is, furthermore, assumed that the additional return on capital will be earned by including a cost-of-capital adjustment in insurance premiums.

If we denote by CoC the cost of capital associated with the insurance company, then, using the notation

of the previous section, the required profit for its insurance portfolio $Z = \hat{Z} - p$ is:

$$p = CoC \cdot \rho(\hat{Z}) \quad (44)$$

It should be apparent from the preceding discussion that cost of capital and return on capital are closely linked concepts; in fact evaluation of the former often leads to a target level for the latter.

A capital allocation $\sum_{j=1}^n d_j = \rho(\sum_{j=1}^n \hat{X}_j) = \rho(\hat{Z})$ then yields the required profit for each subportfolio:

$$p_i = CoC \cdot d_i \quad (45)$$

Frictional Capital Costs and the Cost of Default

Cost-of-capital approaches to insurance pricing have been criticized in the literature for a number of reasons, including [6]:

- The return on (or cost of) capital considered may be an inadequate measure of performance, as the total shareholder return is influenced by other factors too.
- The methodology does not explicitly allow for the potential default of the insurer. Hence, by associating a fixed cost with economic capital, the benefits of increased policyholder security are disregarded.

An alternative approach is to break down the insurance price into three parts [7]:

- the economic (market consistent) value of the insurance liability;
- the frictional cost of holding capital;
- the cost to policyholders in case of the default by the insurer.

The economic value of the liability Z , denoted here by $EV(Z)$, where EV is a linear pricing functional derived by a financial valuation method, e.g., by equilibrium or no-arbitrage arguments. The frictional costs, which may comprise double taxation, agency costs, and the costs of financial distress [7], may be written in their simplest form as a fixed percentage fc of aggregate capital, i.e., $fc \cdot \rho(Z)$.

The loss to policyholders caused by the insurer's default or *policyholder deficit* is given by

$$(Z - \rho(Z))_+ \quad (46)$$

that is, by the excess of the insurer's liabilities over its assets. It is argued by Myers and Read [48] that the economic value of this cost should be removed from the insurance premium, as it corresponds to a loss to policyholders, given the insurer's shareholders' limited liability; an expression similar to equation (46) has thus been termed the *limited liability put option* or *default option*.

The total premium for the insurer can then be calculated as

$$EV(Z) + fc \cdot \rho(Z) - EV[(Z - \rho(Z))_+] \quad (47)$$

On the basis of expression (48), and assuming ρ is a coherent risk measure, marginal costs give us the allocation of the insurance price for subportfolio X_i by

$$EV(X_i) + fc \cdot d_i - EV[(X_i - d_i)\mathbf{I}_{\{Z > \rho(Z)\}}] \quad (48)$$

where $\mathbf{I}_A$ denotes the indicator of set A .

The discussion on allocating capital by the value of the default option is, in the main, motivated by Myers and Read [48]. That paper, as well as [50] that anticipates it, adopts a slightly different approach, whereby, the aggregate capital allocated to subportfolios is exogenously given rather than calculated *via* a risk measure. Moreover, these papers are set in a dynamic framework rather than the simple 1-period one adopted in this contribution.

Incomplete Markets and Risk Measures

As discussed in the section "Risk Measurement and Capital", risk measures have been traditionally used in the insurance industry to calculate prices. The difference $\rho(X) - E[X]$ then is considered as a *safety loading*. When the risk measures are law invariant, such as those considered here, then the safety loading depends only on the distribution of X and does not reflect market conditions. Moreover, as illustrated in [51], using law-invariant risk measures as pricing functionals, e.g., to reflect market frictions, can also be problematic.

In financial economics, securities are priced using no-arbitrage arguments, which, in a complete market, result in the price of a risk equal to the cost of its replication by traded instruments. Such a market consistent approach to valuation of insurance liabilities is propagated in the context of Solvency II [5]. However, insurance markets are typically incomplete, meaning that no exact replication can be achieved by trading. This motivates approaches

where the price equals the cost of replication with some acceptable level of accuracy. Two examples of incomplete market approaches based on partial replication are mean variance [52] and quantile hedging [53].

The formulation of such pricing methods often utilizes risk measures to define the quality of replication. Different calibrations of these risk measures, corresponding to different levels of risk aversion, give rise to a range of possible prices. It can thus be argued that there is a continuum between pricing and holding risk capital in incomplete markets, with the main difference being the degree of risk aversion considered in each application. (A rather different approach to incomplete market pricing is based on indifference arguments in a dynamic setting is [54], which is closely related to the risk measures discussed in the section “Indifference Arguments.”) Hence one could consider financial pricing of insurance as being a dynamic extension of the traditional actuarial risk measures or premium principles, taking place in a richer economic environment where dynamic trading is possible.

References

- [1] Markowitz, H. (1952). Portfolio selection, *Journal of Finance* **7**, 77–91.
- [2] Goovaerts, M.J., de Vylder, F. & Haezendonck, J. (1984). *Insurance Premiums*, Elsevier Science.
- [3] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**(3), 203–228.
- [4] Basel Committee on Banking Supervision. (2003). *The New Basel Capital Accord*. Bank of International Settlements.
- [5] European Commission. (2002). *Considerations on the Design of a Future Prudential Supervisory System*, MARKT/2535/02–EN. European Commission.
- [6] Exley, C.J. & Smith, A.D. (2006). *The Cost of Capital for Financial Firms*, presented to the Institute of Actuaries, 23 January 2006.
- [7] Hancock, J., Huber, P. & Koch, P. (2001). *The Economics of Insurance: How Insurers Create Value for Shareholders*, Swiss Re Technical Publishing.
- [8] Bühlmann, H. (1970). *Mathematical Methods in Risk Theory*, Springer, Berlin.
- [9] Denuit, M., Dhaene, J., Goovaerts, M.J. & Kaas, R. (2005). *Actuarial Theory for Dependent Risk: Measures, Orders and Models*, John Wiley & Sons, New York.
- [10] Embrechts, P., McNeil, A. & Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls, in *Risk Management: Value at Risk and Beyond*, M.A.H. Dempster, ed, Cambridge University Press, Cambridge.
- [11] Gerber, H.U. (1974). On additive premium calculation principles, *ASTIN Bulletin* **7**(3), 215–222.
- [12] Föllmer, H. & Schied, A. (2002). Robust preferences and convex risk measures, in *Advances in Finance and Stochastics*, K. Sadmann, ed, Springer, Berlin.
- [13] Financial Services Authority (2006). Prudential sourcebook for Insurers (ISPRU) in *The Full Handbook, Financial Services Authority*, <http://fsahandbook.info/FSA/html/handbook>.
- [14] Wirth, J.L. & Hardy, M.R. (1999). A synthesis of risk measures for capital adequacy, *Insurance Mathematics and Economics* **25**(3), 337–347.
- [15] Wang, S.S. (1996). Premium calculation by transforming the premium layer density, *ASTIN Bulletin* **26**(1), 71–92.
- [16] Denneberg, D. (1990). Distorted probabilities and insurance premiums, *Methods of Operations Research* **63**, 3–5.
- [17] Denneberg, D. (1994). *Non-Additive Measure and Integral*, Kluwer Academic Publishers, Dordrecht.
- [18] Schmeidler, D. (1989). Subjective probability and expected utility without additivity, *Econometrica* **57**(3), 571–587.
- [19] Acerbi, C. (2002). Spectral measures of risk: a coherent representation of subjective risk aversion, *Journal of Banking and Finance* **26**(7), 1505–1518.
- [20] Goovaerts, M., Dhaene, J. & Kaas, R. (2003). Economic capital allocation derived from risk measures, *North American Actuarial Journal* **7**(2), 44–59.
- [21] Wang, S.S., Young, V.R. & Panjer, H.H. (1997). Axiomatic characterization of insurance prices, *Insurance Mathematics and Economics* **21**(2), 173–183.
- [22] Föllmer, H. & Schied, A. (2002). Convex measures of risk and trading constraints, *Finance and Stochastics* **6**(4), 429–447.
- [23] Deprez, O. & Gerber, H.U. (1985). On convex principles of premium calculation, *Insurance Mathematics and Economics* **4**(3), 179–189.
- [24] Tsanakas, A. & Desli, E. (2003). Risk measures and theories of choice, *British Actuarial Journal* **9**(4), 959–991.
- [25] Müller, A. & Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*, John Wiley & Sons.
- [26] Bäuerle, N. & Müller, A. (2006). Stochastic orders and risk measures: consistency and bounds, *Insurance Mathematics and Economics* **38**(1), 132–148.
- [27] Dhaene, J., Denuit, M., Goovaerts, M.J., Kaas, R. & Vyncke, D. (2002). The concept of comonotonicity in actuarial science and finance: theory, *Insurance Mathematics and Economics* **31**(1), 3–33.
- [28] Dhaene, J. & Goovaerts, M.J. (1996). Dependency of risks and stop-loss order, *ASTIN Bulletin* **26**(2), 201–212.
- [29] Kusuoka, S. (2001). On law-invariant coherent risk measures, *Advances in Mathematical Economics* **3**, 83–95.
- [30] von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*, Princeton University Press.

- [31] Yaari, M. (1987). The dual theory of choice under risk, *Econometrica* **55**(1), 95–115.
- [32] Denuit, M., Dhaene, J., Goovaerts, M., Kaas, R. & Laeven, R. (2006). Risk measurement with equivalent utility principles, *Statistics and Decisions* **24**(1), 1–25.
- [33] Huber, J. (1981). *Robust Statistics*, John Wiley & Sons.
- [34] Gilboa, I. & Schmeidler, D. (1989). Maxmin expected utility with a non-unique prior, *Journal of Mathematical Economics* **18**, 141–153.
- [35] Gerber, H.U. & Goovaerts, M.J. (1981). On the representation of additive principles of premium calculation, *Scandinavian Actuarial Journal* **4**, 221–227.
- [36] Goovaerts, M., Kaas, R., Laeven, R. & Tang, Q. (2004). A comonotonic image of independence for additive risk measures, *Insurance Mathematics and Economics* **35**(3), 581–594.
- [37] Bühlmann, H. (1985). Premium calculation from top down, *ASTIN Bulletin* **15**(2), 89–101.
- [38] Lemaire, J. (1984). An application of game theory: cost allocation, *ASTIN Bulletin* **14**(1), 61–81.
- [39] Denault, M. (2001). Coherent allocation of risk capital, *Journal of Risk* **4**(1), 1–34.
- [40] Aubin, J.-P. (1981). Cooperative fuzzy games, *Mathematics of Operations Research* **6**(1), 1–13.
- [41] Aumann, R.J. & Shapley, L.S. (1974). *Values of Non-Atomic Games*, Princeton University Press.
- [42] Billera, L.J. & Heath, D.C. (1982). Allocation of shared costs: a set of axioms yielding a unique procedure, *Mathematics of Operations Research* **7**(1), 32–39.
- [43] Mirman, L. & Tauman, Y. (1982). Demand compatible equitable cost sharing prices, *Mathematics of Operations Research* **7**(1), 40–56.
- [44] Tasche, D. (2004). Allocating portfolio economic capital to sub-portfolios, *Economic Capital: A practitioner's Guide*, Risk Books, pp. 275–302.
- [45] Tsanakas, A. (2004). Dynamic capital allocation with distortion risk measures, *Insurance Mathematics and Economics* **35**, 223–243.
- [46] Dhaene, J., Tsanakas, A., Valdez, E. & Vanduffel, S. (2005). Optimal capital allocation principles, *9th International Congress on Insurance: Mathematics and Economics*, Quebec, July 6–8.
- [47] Laeven, R.J.A. & Goovaerts, M.J. (2003). An optimization approach to the dynamic allocation of economic capital, *Insurance Mathematics and Economics* **35**(2), 299–319.
- [48] Myers, S.C. & Read, J.A. (2001). Capital allocation for insurance companies, *Journal of Risk and Insurance* **68**(4), 545–580.
- [49] Brealey, R.A. & Myers, S.C. (2002). *Principles of Corporate Finance*, 7th Edition, McGraw-Hill.
- [50] Merton, R. & Perold, A. (1993). Theory of risk capital in financial firms, *Journal of Applied Corporate Finance* **6**(3), 16–32.
- [51] Castgnoli, E., Maccheroni, F. & Marinacci, M. (2004). Choquet insurance pricing: a caveat, *Mathematical Finance* **14**(3), 481–485.
- [52] Schweizer, M. (1992). Mean-variance hedging for general claims, *Annals of Applied Probability* **2**, 171–179.
- [53] Föllmer, H. & Leukert, P. (1999). Quantile hedging, *Finance and Stochastics* **3**(3), 251–273.
- [54] Young, V.R. & Zariphopoulou, T. (2002). Pricing dynamic insurance risks using the principle of equivalent utility, *Scandinavian Actuarial Journal* **4**, 246–279.

Related Articles

Mathematical Models of Credit Risk

Model Risk

Risk-Neutral Pricing: Importance and Relevance

Structural Models of Corporate Credit Risk

ANDREAS TSANAKAS

Risk–Benefit Analysis for Environmental Applications

In a risk–benefit analysis, the issue examined is the trade-off between losses suffered from illnesses or diseases, due to exposures to an environmental contaminant, and losses due to toxic effects induced by environmental remediation, e.g., chemicals used in cleanup or disinfectants used to reduce microbial exposures (*see* **Environmental Hazard; Environmental Health Risk**). Perhaps, this process is more appropriately labeled a risk–risk analysis. An additional societal complication arises when risks and benefits are incurred by different individuals. A risk management or societal question is whether the benefits (e.g., reductions of severity or incidences of illnesses, diseases, or death) outweigh the risks introduced by interventions (e.g., by-products from disinfectants used for drinking water treatment) (*see* **Enterprise Risk Management (ERM)**). The solution depends not only on the incidences of adverse health effects, but also on their relative costs (e.g., medical expenses and/or number of workdays lost). A risk–benefit issue also arises for drugs with unwanted side effects and for vitamins and essential minerals that become toxic at high doses. The total relative cost or loss due to an adverse health effect is the incidence of the effect times its relative cost or loss. In the following section, a quantitative procedure is described for estimating the optimum dose for a chemical intervention, drug, vitamin, or essential mineral that minimizes the overall expected relative loss from the risk of illness or disease due to environmental exposure and the expected relative loss induced by chemical, drug, or dietary intervention.

Method for Optimum Dose of Intervention

Let d represent the dose of a chemical introduced to reduce the incidence of an adverse effect. Let $P_1(d)$ represent the expected probability of the adverse effect in the presence of an intervention dose of d units. For example, $P_1(d)$ might represent the probability of illness for a person exposed at d units of a chemical treatment (e.g., disinfectant, drug, vitamin, or essential mineral). It is assumed that $P_1(d)$

is a monotonically decreasing function of d , i.e., $0 < P_1(d) \leq P_1(0)$, where $P_1(0)$ is the background incidence without intervention. Let c_1 represent the expected cost (loss) due to the occurrence of the adverse effect. Loss is expressed as some appropriate metric, e.g., monetary cost of treatment of a disease, length of hospital stay, lost workdays, or life shortening. Then, the expected loss due to the adverse event is the product of the probability of the adverse event times the loss per event, $E_1(\text{loss}) = c_1 P_1(d)$, as a monotonically decreasing function of d .

Let $P_2(d)$ represent the probability of a different adverse effect (e.g., toxic effect, illness, or disease) induced by a dose, d , of an intervention agent. It is assumed that $P_2(d)$ is a monotonically increasing function of d . Note that $P_1(d)$ and $P_2(d)$ must be expressed on the same basis, e.g., probability per serving or per lifetime. Let c_2 represent the cost (loss) due to an adverse effect induced by the intervention agent, where c_2 must be expressed in the same metric as c_1 . The expected loss induced by the intervention product, $E_2 = c_2 P_2(d)$, is a monotonically increasing function d .

Chen and Gaylor [1] used the principle of minimum expected loss for setting the optimum beneficial dose of a compound accompanied by a potential toxic effect.

The total expected loss from a product treated with a dose, d , of an intervention product is

$$E(\text{loss}) = c_1 P_1(d) + c_2 P_2(d) \quad (1)$$

This results in a U-shaped curve as a function of dose. Let $E'(\text{loss})$ represent the first derivative of $E(\text{loss})$ with respect to d . The expected loss is minimized when

$$E'(\text{loss}) = c_1 P'_1(d) + c_2 P'_2(d) = 0 \quad (2)$$

The value of d (optimum dose) that satisfies this expression

$$\frac{P'_2(d)}{P'_1(d)} = -\frac{c_1}{c_2} \quad (3)$$

minimizes the total expected loss, which depends on the relative (not the absolute) losses from the initial and induced effects. It would appear that the values of relative losses may be more readily determined than absolute losses. Relative losses are quite amenable to any subsequent sensitivity analysis.

The optimum dose depends on the slopes (first derivatives) of the dose–response curves, $P_1(d)$ and

$P_2(d)$. Thus, as long as the mathematical model used for $P_1(d)$ and $P_2(d)$ approximates the true dose–response slope at low doses, the estimate of the optimum dose and relative expected loss will be fairly accurate. It is advantageous that the choices of the models are not crucial as long as the models approximate the slope of the dose–response curves in the region of the optimum dose.

Further, suppose it is of interest to find the optimum dose that minimizes the total incidence of adverse events, apart from any losses (costs) per event. Then, set $c_1 = c_2 = 1$ and the optimum dose is simply the value of d that satisfies $P'_2(d) = -P'_1(d)$.

Example for Optimum Dose of Intervention

Suppose the background probability of an illness or disease without intervention, i.e., $d = 0$, is $P_1(0) = 0.001$. Suppose a dose of one unit of the intervention product reduces the probability of the illness or disease by a factor of 2. This property is described by the negative exponential model

$$P_1(d) = 0.001 \cdot e^{-0.693d} \quad (4)$$

with a first derivative of

$$P'_1(d) = -0.000693e^{-0.693d} \quad (5)$$

Suppose the illness or disease induced by the intervention product is described by a logistic model and occurs with a background incidence of 0.0001 and one unit of the intervention product triples the background incidence

$$P_2(d) = \frac{1}{[1 + e^{(9.21-1.1d)}]} \quad (6)$$

with a first derivative of

$$P'_2(d) = \frac{1.1e^{(9.21-1.1d)}}{[1 + e^{(9.21-1.1d)}]^2} \quad (7)$$

From equation (3), the optimum dose (D) that minimizes the total expected loss is given by the solution of $P'_2(D)/P'_1(D) = -(c_1/c_2)$, giving

$$\frac{1.1e^{(9.21-1.1D)}}{[1 + e^{(9.21-1.1D)}]^2 \cdot (-0.000693e^{-0.693D})} = -c_1/c_2 \quad (8)$$

Table 1 Optimum dose (D) for various values of relative loss (c_1/c_2), estimated associated incidence of the environmentally caused disease, $P_1(D)$, the incidence of the disease induced by the intervention, $P_2(D)$, the total expected relative loss, $C(D) = [c_1 P_1(D) + c_2 P_2(D)]$, at the optimum dose, and the total expected loss for no intervention, $C(0) = [c_1 P_1(0) + c_2 P_2(0)]$, with $d = 0$

c_1/c_2	D	$P_1(D)$	$P_2(D)$	$C(D)$	$C(0)$	$C(D)/C(0)$
1/4	0.25	0.00084	0.00013	0.00136	0.00140	0.97
1/2	0.64	0.00064	0.00020	0.00104	0.00120	0.87
1/1	1.04	0.00049	0.00031	0.00080	0.00110	0.73
2/1	1.42	0.00037	0.00048	0.00122	0.00210	0.58
4/1	1.81	0.00029	0.00073	0.00189	0.00410	0.46

In this example, there is no simple analytical solution for the optimum dose (D) that minimizes the total expected loss (cost) from both illnesses or diseases. However, numerical solutions are readily attainable. Some results are illustrated in Table 1 for various values of the ratio of the relative loss (cost) of the environmentally caused illness or disease (c_1) and the relative loss (cost) induced by the intervention.

Uncertainty in the Calculation of the Optimum Dose of Intervention

In the example, if the dose–response models, $P_1(d)$ and $P_2(d)$, are correct and $c_1/c_2 = 1$, the optimum dose is 1.04 units with an expected total loss (cost) of 0.00080. However, if the ratio of the slopes of the dose responses, $P'_2(d)/P'_1(d)$, is overestimated by a factor of 4 (equivalently c_1/c_2 is underestimated by a factor of 4), a dose of 0.25 units would be selected with a total expected loss of $(0.00084 + 0.00013) = 0.00097$. This represents about a 21% increase in the total expected cost $(0.00097/0.00080)$. Similarly, if the ratios of the slopes of the dose responses is underestimated by a factor of 4 (equivalently c_1/c_2 is overestimated by a factor of 4), a dose of 1.81 units would be selected with a total expected loss of $(0.00029 + 0.00073) = 0.00102$, which is 27.5% larger than the expected total loss at the optimum dose. In this example, the total expected cost is not very sensitive to the choice of models and resulting slopes or the choice of c_1/c_2 .

Discussion

As c_1 increases relative to c_2 , the optimum dose increases, resulting in a decrease in the incidence of the initial illness or disease and accompanied by an increase in the illness or disease induced by intervention and increased reduction in the total expected loss relative to the expected total loss without intervention, $C(D)/C(0)$. If the loss (cost) of intervention is high, the optimum dose will be low and little gain over the *status quo* is accomplished. If the loss (cost) of intervention is low, then the optimum dose will be large and substantial reduction in cost relative to the *status quo* may be realized.

Additional examples with various dose–response relationships are presented by Gaylor [2].

References

- [1] Chen, J.J. & Gaylor, D.W. (1986). The use of decision-theoretic approach in regulating toxicity, *Regulatory Toxicology and Pharmacology* **6**, 274–283.
- [2] Gaylor, D.W. (2005). Risk/benefit assessments of human diseases: optimum dose for intervention, *Risk Analysis* **25**, 161–168.

Related Articles

Dose–Response Analysis

Environmental Risk Regulation

DAVID W. GAYLOR

Risk-Neutral Pricing: Importance and Relevance

In seminal papers, Black and Scholes [1] and Merton [2] (*see* **Default Correlation; Default Risk**) derived the price of an option analytically, under some regularity assumptions on the market behavior, transactions, growth etc. Although these assumptions are not often applicable in practice, however, these papers changed the way of thinking of financial markets all over the world. One of the main assumptions in their derivation, along with others, is the sublime introduction of risk-neutral measures (*see* **Equity-Linked Life Insurance**) in their calculations. They have shown that whether the market goes up or down, one can keep a portfolio risk neutral using certain strategies, such as, Δ -hedging. Finally, they calculated the price of an option, assuming growth of the portfolio at the risk-free interest rate, which gives rise to a risk-neutral measure. It became apparent that whatever is the market behavior, under risk-neutral measure, the asset price follows a discounted martingale (*see* **Insurance Pricing/Nonlife; Multistate Models for Life Insurance Mathematics**) which indicates the fair pricing. Thus, the central theme of option pricing becomes finding an equivalent risk-neutral measure under which the asset price process is a discounted martingale.

In the following, we first describe the binomial model (*see* **Volatility Smile**) due to Cox, Ross, and Rubinstein (CRR) [3] and from there analyze how it leads to the risk-neutral measure under which the asset price becomes a discounted martingale. Next, we describe the Black–Scholes [1] and Merton [2] (BSM) equations and argue that, using a suitable change of variable, the equation can be solved by using the Feynman–Kac formula relating to the heat equation in partial differential equations (PDEs). Thus, the risk-neutral measure becomes the fundamental solution of the transformed heat equation. Later, we generalize this to the multiasset scenario and models with (time varying) stochastic volatility; we then sum up with brief concluding remarks.

Binomial Model and the Risk-Neutral Measure

We first assume the following:

Assumptions

1. There are no market imperfections (no transaction costs, taxes, constraints on short sales; trading is continuous and frictionless).
2. There is unlimited borrowing and lending at rate r . Bond yield $dD = rD dt$.
3. No-arbitrage opportunity exists, i.e., no-arbitrage principle (*see* **Pricing of Life Insurance Liabilities; Premium Calculation and Insurance Pricing; Risk Measures and Economic Capital for (Re)insurers; Weather Derivatives**) is adopted.

Consider the one step binomial tree of CRR [3] with the following standard notations.

f : value of option, S : price of a stock, r : risk-free rate of interest, T : time elapsed, u : proportion of upward movement for the stock price and, d : proportion of downward movement for the same.

The activity of eliminating risk while taking a financial position is called *hedging* (*see* **Equity-Linked Life Insurance; Securitization/Life; Structured Products and Hybrid Securities; Statistical Arbitrage**). It is usually achieved by taking a simultaneous position whose risk pattern is as closely opposite to that of the original position as possible. When this is achieved perfectly, the corresponding hedge is called a *perfect hedge*, and the resultant financial portfolio is called a *riskless portfolio*. In real life, constructing such a riskless portfolio is possible only for a very short span of time.

In the context of the binomial tree, one uses portfolio weights $(-1, \Delta)$ for options and stocks. Here Δ is found, assuming that the portfolio constructed is riskless at the next unit of time, i.e., $\Delta S_u - f_u = \Delta S_d - f_d \Rightarrow \Delta = (f_u - f_d)/(S_u - S_d)$. Thus, this Δ hedging is the perfect hedge for the portfolio. Here, f_u and S_u are the options and stock prices corresponding to the up movement in one unit of time and, similarly, f_d and S_d are the corresponding prices for the down movement. To make the portfolio arbitrage-free, it should grow at the risk-free interest rate. Thus, one arrives at the next unit of time T , $e^{rT}(\Delta S_0 - f_0) = \Delta S_u - f_u$. Solving this equation

for f_0 , it is seen that

$$E(f_T) = pf_u + (1 - p)f_d = f_0 e^{rT} \quad (1)$$

and

$$E(S_T) = pS_0u + (1 - p)S_0d = S_0 e^{rT} \quad (2)$$

assuming $S_u = S_0u$ and $S_d = S_0d$, where $p = \frac{e^{rT} - d}{u - d}$. Thus, with this value of p , the expected return on the stock equals the risk-free rate.

However, this equation does not depend on the risk preference of the individuals. Thus, it appears as though we are in a world consisting of risk-neutral individuals. This is an example of a more general principle called *risk-neutral valuation* (see **Options and Guarantees in Life Insurance**) which assumes that we are operating in a “risk-neutral world”.

It is to be noted that this p is the probability of an up movement in a risk-neutral world. This may not be equal to the probability of an up movement in the real world.

Using risk-neutral valuation is convenient because we simply equate the expected return to the risk-free rate.

To generalize this model to n -steps with a finer grid of time points, say, in $\delta T = T/n$ observe, $S_{T+t} = S_t u^X d^{n-X}$, where X is the number of up moves in the market in $(t, T+t]$. Note that $ud = 1$ and $\log u = \sigma \sqrt{T/n}$ where σ is the volatility of the market. Thus $u^X d^{n-X} = e^{X\sigma \sqrt{T/n}} e^{-(n-X)\sigma \sqrt{T/n}} = e^{2X\sigma \sqrt{T/n}} e^{-\sigma \sqrt{nT}} = e^{2\sigma \sqrt{nT}(X/n - 1/2)}$. Assuming that the moves in n steps are independent and each move has two outcomes (up or down) with fixed probabilities, p for going up and $1 - p$ for going down, it follows that the distribution of X is Binomial (n, p) . Hence X follows binomial (n, p) distribution, where $p = [e^{rT/n} - e^{-\sigma \sqrt{T/n}}] / [e^{\sigma \sqrt{T/n}} - e^{-\sigma \sqrt{T/n}}] = r(T/n) + \sigma \sqrt{T/n} - (\sigma^2 T/2n) / [2\sigma \sqrt{T/n}]$ (ignoring higher powers of $(1/n)$ which is equivalent to $(1/2) + (r/2\sigma)\sqrt{T/n} - (\sigma/4)\sqrt{T/n}$. Note that, as n becomes large, X approaches the normal distribution. Thus, the mean of the normal distribution becomes

$$\begin{aligned} E \left[2\sigma \sqrt{nT} \left(\frac{X}{n} - \frac{1}{2} \right) \right] \\ = 2\sigma \sqrt{nT} \left[\left(\frac{r}{2\sigma} \right) \sqrt{\frac{T}{n}} - \left(\frac{\sigma}{4} \right) \sqrt{\frac{T}{n}} \right] \\ = \left(r - \frac{\sigma^2}{2} \right) T \end{aligned} \quad (3)$$

and the variance

$$\begin{aligned} \text{Var} \left[2\sigma \sqrt{nT} \left(\frac{X}{n} - \frac{1}{2} \right) \right] &= \frac{4\sigma^2 nT p(1-p)}{n} \\ &= 4\sigma^2 T \left(\frac{1}{4} - \left[\left(\frac{r}{2\sigma} \right) \sqrt{\frac{T}{n}} - \left(\frac{\sigma}{4} \right) \sqrt{\frac{T}{n}} \right]^2 \right) \\ &= \sigma^2 T - \left[\left(\frac{r}{2\sigma} \right) - \left(\frac{\sigma}{4} \right) \right]^2 \frac{4(\sigma T)^2}{n} \longrightarrow \sigma^2 T \end{aligned} \quad (4)$$

Therefore, as the subinterval length becomes finer and finer, i.e., as n tends to infinity, the asymptotic distribution of $2\sigma \sqrt{nT}(X/n - 1/2)$ becomes $N((r - \sigma^2/2)T, \sigma^2 T)$. Thus, the distribution of $\log(S_T) - \log(S_0)$ becomes $N((r - \sigma^2/2)T, \sigma^2 T)$. Since this argument can be carried out for any interval $(T_0, T]$, $T > T_0 \geq 0$, for $T_1 < \dots < T_k$, on the intervals $(T_i, T_{i+1}]$, it follows that that distribution of $\log(S_{T_{i+1}}) - \log(S_{T_i})$ is $N((r - \sigma^2/2)(T_{i+1} - T_i), \sigma^2(T_{i+1} - T_i))$ and for disjoint intervals they are independent. Now taking a linear transformation of these variables as $Y = AZ$, where components of Z are differences of log prices, and hence are independent and, for $i \geq 1$, Z_{i+1} has mean $(r - \sigma^2/2)(T_{i+1} - T_i)$ and variance $\sigma^2(T_{i+1} - T_i)$ with Z_1 having mean $(r - \sigma^2/2)T_1$ and variance $\sigma^2 T_1$, where

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 1 & 1 & 1 & \dots & 1 \end{bmatrix} \quad (5)$$

the joint distribution of $\{\log(S_{T_i}) : i = 1, \dots, k\}$ becomes k -variate normal with mean $(r - \sigma^2/2)T_i$ and variance $\sigma^2 T_i$, for $i = 1, \dots, k$ and the (i, j) th element of the correlation matrix $\sqrt{T_i/T_j}$ for $i \leq j$, as $\log(S_{T_j}) = \log(S_{T_i}) + [\log(S_{T_j}) - \log(S_{T_i})]$, and the first term is independent of the second term. Since the finite-dimensional distribution of $\log(S_T/S_0)$ agrees with that of Brownian motion, with means and variances as mentioned above, and both are continuous processes, one can now perceive S_T as distributionally equivalent to $S_0 \exp \{(r - \sigma^2/2)T + \sigma W_T\}$, which is the risk-neutral model, where W_T is the standard Brownian motion. Thus the price of an European option (see

Insurance Pricing/Nonlife), under the above assumption, can be found by finding the expectation of the payoff with respect to the risk-neutral measure of the S_T process, i.e., with respect to the measure of $S_0 \exp \{(r - \sigma^2/2)T + \sigma W_T\}$ process. (For an introductory exposition to binomial trees and option pricing, see Ross [4].)

Notice that the above argument and the use of Itô's lemma assert that, under risk-neutral measure, $\{S_t\}$ satisfies the following stochastic differential equation (SDE).

$$dS_t = rS_t dt + \sigma S_t dW_t \quad (6)$$

Had we started with the market growth (probability) model, i.e., $p_\mu = \frac{e^{\mu T} - d}{u - d}$, our argument would have led to an SDE,

$$dS_t = \mu S_t dt + \sigma S_t dW_t \quad (7)$$

which is assumed to be the market model of stock price in Black–Scholes.

Now, similar to the discrete time model, if the portfolio weights for options and stocks are $(-1, \Delta)$, then the value of the portfolio is $V = -f + \Delta S$. To keep the portfolio riskless at any time t , one should observe that $\partial V / \partial S = 0$. Thus, solving Δ , one gets $\Delta = \partial f / \partial S$, which is the continuous version of the discrete analog $\frac{f_u - f_d}{S_u - S_d}$. Then, at a time t , growth of the value should be at the rate of risk-free interest, $V_t = e^{rt} V_0$, i.e., $dV_t = rV_t dt$, to avoid arbitrage opportunity. Now, from the market model, $dV = -df + \frac{\partial f}{\partial S} dS$, using Itô's lemma on f , one finds $df = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial S} dS + \frac{1}{2} \frac{\partial^2 f}{\partial S^2} (dS)^2$. Thus, $-df + \frac{\partial f}{\partial S} dS = -\frac{\partial f}{\partial t} dt - \frac{1}{2} \frac{\partial^2 f}{\partial S^2} \sigma^2 S^2 dt$. Therefore, using the risk-free model of V , i.e., $dV_t = rV_t dt = r(-f + \frac{\partial f}{\partial S} S) dt$, one arrives at the Black–Scholes [1]–Merton [2] (BSM) differential equation

$$\frac{\partial f}{\partial t} + rS \frac{\partial f}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 f}{\partial S^2} = rf \quad (8)$$

Note that this equation does not involve any information on the risk attitude of the investors. The Black–Scholes–Merton (BSM) differential equation would depend on risk preference if it involved μ (expected rate of return on the stock). Thus, in this context, for all practical purposes we can assume that all investors are risk neutral. Thus the expected rate of return will simply be the rate r . Making a change of

variable in this equation as

$$\tau = T - t, z = a \log S + b\tau, \text{ and } f = e^{c\tau} g(z, \tau) \quad (9)$$

for a, b and c to be determined later, we get

$$\begin{aligned} \frac{\partial f}{\partial t} &= \frac{\partial e^{c\tau}}{\partial t} g + e^{c\tau} \frac{\partial g}{\partial \tau} \frac{\partial \tau}{\partial t} + e^{c\tau} \frac{\partial g}{\partial z} \frac{\partial z}{\partial t} \\ &= -ce^{c\tau} g - e^{c\tau} \frac{\partial g}{\partial \tau} - be^{c\tau} \frac{\partial g}{\partial z} \end{aligned} \quad (10)$$

$$\frac{\partial f}{\partial S} = e^{c\tau} \frac{\partial g}{\partial z} \frac{\partial z}{\partial S} = \frac{ae^{c\tau}}{S} \frac{\partial g}{\partial z} \quad (11)$$

$$\begin{aligned} \frac{\partial^2 f}{\partial S^2} &= \frac{\partial}{\partial S} \left(\frac{ae^{c\tau}}{S} \right) \frac{\partial g}{\partial z} + \frac{ae^{c\tau}}{S} \frac{\partial}{\partial S} \left(\frac{\partial g}{\partial z} \right) \\ &= -\frac{ae^{c\tau}}{S^2} \frac{\partial g}{\partial z} + \frac{a^2 e^{c\tau}}{S^2} \frac{\partial^2 g}{\partial z^2} \end{aligned} \quad (12)$$

Thus, the BSM equation becomes

$$\begin{aligned} 0 &= \frac{\partial f}{\partial t} + rS \frac{\partial f}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 f}{\partial S^2} - rf \\ &= -(r + c)e^{c\tau} g - e^{c\tau} \frac{\partial g}{\partial \tau} + \left(ar - b - \frac{1}{2} a \sigma^2 \right) \\ &\quad \times e^{c\tau} \frac{\partial g}{\partial z} + \frac{1}{2} \sigma^2 a^2 e^{c\tau} \frac{\partial^2 g}{\partial z^2} \end{aligned} \quad (13)$$

Now taking $c = -r, b = a(r - \frac{1}{2}\sigma^2)$ and $a = 1$ and dividing both side by, $e^{c\tau}$ we obtain

$$\frac{\partial g}{\partial \tau} = \frac{1}{2} \sigma^2 \frac{\partial^2 g}{\partial z^2} \quad (14)$$

Thus, having an initial data, $g(z, 0) = f(S, T) =$ payoff, say, one can solve the initial value problem of the heat equation, using the Feynman–Kac formula as

$$\begin{aligned} g(z, \tau) &= E(g(Z_\tau, \tau) | Z_0 = z) = E(\text{payoff}) \\ &= \int_{-\infty}^{\infty} \text{payoff}(y) p(z, \tau, y) dy \end{aligned} \quad (15)$$

where $p(z, \tau, y)$ is the fundamental solution of the equation and is given by $p(z, \tau, y) = \frac{1}{\sigma \sqrt{2\pi\tau}} \exp\{-(y - z)^2 / 2\sigma^2 \tau\}$ which is the density of the Brownian motion, starting at z with drift zero and diffusion coefficient σ^2 . Now, getting back to the original variable, one arrives at the solution that is described above with respect to the risk-neutral measure (for details, see Kallianpur and Karandikar [5]).

Though risk-neutral valuation is an artificial device for obtaining solutions to the BSM differential equation, the solutions are valid in all worlds. When we move to the real world, both the growth rate of a stock and the discount rate changes but fortunately, they do so in opposite directions and at exactly equal magnitudes to offset each other.

This principle can then be applied to pricing forward contracts or options with known dividend yields also.

Equivalent Martingale Measure (EMM)

In the last section, we discussed the asset pricing theories implicitly, by using equivalence of no-arbitrage principle and the existence of a measure under which the price process is a discounted martingale. Naturally, this leads us to the discussion of existence of an equivalent martingale measure (EMM) (*see Insurance Pricing/Nonlife; Mathematical Models of Credit Risk*) in a broader perspective. EMM is a probability distribution under which the price processes is adjusted to be consistent with no-arbitrage principle and attain risk neutrality. Harrison and Kreps [6] show that under the EMM, the stock price process is a martingale. Accordingly, the risk-neutral pricing method is also called the *martingale pricing technique*.

Consider derivatives dependent only on one variable θ : $\frac{d\theta}{\theta} = m dt + s dW$. Suppose, for simplicity, f_1 and f_2 are two derivatives $f_j = f_j(\theta, t)$, $j = 1, 2$.

$$\frac{df_j}{f_j} = \mu_j dt + \sigma_j dW, \quad \mu_j = \mu_j(\theta, t),$$

$$\sigma_j = \sigma_j(\theta, t), \quad j = 1, 2 \quad (16)$$

No-arbitrage principle and usual riskless portfolio construction implies

$$\frac{\mu_1 - r}{\sigma_1} = \frac{\mu_2 - r}{\sigma_2} = \lambda \quad (17)$$

where λ is called the *market price of risk*. Then the model becomes

$$\frac{df_j}{f_j} = (r + \lambda \sigma_j) dt + \sigma_j dW \quad (18)$$

This can be extended to several state variables $\theta_1, \theta_2, \dots, \theta_n$ as follows: $\frac{d\theta_i}{\theta_i} = m_i dt + s_i dW_i$, $i = 1, \dots, n$, and $f_j = f_j(\theta_1, \theta_2, \dots, \theta_n)$: $\frac{df_j}{f_j} = \mu_j dt + \sum_{i=1}^n \sigma_{ij} dW_i$. We similarly have $\mu_j - r = \sum_{i=1}^n \lambda_i \sigma_{ij}$: extent of excess return required (see Hull [7]).

The following example illustrates the EMM result:

Consider two traded securities f and g and a derivative $\phi = \frac{f}{g}$. Suppose the market price of risk is $\lambda = \sigma_g$, volatility of g . Then

$$\frac{df}{f} = (r + \sigma_g \sigma_f) dt + \sigma_f dW \quad (19)$$

$$\frac{dg}{g} = (r + \sigma_g^2) dt + \sigma_g dW \quad (20)$$

Now, it can be shown that

$$d\phi = d\left(\frac{f}{g}\right) = (\sigma_f - \sigma_g) \frac{f}{g} dW \quad (21)$$

$$\text{or} \quad \frac{d\phi}{\phi} = (\sigma_f - \sigma_g) dW \quad (22)$$

which is a martingale (this can be posed as an application of Girsanov's theorem).

When there are no-arbitrage opportunities, $\phi = \frac{f}{g}$ is a martingale for any security f , for some choice of market price of risk that is given by the volatility of the numeraire security g . This world is referred to as forward risk neutral with respect to g .

Models of Changing Volatility

Empirical studies show that GBM is not an appropriate model for some financial instruments: skewness, excess kurtosis, serial correlation, and time-varying volatilities are some of the observed features of data that do not conform with the properties of GBM.

Among the solutions proposed, time-varying volatility models have received the most attention as σ plays such a crucial role in the BSM formulation.

If $\sigma = \sigma(t)$, a known function of time, then in the BSM formula, we can replace σ by $\int_t^T \sigma_s ds$. But if σ is stochastic, e.g., given by the model

$$\frac{d\sigma}{\sigma} = \mu dt + \sigma dW_p \quad (23)$$

and

$$d\sigma = \alpha(\sigma) dt + \beta(\sigma) dW_\sigma \quad (24)$$

with $dW_p dW_\sigma = \rho dt$ (instantaneous correlation) and α and β are specified functions, then such models are called *stochastic volatility models*. These models represent an incomplete market, where pricing depends on risk premium of the volatility. Even so, for some

models, a pricing formula may be derived (see, for instance, Heston [8]). Derivation of the pricing rule is not always feasible, as formation of a replicating portfolio (see **Solvency**) is not, in general, possible. If $\rho = 0$ then, this can be done as in the following model due to Hull and White [9]:

$$\frac{dp}{p} = \mu dt + \sigma dW_p \quad (25)$$

and

$$\frac{d\sigma^2}{\sigma^2} = \alpha dt + \xi dW_\sigma \quad (26)$$

With these models, one can price the derivative using the EMM technique, with volatility as a non-traded security.

Direct estimation in continuous time is very difficult but a discrete approximation by Nelson [10, 11], which is inspired by pioneering works in the literature on ARCH (see **Extreme Value Theory in Finance**), GARCH, and EGARCH formulation by Engle [12] and later by Bollerslev [13], Bollerslev *et al.* [14], and others, works well in practice. Therefore, the other possibility is to start with a discrete-time model for the dynamics of the asset price, governed by an ARCH model, and then, to derive the corresponding formulae. This is, naturally, easier to implement. However, some issues like nonlinearity and state dependence are very hard to handle theoretically in the discrete setting, although they are easier in the continuous-time framework. With the advent of powerful personal computers, numerical methods with various sophisticated discretization techniques using binomial/trinomial trees, ARCH/GARCH models are not beyond the realm of achievement any more [15].

Conclusion

The concept of the risk-neutral measure is a very powerful and important tool in financial modeling. The basis of the approach lies in the *Feynman–Kac heat PDE*, and it has now become a commonly used term in financial markets thanks to the seminal work of Black and Scholes [1, 2]. With the advent of powerful computers, this has become an even more important

tool, nowadays. We have given a brief description of the model, its variants, and applications. A detailed study would be voluminous, and the interested reader is referred to the literature cited above.

References

- [1] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**(3), 637–654.
- [2] Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.
- [3] Cox, J.C., Ross, S.A. & Rubinstein, M. (1979). Option pricing: a simplified approach, *Journal of Financial Economics* **7**, 229–263.
- [4] Ross, S.M. (1999). *An Introduction to Mathematical Finance: Options and Other Topics*, Cambridge University Press, Cambridge.
- [5] Kallianpur, G. & Karandikar, R.L. (2000). *Introduction to Option Pricing Theory*, Birkhäuser, Boston.
- [6] Harrison, M. & Kreps, D. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- [7] Hull, J.C. (2000). *Options, Futures, and Other Derivatives*, 4th Edition, Prentice Hall International, London.
- [8] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *The Review of Financial Studies* **6**, 327–343.
- [9] Hull, J.C. & White, A. (1987). The pricing of options on assets with stochastic volatilities, *The Journal of Finance* **42**(2), 281–300.
- [10] Nelson, D.B. (1990). ARCH models as diffusion approximations, *Journal of Econometrics* **45**(1–2), 7–38.
- [11] Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: a new approach, *Econometrica* **59**(2), 347–370.
- [12] Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica* **50**(4), 987–1007.
- [13] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**(3), 307–327.
- [14] Bollerslev, T., Engle, R.F. & Nelson, D.B. (1994). Arch models, *Handbook of Econometrics*, North-Holland, Amsterdam, Vol. IV, pp. 2959–3038.
- [15] Duffie, D. (1996). *Dynamic Asset Pricing Theory*, 2nd Edition, Princeton University Press, Princeton.

GOPAL K. BASAK AND DIGANTA MUKHERJEE

Role of Alternative Assets in Portfolio Construction

Motivation

In this paper, we discuss the role of alternative investments within the context of asset allocation for long-term investors. We define alternative assets as the most popular private securities/contracts – hedge funds (*see* **Asset–Liability Management for Life Insurers; Alternative Risk Transfer**); managed futures (commodities, currency, and fixed income); and private equity (venture capital and leveraged buy-outs). These asset categories cover a wide assortment of investment strategies [1–4]. For example, there are over a dozen subcategories of hedge funds. Unfortunately, it has been difficult to measure the annual temporal performance of private equity for portfolio models; we do not focus on these securities. The recommendations apply to fully integrated risk-management systems (i.e., asset and liability management) with suitable extensions [5].

Alternative asset categories have become increasingly popular with institutional and wealthy individual investors since the recession in 2000–2001. The trend has been caused by several interrelated factors, including the superior performance achieved by leading university endowments over the past decade, and the need to recover lost surpluses by pension trusts (among others). Top university endowments in the United States (e.g., Yale, Harvard, Princeton, and Stanford Universities) and other leading institutional investors have achieved 15 to over 20% annual returns over the past decade by shifting a large proportion of their capital to private investments. In contrast, especially since 2000, numerous pension trusts have fallen behind, with many funding ratios dropping to the 75–80% range [6].

A major benefit of the alternative investments involves the generation of return patterns that differ from factors that affect equity and bond markets. In particular, stocks and bonds are largely driven by three generic factors: (a) government (default free) interest rates, (b) corporate earnings as a proxy for the level of economic activities, and (c) a risk premium [7, 8]. Thus, an investor's diversification is limited because of the dependence on a relatively

small number of underlying driving factors. Diversification becomes much less during periods of economic instability and contagion owing to an increase in risk premium.

A second potential benefit of alternative assets, especially for private markets such as venture capital, is the ability to increase leverage while smoothing price variations over several years. By their structure, some private market securities, e.g. early stage ventures, are not subject to the fluctuations of liquid market-based instruments. Owing to lack of reporting on reliable returns on a regular basis, there is difficulty in analyzing these asset categories within an optimal portfolio model. Future research should be aimed at this domain (see section titled “Summary and Future Directions”). Accordingly, we focus on nontraditional assets possessing marketable securities in this paper.

For simplicity, we discuss the role of alternative investments for asset-only allocation models. To properly address an investor's circumstance, we advocate a comprehensive asset and liability model (ALM) (*see* **Actuary; Risk Measures and Economic Capital for (Re)insurers**) such as, among others, described in Consigli and Dempster [9, 10], Mulvey *et al.* [11], and Ziemba and Mulvey [5].

Multiperiod Portfolio Models

This section provides a brief explanation on benefits of adopting multiperiod models, especially fixed-mix policy rules, for portfolio construction. There are distinct advantages (*see* **Credit Migration Matrices; Asset–Liability Management for Nonlife Insurers**) of a multiperiod horizon as compared with a static buy-and-hold framework [12, 13]. First, the multiperiod model can address a number of significant real-world issues, such as transaction costs (e.g. taxes) and changing economic environments (growth *versus* recession) with nonconstant correlation and covariance matrices. For instance, stock and bond returns are generally positively related under normal economic conditions, whereas these returns can become negatively related during and after a recession. A multiperiod portfolio model can show the impact of these changing conditions on the investor's future wealth in an integrated risk fashion. Also, the performance of a multiperiod model can be greater than the performance of a buy-and-hold model for

comparable planning horizons because of the gains attained by rebalancing the portfolio (*see Equity-Linked Life Insurance; Informational Value of Corporate Issuer Credit Ratings*) at selected time junctures. References [13–16] discuss the nature of the rebalancing gains.

There is a major drawback, however, to implement a multiperiod model: the model can be non-convex, which makes it difficult to attain the optimal strategy. Instead of complicated optimization techniques, we provide a simple, yet efficient, approach – fixed-mix policy rule – to illustrate the benefits of multiperiod horizon. The fixed-mix strategy always applies the same weights at the beginning of the time period to constituents, in contrast to the buy-and-hold approach, where the weights vary as the prices of constituents change over time. Also, the fixed-mix strategy can serve as a benchmark for other dynamic strategies.

Early on, Samuelson [17] and Merton [18] showed that the fixed-mix investment rule is optimal under certain restrictive assumptions. Mulvey *et al.* [19], among others, presents a clear illustration for the connection between the fixed-mix rule and rebalancing gain. For simplicity, let us assume that there is one stock and one risk-free asset. Suppose the stock price process P_t follows a geometric Brownian motion that can be represented by the equation

$$dP_t = \alpha P_t dt + \sigma P_t dz_t \quad (1)$$

where α is drift, σ is volatility, and z_t is a Brownian motion with mean zero and variance t . Similarly, risk-free asset B_t follows the same process with drift equal to zero and volatility equal to zero. Then the stochastic differential equation (*see Numerical Schemes for Stochastic Differential Equation Models*) for B_t can be written as

$$dB_t = rB_t dt \quad (2)$$

Now, assume that we invest η in the stock and $(1 - \eta)$ in the risk-free asset with the fixed-mix policy rule. Then the wealth process of the portfolio W_t can be expressed as

$$\frac{dW_t}{W_t} = \frac{\eta dP_t}{P_t} + \frac{(1 - \eta) dB_t}{B_t} \quad (3)$$

After substituting for P_t and B_t in the equation, one can show that the growth rate of the portfolio is

$$\gamma_w = \eta\alpha + (1 - \eta)r - \frac{\eta^2\sigma^2}{2} \quad (4)$$

For simplicity, we assume that the growth rate of the stock and the risk-free asset are equal.^a Then the growth rate of the portfolio can be rewritten as

$$\gamma_w = r + \frac{(\eta - \eta^2)\sigma^2}{2} \quad (5)$$

If $0 < \eta < 1$, this quantity is greater than the growth rate of the buy-and-hold approach by $(\eta - \eta^2)\sigma^2/2$. The quantity corresponds to the rebalancing gain due to the application of the fixed-mix policy rule as compared with buy-and-hold.

Many investors have applied versions of the fixed-mix rules with practical successes [5, 11, 16, 20–22]. For example, the famous 60/40 norm (60% equity and 40% bonds) falls under this policy. Here, at each period, we rebalance the portfolio to 60% equity and 40% bond. Another good example is Standard & Profit 500 equal-weighted index (S&P EWI) by Rydex Investments. As opposed to traditional cap-weighted S&P index, stocks have the same weight (1/500) and the index is rebalanced semiannually to maintain the weights over time. To illustrate the benefits of applying the fixed-mix policy rule, during 1994–2005, S&P EWI achieved 2% excess return with only 0.6% extra volatility compared to S&P 500 index. Figure 1 illustrates log prices of S&P 500 and S&P EWI for last 4 years.

We close this section by discussing desirable properties of assets in order to achieve rebalancing gain. First, suppose two assets in the derivation above are perfectly correlated. Then, it can be easily shown that the rebalancing gain is zero. From this, it is evident that diversification among assets plays a major role to achieve an excess growth rate. This observation suggests that dynamic diversification is essential in order to produce extra gains *via* multiperiod approaches. Also, as always, diversification provides a source of reducing portfolio risk. Second, given a set of independent assets, the rebalancing gain $((\eta - \eta^2)\sigma^2/2)$ increases as the volatilities of assets increase. To benefit from rebalancing gain, the volatility of each asset should be reasonably high. In this context, the traditional Sharpe ratio (*see Axiomatic Models of Perceived Risk; Longevity Risk and Life Annuities*) might not be a good measure for individual asset in terms of multiperiod portfolio management, even though it remains valid at the portfolio level. Additionally, low transaction costs (fees, taxes, etc.) are desirable because applying the fixed-mix policy rule requires portfolio rebalancing.

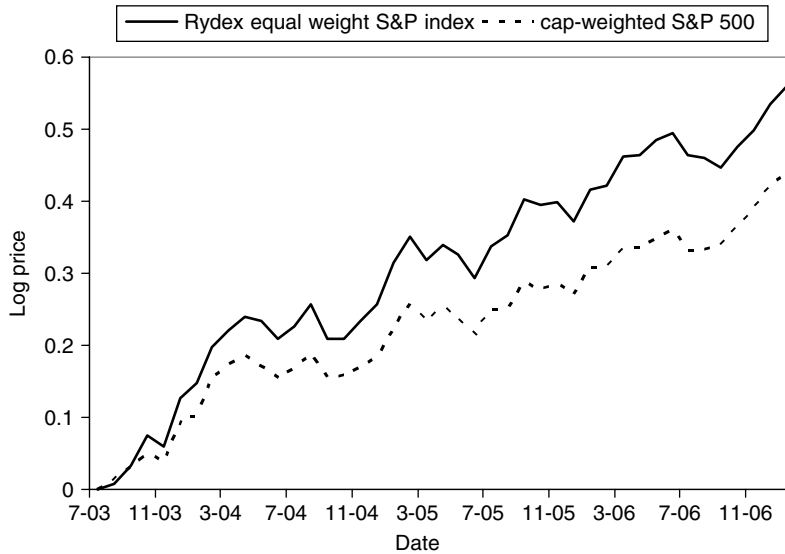


Figure 1 Log prices of S&P 500 index and S&P EWI during July 2003 to December 2006. This figure illustrates the log-price processes for S&P EWI and S&P 500 from July 2003 to December 2006. Each index is scaled to have a log price of zero at the beginning of the sample period. In term of the total return, S&P EWI outperformed S&P 500 index for last 4 years and this performance difference between the two assets can be interpreted as a rebalancing gain due to the fixed-mix policy rule

In summary, the properties of the best ingredients for the fixed-mix rule are (a) relatively good performance (positive expected return), (b) relatively low correlations among assets, (c) reasonably high volatility, and (d) low transaction costs. For more detailed discussion, see Fernholz [14] and Mulvey *et al.* [19].

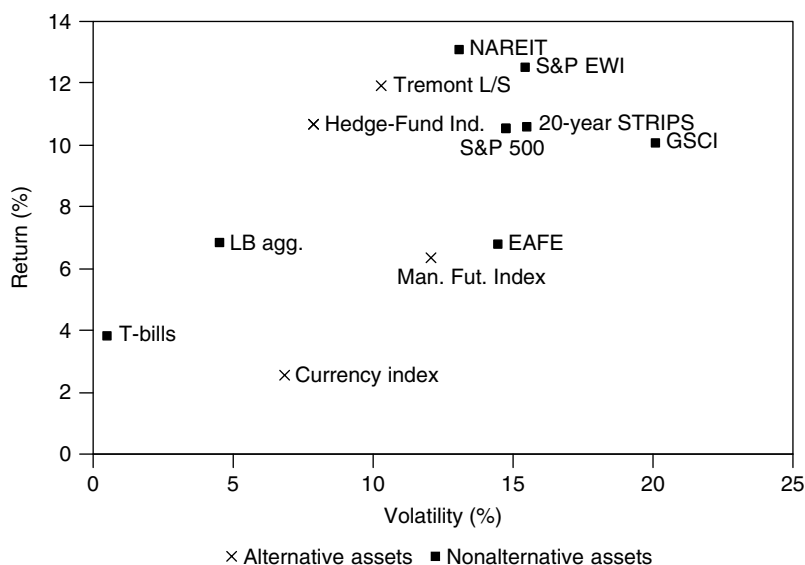
Historical Perspective

In this section, we describe the performance of several major categories of alternative investments along with traditional assets over a recent 12-year historical period. While there are obvious limitations of evaluating historical performance, there is benefit for observing the past patterns. Figure 2 and Table 1 depict historical performance of some of the major traditional asset categories and alternative assets including an aggregate hedge-fund index, managed futures index, a long/short equity fund index, and a currency-based index.

There are several general observations. First, the historical record is limited for most alternative assets. For instance, before 1994, the hedge-fund industry was quite small and the indices did not adjust for

survivor bias. In the future, because of the explosive growth in hedge funds, returns may decrease by reducing the attainable edge for certain strategies, such as statistical arbitrage (*see Statistical Arbitrage; Ethical Issues in Using Statistics, Statistical Methods, and Statistical Sources in Work Related to Homeland Security*). Given these limitations, the overall performance of the alternative assets was generally contained within the coverage of traditional equities, bonds, and real-estate investment trusts (REITs). The returns and risks are roughly compatible between traditional and alternative indices.

Several studies have conducted performance attribution of the return patterns of selected alternative assets, mostly hedge funds [2, 23, 24]. This research helps investors understand the process undertaken by their portfolio managers (at least at a high level). If an investor can find a pattern that is a reasonably consistent match, the portfolio manager could be compensated by comparison to this benchmark, perhaps, in addition to the usual absolute return benchmarks. Also, the development of economic scenario generating systems requires a linkage of economic factors to the returns of the asset categories (both traditional and alternative) [10].



Traditional asset classes		
Type	Name	Description
Equity	S&P 500	Standard and Poor 500 index: An unmanaged cap-weighted index of 500 domestic stocks.
	S&P EWI	Rydex S&P 500 equal weighted index: The fixed-mix version (equal weight) of S&P 500 index.
	EAFE	Morgan Stanley equity index for Europe, Australia, and the Far East: An unmanaged cap-weighted index of overseas stocks.
Bond	LB Agg.	Lehman long aggregate bond index: an unmanaged index of government & corporate bonds, mortgage-backed securities and asset-backed securities.
	Strips	20-year U.S. government zero-coupon bonds.
	T-bill	U.S. government 30-day Treasury bill
Real Estate	NAREIT	National association of real-estate investment trusts: an unmanaged index of U.S. real estates.
Commodity	GSCI	Goldman Sachs commodity index: a composite index of long-only commodity futures
Alternative asset classes		
Type	Name	Description
Hedge Fund	Hedge. Fund Ind.	Tremont hedge-fund aggregate index: an asset-weighted hedge-fund index which is net of fees and expenses.
	Man. Fut. Index	Tremont managed futures index: an asset-weighted hedge-fund index of investments in listed bond, currency, equity and commodity futures markets.
	Tremont L/S	Tremont long/short equity index: an asset-weight hedge-fund index of investments on both the long and short sides of equity markets.
Currency	Currency Index	Reuters-CRB Currencies Index: an index of five currency futures (BP, EC, CD, SF and JY).

Figure 2 Performance of alternative and nonalternative asset categories (1994–2005). Among 12 categories, four assets – Tremont hedge-fund aggregate index (Hedge-Fund Ind.), Tremont long/short equity index (Tremont L/S), currency index, and Tremont managed futures index (Man. Fut. Index) – are classified as alternative asset classes. Unlike traditional assets, each fund in these categories has a specific benchmark. That is, the money manager of a specific fund is asked to outperform the corresponding benchmark such as S&P 500 or Russell 1000, while constructing her portfolio similar to the benchmark. Thus, performance of such alternative assets is highly dependent on their underlying benchmark. Therefore, a direct comparison between traditional and alternative assets is not straightforward. However, since we focus on the alternative assets as a genuine source of diversification rather than superior performance, we illustrate the historical performance of such assets along with traditional ones to give a general idea to the readers. See end note (a) for the detailed explanation of each asset

Table 1 Summary statistics for historical performance of popular asset categories. In this table, investment performance of each asset category for whole sample period (top), the first 6-year (middle), and the second 6-year (bottom) is shown. Assets with relatively high maximum drawdown are highlighted. For detailed description of each asset, see the legend in Figure 1

	Annualized return (%)	Standard deviation (%)	Sharpe ratio	Maximum drawdown (%)	Return/ Drawdown
<i>Whole sample period (1994–2005)</i>					
S&P 500	10.5	14.8	0.45	44.7	0.24
LB agg.	6.8	4.5	0.66	5.3	1.29
EAFE	6.8	14.5	0.20	47.5	0.14
T-bills	3.8	0.5	0.00	0.0	N/A
NAREIT	13.1	13.1	0.71	26.3	0.50
GSCI	10.1	20.1	0.31	48.3	0.21
Hedge-Fund Ind.	10.7	7.9	0.87	13.8	0.77
Man. Fut. Index	6.4	12.1	0.21	17.7	0.36
Currency index	2.6	6.8	−0.18	28.7	0.09
Tremont L/S	11.9	10.3	0.78	15.0	0.79
S&P EWI	12.5	15.4	0.56	30.3	0.41
20-year STRIPS	10.6	15.5	0.43	22.8	0.46
<i>First subperiod (1994–1999)</i>					
S&P 500	23.5	13.6	1.37	15.4	1.53
LB agg.	5.9	4.0	0.24	5.2	1.14
EAFE	12.3	13.8	0.54	15.0	0.82
T-bills	4.9	0.2	0.00	0.0	N/A
NAREIT	6.5	12.0	0.13	26.3	0.25
GSCI	4.7	17.4	−0.01	48.3	0.10
Hedge-Fund Ind.	14.1	9.9	0.93	13.8	1.02
Man. Fut. Index	5.5	11.5	0.05	17.7	0.31
Currency index	0.1	6.7	−0.73	20.4	0.00
Tremont L/S	18.5	11.6	1.18	11.4	1.62
S&P EWI	17.1	13.7	0.89	19.9	0.86
20-year STRIPS	7.2	14.7	0.16	22.8	0.32
<i>Second subperiod (2000–2005)</i>					
S&P 500	−1.1	15.2	−0.25	44.7	−0.03
LB Agg.	7.7	5.1	0.99	5.3	1.47
EAFE	1.5	15.1	−0.08	47.5	0.03
T-bills	2.7	0.5	0.00	0.0	N/A
NAREIT	20.0	13.9	1.24	15.3	1.31
GSCI	15.7	22.5	0.57	35.4	0.44
Hedge Fund Ind	7.4	5.1	0.92	7.7	0.96
Man. Fut. Index	7.2	12.7	0.35	13.9	0.52
Currency Index	5.1	7.0	0.34	15.3	0.33
Tremont L/S	5.6	8.6	0.34	15.0	0.37
S&P EWI	8.1	17.0	0.31	30.3	0.27
20-year STRIPS	14.0	16.3	0.69	19.1	0.73

Next, we briefly review the historical performance of the asset categories over two distinct subperiods. We designate the first period January 1, 1994 to December 31, 1999 as “high equity”, whereas the second period January 1, 2000 to December

31, 2005 indicates “low equity”. Over the entire 12-year period, the annual returns for the asset categories range from low = 2.6% (for currencies) to high = 13.1% (for REITs). Many assets display disparate behavior over the two 6-year subperiods:

the Goldman Sachs commodity index (GSCI) and REITs had their worst showing during the first subperiod – the lowest returns and highest drawdown values, whereas EAFE and S&P 500 display the opposite results – high returns in the first subperiod. As a general observation, investors should be ready to encounter sharp drops in individual asset categories. Drawdown for half of the categories lies in the range 26–48% (Table 1).

Two of the highest historical return-to-risk ratios occurred in the hedge-fund categories: (a) the Tremont aggregate hedge-fund index (0.87) and (b) the Tremont long/short index (0.78). In both cases, returns are greater than the S&P 500 index with much lower volatility. As mentioned, this performance has led to increasing interest in hedge funds. Many experts believe that the median future returns for hedge funds are likely to be lower than historical values – due in part to the large number of managers entering the domain. In fact, low volatility may be a detriment for increasing overall portfolio performance since it limits the rebalancing gains. There are advantages in combining assets with modest return-to-risk ratios and reasonable returns in a rebalanced portfolio, when the lower ratio is caused by higher volatility.

In summary, alternative assets have displayed solid performance over the 12-year period, 1994–2005, especially the Tremont aggregate hedge-fund and long/short indices. In both cases, however, the returns in the second period, while remaining positive, fell substantially, partially because of the lower returns of equities. In contrast, the currency index and managed futures showed the opposite relationship – higher returns in the second period. The latter assets showed countercyclical behavior as compared with equities. As mentioned, there is some concern that returns will drop further with the recent expansion of the alternative investment universe. Even in this environment, alternative investments can provide benefits to the investor as a novel source of diversification, as we will see in the next section.

The Role of Alternative Assets in the Portfolio Management

As mentioned, there is evidence that private markets can generate superior returns as compared

with many public markets [25]. Unfortunately, for most investors, top opportunities are rarely available without special access privileges. These accessibility issues are slowly receding with the recent introduction of tradable hedge funds and related instruments (such as active exchange traded funds), which allow individual investors to gain a portion of the median hedge-fund returns.

Importantly, alternative investments can provide the benefits of wide diversification and leverage to achieve superior performance. In this section, we are less concerned with superior performance; we use the alternative assets to provide additional sources of diversification – above and beyond that dictated by equities and bonds.

The most comprehensive approach for evaluation of alternative assets in a portfolio is to apply an integrated risk-management system on a set of investment vehicles, which includes alternative asset classes. However, such an approach is beyond the scope of this article. Thus, rather than conducting an ALM optimization, we apply the fixed-mix rule to the assets mentioned in the previous section. More specifically, the analysis will construct three portfolios: (P1) a buy-and-hold portfolio of only traditional assets, (P2) a buy-and-hold portfolio of traditional and alternative assets, and (P3) a fixed-mix portfolio of both traditional and alternative assets. In this regard, we use the fixed-mix rule at two levels. First at the stock selection level, we substitute an equal-weighted S&P 500 index for the capital-weighted S&P 500 fund. The equal-weighted index has generated better performance over the standard S&P 500 index [21], as would be expected owing to the additional returns gotten from rebalancing the mix. Then, the portfolio is rebalanced monthly to fulfill the fixed-mix policy rule at the asset selection level. For simplicity, assets are weighted equally for all three portfolios. Table 2 summarizes these strategies.

We first compare P1 and P2 to illustrate the diversification benefits from the alternative asset categories. The two leftmost columns of Table 3 show the resulting performance. Here, the historical performance of P1 and P2 is 9.9 and 9.4% per year, with annualized volatility equal to 7.9 and 6.6%, respectively. As expected, alternative assets serve as a novel source of diversification, resulting higher risk-reward ratios for P2. The benefit of including alternative assets becomes even greater when the fixed-mix rule is used. Among three portfolios,

Table 2 Portfolio description

Portfolio	Description	Constituents
P1	Traditional assets only (buy-and-hold)	Traditional assets: S&P 500, LB bond, EAFE, NAREIT, GSCI, STRIPS
P2	With alternative assets (buy-and-hold)	Traditional assets: S&P 500, LB bond, EAFE, NAREIT, GSCI, STRIPS Alternative assets: man futures, hedge-fund index, L/S index, currency
P3	With alternative assets (fixed mix)	Traditional assets: S&P EWI, LB bond, EAFE, NAREIT, GSCI, STRIPS Alternative assets: man futures, hedge-fund index, L/S index, currency

P3 shows the best performance in most of performance measures. Clearly, wide diversification pays off in terms of reducing the portfolio's overall risk – volatility and maximal drawdown. The maximum drawdown for P3 is mere 6.4%, which is almost half of that for P1. Also, improvements in return–risk ratios are significant, especially the return–drawdown ratio (from 0.89 to 1.54). It is also worth noting that the performances of P3 in each of the two sub-periods (1994–1999 and 2000–2005) are relatively similar, which implies that it provides more reliable outcomes.

Next, we take three portfolios and apply selected degrees of leverage at several values: 20, 50, and 100%. Leverage is achieved in the conventional way by borrowing money at the T-bill rate and putting it accordingly on the constituents. Because three portfolios in consideration do not include T-bill, the relative weights do not change as the portfolio is levered. The returns increase for each juncture with increasing risks (as measured by volatility and drawdown). However, for the fixed-mix portfolio with alternative assets (P3), the overall risks are quite reasonable even at the 100% leverage – 12.4% annualized volatility and 14.4% drawdown – resulting in better risk–reward ratio than the other portfolios. Interestingly, P3 outperforms P1 with no leverage in terms of return–risk ratios, even at 100% leverage. Efficient frontiers in Figure 3 clearly illustrate this point. See Mulvey *et al.* [19] for further improvements via overlay strategies.

The historical results suggest that investors can benefit by including alternative assets in their portfolio. First, an investor with access to the top deals can achieve truly superior performance – for example,

Renaissance Technologies' annual return of over 35% after fees since 1989. Similarly, the leading U.S. university endowments have shown that private investments can be highly profitable. But also significantly, alternative assets offer the benefits of combining wide diversification and targeted leverage. These advantages are more readily available for most investors than gaining access to the top private investments.

There are two qualifiers for this empirical study: (a) the historical performance of alternative investments may not correspond to future performance owing to, among others, the increase in the number of hedge funds existing today, and (b) it can be difficult to rebalance a portfolio because of restriction on the entry and exit of capital within many of the private markets. Accordingly, the empirical results should be treated as an illustration of possible benefits. This issue is expected to be partially resolved in near future owing to emergence of new financial instruments such as active exchange traded funds. The main message remains – alternative investments can provide increasing diversification benefits because of the uniqueness of the return patterns.

Summary and Future Directions

The top alternative investments have delivered superior performance over the past 10–15 years, as shown by the returns of leading university endowments and the consistently high returns of selected hedge funds. Unfortunately, most investors are unable to gain access to these opportunities at this time.

The report suggests that, with careful risk management, investor performance can be improved by adding alternative assets to a portfolio of traditional

Table 3 Historical results with different leverage values applied to portfolios. In this table, historical investment performance of three portfolios is illustrated for the whole sample period (1994–2005, top row), the first 6-year subperiod (1994–1999, middle row), and the second subperiod (2000–2005, bottom row). Performance of P1, P2, and P3 are shown in the left, middle, and right column, respectively. As anticipated, P3 outperforms both P1 and P2, which depicts the benefits of including alternative assets as well as adopting multi-period models. Performance improvements are most significant in the return–drawdown ratio. Also, the fixed-mix portfolio with alternative assets (P3) shows the best return–risk ratios as it gets levered up

	P1: traditional assets only (buy-and-hold)				P2: with alternative assets (buy-and-hold)				P3: with alternative assets (fixed mix)				
	0%	20%	50%	100%	0%	20%	50%	100%	0%	20%	50%	100%	
Leverage													
1994–2005	Return	9.9%	11.0%	12.8%	15.5%	9.4%	10.4%	12.1%	14.7%	9.8%	10.9%	12.7%	15.6%
	Volatility	7.9%	9.5%	11.9%	15.9%	6.6%	7.9%	9.9%	13.2%	6.2%	7.4%	9.3%	12.4%
	Sharpe ratio	0.76	0.76	0.75	0.74	0.84	0.84	0.83	0.82	0.96	0.96	0.95	0.95
	Drawdown (DD)	11.1%	14.1%	18.6%	25.7%	7.1%	9.3%	12.7%	18.3%	6.4%	8.0%	10.4%	14.4%
	Return/DD	0.89	0.78	0.69	0.61	1.31	1.12	0.95	0.80	1.54	1.37	1.22	1.08
1994–1999	Return	11.1%	12.3%	14.0%	16.9%	10.9%	12.1%	13.9%	16.7%	9.6%	10.5%	11.9%	14.1%
	Volatility	7.9%	9.5%	11.8%	15.7%	6.8%	8.1%	10.1%	13.5%	6.3%	7.5%	9.4%	12.5%
	Sharpe ratio	1.40	1.30	1.19	1.07	1.61	1.49	1.37	1.24	1.53	1.40	1.26	1.13
	Drawdown (DD)	9.1%	11.0%	14.0%	19.1%	7.1%	8.7%	11.0%	14.9%	6.4%	8.0%	10.4%	14.4%
	Return/DD	1.22	1.11	1.00	0.88	1.53	1.39	1.25	1.12	1.51	1.32	1.14	0.98
2000–2005	Return	10.9%	12.5%	15.0%	19.0%	9.2%	10.5%	12.4%	15.6%	9.9%	11.3%	13.5%	17.1%
	Volatility	7.9%	9.5%	11.9%	15.8%	6.5%	7.8%	9.7%	13.0%	6.2%	7.4%	9.3%	12.4%
	Sharpe ratio	1.38	1.32	1.26	1.20	1.42	1.35	1.27	1.20	1.61	1.53	1.45	1.38
	Drawdown (DD)	7.3%	9.4%	12.6%	17.8%	5.3%	7.0%	9.6%	13.9%	4.7%	6.3%	8.7%	12.6%
	Return/DD	1.50	1.33	1.19	1.07	1.75	1.49	1.29	1.12	2.10	1.81	1.56	1.35

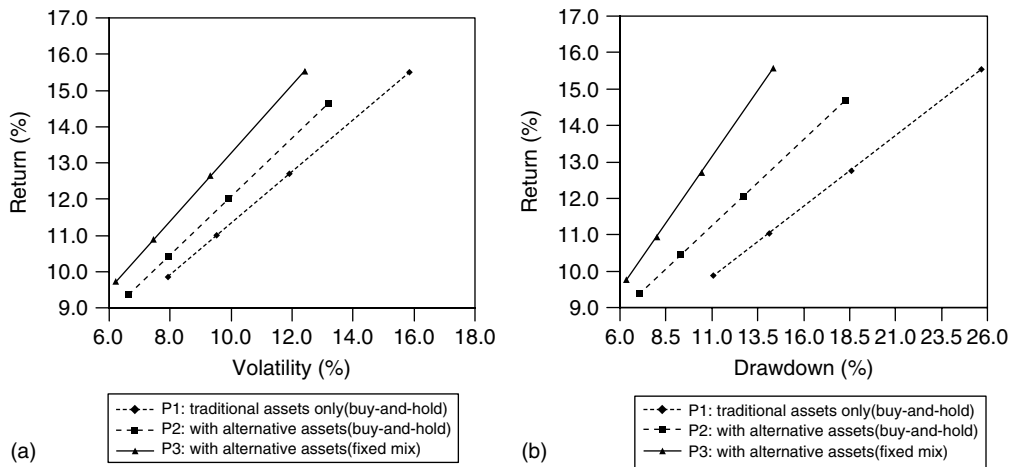


Figure 3 Efficient frontiers (see *Asset–Liability Management for Nonlife Insurers*) of the portfolios with/without alternative assets. Figure (a) illustrates efficient frontiers in volatility–return plane, while (b) is drawn in maximum drawdown–return plane. The efficient frontier of P3 contains those of P1 and P2 in both cases, which clearly exhibit the role of alternative assets in portfolio construction

asset categories. Alternative assets can provide reasonable performance with less dependency on the usual economic factors such as corporate earnings, interest rates, and risk premium. The novel return patterns provide a substantial benefit for increasing diversification. For long-term investors, wide diversification can be coupled with target leverage to increase portfolio performance. Rebalancing gains are also available for selected investors. As always, investors should carefully analyze their potential risks and rewards in an integrated, anticipatory fashion.

What are directions for future research? First, we can continue to search for assets with novel sources of returns (as compared with stocks, bonds, and money market securities). A prime example involves weather-related products. Ideally, the emerging securities would develop in liquid markets so that investor has valid market prices and can achieve rebalancing gains.

In addition, research can be aimed at improving the modeling/pricing of private securities. Current approaches, such as the internal rate of returns for seasoned (vintage year) ventures, are not so helpful for the problem of seeking an optimal asset allocation. Approaches developed for asset allocation (and integrated risk management) under traditional categories will need to be extended for the inclusion of their privately held investments/securities.

Undoubtedly, long-term, multiperiod financial planning models for individual investors will continue to grow in popularity. The aging population of wealthy individuals will require assistance as they approach retirement and also for estate planning purposes. The U.S. government has recently passed legislation that makes it easier for financial organizations to provide probabilistic investment advice. This change in regulation has already led to implementation of a number of stochastic planning systems (similar to the ones discussed in this paper). Individual and institutional investors alike can benefit by applying integrated risk-management systems in conjunction with a full set of traditional and alternative asset categories.

End Notes

- ^a. This assumption is *not* required to illustrate the rebalancing gain, but it makes the illustration simpler and easier to understand.

References

- [1] Anson, M.J.P. (2006). *Handbook of Alternative Assets*, 2nd Edition, John Wiley & Sons, Hoboken.
- [2] Feng, W. & Hsieh, D.A. (1999). A primer on hedge funds, *Journal of Empirical Finance* 6, 309–311.

- [3] Jaeger, R.A. (2003). *All About Hedge Funds*, McGraw-Hill.
- [4] Lhabitant, F.-S. (2002). *Hedge Funds, Myths and Limits*, John Wiley & Sons.
- [5] Ziemba, W. & Mulvey, J. (eds) (1998). *World Wide Asset and Liability Modeling*, Cambridge University Press, Cambridge.
- [6] Mulvey, J.M., Simsek, K.D., Zhang, Z., Fabozzi, F. & Pauling, W. (2006). Assisting defined-benefit pension plans, *Operations Research* (to appear).
- [7] Bakshi, G. & Chen, Z. (2005). Stock valuation in dynamic economies, *Journal of Financial Markets* **8**(2), 111–151.
- [8] Chen, Z. & Dong, M. (2001). *Stock Valuation and Investment Strategies*, Yale ICF Working Paper No. 00–46.
- [9] Consigli, G. & Dempster, M.A.H. (1998). Dynamic stochastic programming for asset-liability management, *Annals of Operations Research* **81**, 131–161.
- [10] Consigli, G. & Dempster, M.A.H. (1998). The CALM stochastic programming model for dynamic asset-liability management, in *World Wide Asset and Liability Modeling*, W. Ziemba & J. Mulvey, eds, Cambridge University Press, Cambridge, pp. 464–500.
- [11] Mulvey, J.M., Gould, G. & Morgan, C. (2000). An asset and liability management system for Towers Perrin-Tillinghast, *Interfaces* **30**, 96–114.
- [12] Mulvey, J.M., Pauling, B. & Madey, R.E. (2003a). Advantages of multiperiod portfolio models, *Journal of Portfolio Management* **29**, 35–45.
- [13] Luenberger, D. (1997). *Investment Science*, Oxford University Press, New York.
- [14] Fernholz, R. (2002). *Stochastic Portfolio Theory*, Springer-Verlag, New York.
- [15] Fernholz, R. & Shay, B. (1982). Stochastic portfolio theory and stock market equilibrium, *Journal of Finance* **37**, 615–624.
- [16] Mulvey, J.M., Kaul, S.S.N. & Simsek, K.D. (2004). Evaluating a trend-following commodity index for multi-period asset allocation, *Journal of Alternative Investments* **7**(1), 54–69.
- [17] Samuelson, P.A. (1969). Lifetime portfolio selection by dynamic stochastic programming, *Review of Economics Statistics* **51**, 239–246.
- [18] Merton, R.C. (1969). Lifetime portfolio selection under uncertainty: the continuous-time case, *Review of Economics Statistics* **51**, 247–257.
- [19] Mulvey, J.M., Ural, C. & Zhang, Z. (2007). Improving performance for long-term investors: wide diversification, leverage, and overlay strategies, *Quantitative Finance* **7**(2), 175–187.
- [20] Mulvey, J.M. & Thorlacius, A.E. (1998). *The Towers Perrin Global Capital Market Scenario Generation System: CAP Link. World Wide Asset and Liability Modeling*, W. Ziemba & J. Mulvey, eds, Cambridge University Press, Cambridge.
- [21] Mulvey, J.M. (2005). *Essential Portfolio Theory*, A Rydex Investment White Paper (also Princeton University Report).
- [22] Perold, A.F. & Sharpe, W.F. (1988). Dynamic strategies for asset allocation, *Financial Analysts Journal* **1/2**, 16–27.
- [23] Feng, W. & Hsieh, D.A. (2000). Performance characteristics of hedge funds, and commodity funds: natural versus spurious biases, *Journal of Financial and Quantitative Analysis* **35**, 291–307.
- [24] Feng, W. & Hsieh, D.A. (2002). Benchmarks of hedge funds performance: information contents and measurement biases, *Financial Analyst Journal* **58**, 22–34.
- [25] Swensen, D.F. (2000). *Pioneering Portfolio Management*, The Free Press, New York.

JOHN M. MULVEY AND WOO CHANG KIM

Role of Risk Communication in a Comprehensive Risk Management Approach

The ability of humankind to intuitively assess and manage risks has been fundamental for human survival and evolution. Those who were adept at recognizing risk and learning from danger survived to reproduce, whereas those who could not inevitably perished from avoidable environmental hazards [1]. However, the risks of concern to people have changed dramatically in the last 30–40 years. While people were previously concerned about major risks with visible and often drastic consequences (such as polio and physical hazards associated with various types of employment like mining), resolution of these high magnitude hazards resulted in a shift of concern to risks that were less visible, more poorly understood, and inherently more frightening because of their unknown nature (such as pesticides). This gave rise to a more “formal” practice of quantitative risk assessment in the early 1980s [2] (*see Environmental Health Risk Assessment; Environmental Hazard; Risk and the Media*).

The assessment and subsequent management of these risks has burgeoned into a distinct field of scientific study and public policy. In recent years, the effective management of risks to human health posed by potentially hazardous environmental contaminants or situations has commanded more public attention and reaction as risk issues and the related scientific information have become more complex and multifaceted (*see Environmental Hazard*). As various agencies and organizations struggled to deal with a public that was increasingly apprehensive and distrustful about these risks, the critical role of effective communication became apparent. *Risk communication* was born as a discrete field of study and practice.

What is Risk Communication?

Risk communication is a marriage of communication theory and risk studies [3]. Although rooted in these disciplines, the field of risk communication has evolved somewhat independently as a distinct area

of scientific inquiry and practice. In the last 25 years researchers and practitioners in a wide range of disciplines have developed theoretical frameworks and models to increase the effectiveness of communications about risk [3–9]. They have also sought to ease the growing tensions between the “experts” and “public” by exploring and conceptualizing the factors dealing with various reactions to and explanations of risk [10, 11].

Risk communication is now acknowledged to be “an interactive process of exchange of information and opinion among individuals”. Successful communication is defined as that which “raises the level of understanding of relevant issues or actions for those involved and satisfies them that they are adequately informed within the limits of available knowledge” [3, p. 2]. However, risk communication has not always been defined in terms of these enlightened reciprocal and inclusive concepts.

Evolution of Risk Communication

The early approaches to risk assessment and risk management generally considered communication to be the final stage of risk management, after the technical risk assessment and selection of risk-management options had been completed (Figure 1). Risk communication was originally thought of as consisting of a one-way message from experts and/or agencies to the public of technical and scientific information, with the primary purpose of persuading people to accept a decision that had already been made (Figure 2). This approach proved to be overwhelmingly unsuccessful, frequently resulting in conflict and frustration on all sides.

The turning point in the approach to risk communication occurred in the late 1980s, when risk practitioners began to reframe the issue of risk communication as a problem in communication theory and practice rather than as simply risk concepts [4]. A milestone in the evolution of current risk communication concepts and practice was the 1989 publication of *Improving Risk Communication*, by a group of committees under the National Research Council [3]. Two subsequent landmark documents in the United States on risk characterization and risk management have been extremely influential internationally in espousing the need to integrate communication and consultation throughout

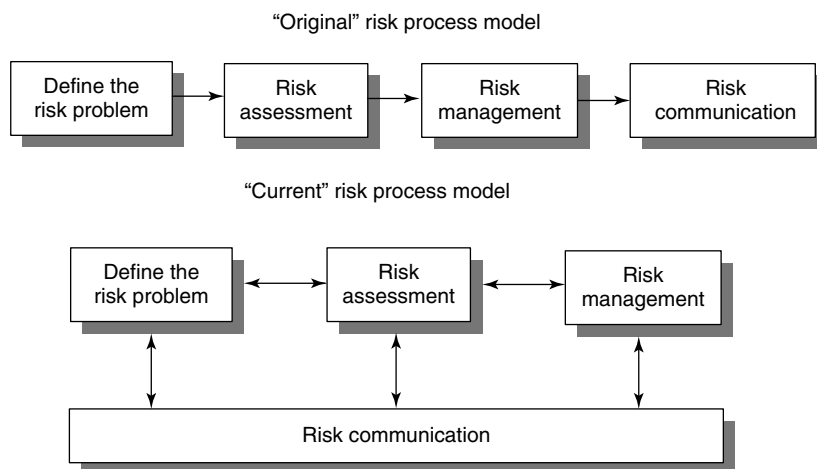


Figure 1 Evolving role of risk communication in risk process models

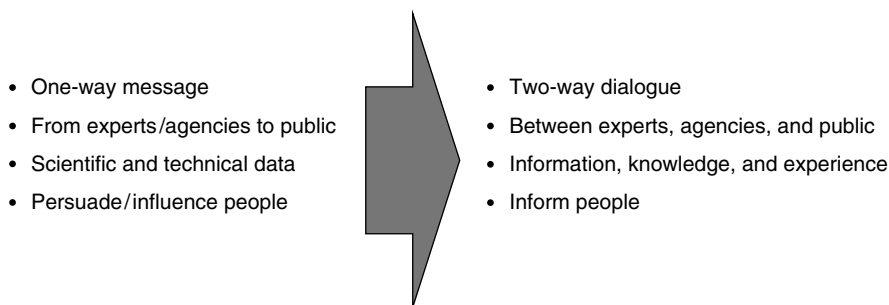


Figure 2 Evolution of risk communication

the risk-management process. *Understanding Risk: Informing Decisions in a Democratic Society* by the US National Research Council National Academy of Science [12] and the *Framework for Environmental Health Risk Management* by the US Presidential/Congressional Commission on Risk Assessment and Risk Management [13] both strongly recommended full collaboration and communication with all interested and affected parties in the early stages of problem formulation and solution development for environmental risk problems so that public values can inform and influence the shaping of risk-management strategies. While most previous models dealt largely with communication *effects*, these models recognized that communication is a *process* that is inextricably linked to the analysis and deliberation of the subject area being communicated.

Risk communication is now generally considered to be a two-way dialogue between experts and the public to exchange all types of information, knowledge, and experience about the risk situation (Figure 2). Its primary purpose is to inform people so that they may make responsible decisions on managing risk. The main tenet of informed and effective modern-day risk communication is the right of people and communities to participate in decisions that affect their lives and those of the people they care about (*see Stakeholder Participation in Risk Management Decision Making*). Communication with interested and affected parties should start with the formulation of the risk problem and continue through the assessment and selection of appropriate risk-management strategies (Figure 1).

Table 1 Risk communication stages^(a)

“All we have to do is get the numbers right.” . . . tell them the numbers.” . . . explain what we mean by the numbers.” . . . show them that they’ve accepted similar risks in the past.” . . . show them that it’s a good deal for them.” . . . treat them nice.” . . . make them our partners.” . . . all of the above”
--

^(a) Reproduced from [14]. © John Wiley & Sons, Inc., 1995

Unfortunately, not everyone has learned from the changing perspectives and approaches to risk communication and the skills that have been developed to put these concepts into practice. Many agencies and industries have felt compelled to repeat the learning process of their predecessors. Fischhoff [14] outlined eight development stages in the evolution of risk communication (Table 1). He concluded that ultimately all of the development stages are necessary for successful risk communication. We need rigorous, defensible science to conduct risk assessments. We need to present these risk estimates to the people involved, and do everything possible to ensure their understanding. We need to provide a context for this information, and to ensure that the information is balanced with respect to risks and potential benefits. And finally, we need to treat people with respect, and to involve them as full partners in the risk process to ensure that all types of information are considered.

Fischhoff further commented that, as in the biological concept that “ontogeny recapitulates phylogeny” (i.e., the development of the individual mimics the evolution of the species), risk practitioners seem bent on recreating this developmental sequence of failures for each new risk problem. His summary of risk communication evolution has resonated strongly with most risk communication practitioners because it so aptly and succinctly embodies the flawed learning progression of this discipline.

Risk Communication as an Integral Part of Risk Management

Rigorous, scientifically based risk assessment is a necessary input to the risk-management process. However, we now know (through a repeated series

of usually frustrating and painful past experiences) that more than good risk science is necessary to ensure successful, accepted risk-management decisions. Other factors (such as economic, social, cultural, ethical, political, and legal considerations) must also be incorporated in risk-management decisions [13]. Ultimately, this complex process of risk management cannot succeed without open, comprehensive, timely, and reciprocal dialogue on the risk issues. Risk communication has thus become the critical key to success for all risk-management initiatives.

However, incorporation of effective risk dialogue from the outset and throughout the risk-management process is still not being consistently achieved. Part of the problem is a lack of understanding by risk assessors and risk managers on the importance of risk communication. Equally problematic is the lack of institutional and organizational commitment to the full incorporation of risk communication in their risk processes [15]. The success of employing good risk communication practice is often measured in part by a relative lack of controversy on a risk issue. Unfortunately, this very success is often translated by organizations as no longer requiring the expenditure of time and resources to communicate because “nothing is going wrong”. When risk communication is not fully incorporated from the beginning of an issue, and the process does become contentious, organizations then (erroneously) assume that suddenly adopting risk communication procedures will resolve the problem. Both of these practices inevitably fail to achieve risk-management objectives, yet managers and organizations continue to resist committing the time and resources into incorporating risk communication as an integral part of their risk-management programs. We need to continue both to learn from

our past experiences and to apply these lessons consistently if we are to genuinely improve our dialogues on risk issues.

References

- [1] Thomas, S.P. & Hrudey, S.E. (1997). *Risk of Death in Canada: What We Know and How We Know It*, The University of Alberta Press, Edmonton.
- [2] U.S. National Research Council (1983). *Risk Assessment in the Federal Government: Managing the Process*, National Academy Press, Washington, DC.
- [3] U.S. National Research Council (1989). *Improving Risk Communication*, National Academy Press, Washington, DC.
- [4] Covello, V.T., von Winterfeldt, D. & Slovic, P. (1987). Communicating scientific information about health and environmental risks: problems and opportunities from a social and behavioral perspective, in *Uncertainty in Risk Assessment, Risk Management, and Decision-Making*, V.T. Covello, L.B. Lave, A. Moghissi & V.R.R. Uppuluri, eds, Plenum Press, New York, pp. 221–239.
- [5] Johnson, B.B. & Covello, V.T. (eds) (1987). *The Social and Cultural Construction of Risk*, Reidel, Dordrecht.
- [6] Leiss, W. & Krewski, D. (1989). Risk communication: theory and practice, in *Prospects and Problems in Risk Communication*, W. Leiss, ed, University of Waterloo Press, Waterloo, pp. 89–112.
- [7] Kasperson, R.E. (1992). The social amplification of risk: progress in developing an integrative framework, in *Social Theories of Risk*, D. Golding & S. Krinsky, eds, Praeger, Westport, pp. 153–178.
- [8] Morgan, M.G., Fischhoff, B., Bostrom, A., Lave, L. & Atman, C.J. (1992). Communicating risk to the public, *Environmental Science and Technology* **26**(11), 2048–2056.
- [9] Chess, C., Salomone, K.L. & Hance, B.J. (1995). Improving risk communication in government: research priorities, *Risk Analysis* **15**(2), 127–135.
- [10] Slovic, P. (1987). Perception of risk, *Science* **236**, 280–284.
- [11] Sandman, P.M. (1989). Hazard versus outrage in the public perception of risk, in *Effective Risk Communication: The Role and Responsibility of Government and Nongovernment Organizations*, V.T. Covello, D.B. McCallum & M.T. Pavlova, eds, Plenum Press, New York, pp. 45–49.
- [12] U.S. National Research Council (1996). *Understanding Risk: Informing Decisions in a Democratic Society*, National Academy Press, Washington, DC.
- [13] U.S. Presidential/Congressional Commission on Risk Assessment and Risk Management (1997). *Framework for Environmental Health Risk Management*, Final Report, Vols. 1 and 2, Washington, DC.
- [14] Fischhoff, B. (1995). Risk perception and communication unplugged: twenty years of process, *Risk Analysis* **15**(2), 137–145.
- [15] Leiss, W. (1996). Three phases in the evolution of risk communication practice, in *The Annals of the American Academy of Political and Social Science: Challenges in Risk Assessment and Risk Management*, H. Kunreuther & P. Slovic, eds, Sage Periodicals Press, Thousand Oaks, pp. 85–94.

CYNTHIA G. JARDINE

Ruin Probabilities: Computational Aspects

Ruin probabilities is a traditional research area within the field of collective risk theory in insurance mathematics. Since the pioneering works of Lundberg and Cramér, it has interested a number of researchers and practitioners. A large and rich literature can be found in journals devoted to actuarial sciences, probability, and statistics. The topic is developed to various degrees in the comprehensive books [1–7].

The present paper is concerned with the numerical evaluation of the ruin or nonruin probabilities, over an infinite horizon and in finite time. We focus primarily on the compound Poisson model, which is the classical model in collective risk theory. A succinct overview of related problems and models is given at the end of the article.

The list of references is nonexhaustive. The absence of reference for a specific result means that the result is standard and can be easily found in the books referred to above.

Compound Poisson Model

A risk model aims at describing and analyzing the time evolution of the reserves (or surplus) of an insurance company. It is built on the basis of two processes: the claim process, which is stochastic because the arrival of claims and their amounts are random by nature, and the premium process, which is quite often considered as deterministic because premiums are collected at a nearly known rate.

The most prominent model is the compound Poisson risk model. Its advantage is that it is mathematically rather simple and relies on intuitive grounds. The underlying assumptions are quite acceptable for many applications, but could be restrictive in some cases. This model also leads to the development of some more realistic and still tractable models.

Specifically, on the one hand, claims arrive according to a Poisson process $\{N(t)\}$ with mean $\lambda > 0$ per unit time. This assumption is natural for a large portfolio of insurance policies, all independent and each having a small probability of claim. The successive claim amounts $\{X_i\}$ are independent random variables with common distribution function

$F(x)$, with $F(0) = 0$, and independent of $\{N(t)\}$. So, claims occur according to a compound Poisson process $\{S(t)\}$ given by $S(t) = \sum_{i=1}^{N(t)} X_i$. On the other hand, the insurer has an initial surplus $u \geq 0$ and receives premiums continuously at a constant rate $c > 0$. This assumption is reasonable in practice, but extension to time-dependent rates is still possible. Combining both claim and premium processes, we see that the surplus at time t is

$$U(t) = u + ct - S(t) \quad (1)$$

Ruin (*see Ruin Theory*) occurs as soon as the surplus becomes negative or null, i.e., at time $T(u) = \inf\{t > 0 : U(t) \leq 0\}$ ($T(u) = \infty$ if ruin does not occur). In other words, $T(u)$ is the first-crossing time of the process $\{S(t)\}$ through the upper linear boundary $y = u + ct$. We denote by $\phi(u, t)$ the probability of nonruin until time t :

$$\begin{aligned} \phi(u, t) &= P[T(u) > t] \equiv P[U(\tau) \\ &= u + c\tau - S(\tau) > 0 \text{ for } 0 < \tau \leq t] \end{aligned} \quad (2)$$

$\psi(u, t) = 1 - \phi(u, t)$ is the probability of ruin before time t . As $t \rightarrow \infty$, equation (2) becomes the ultimate nonruin probability $\phi(u) = P[T(u) = \infty]$, and the ultimate ruin probability is $\psi(u) = 1 - \phi(u)$. These probabilities are the topics of interest hereafter.

Ultimate Ruin Probability

Let $\mu = E(X_1)$ denote the expected claim amount. If $c \leq \lambda\mu$, ruin is sure over infinite horizon ($\psi(u) = 1$ for any $u \geq 0$). This is not surprising since the condition means that, per unit time, the premium is at most equal to the expected aggregate claim. If, on the contrary, $c > \lambda\mu$, ultimate nonruin has a positive probability ($\phi(u) > 0$ for any $u \geq 0$, with $\phi(u) \rightarrow 1$ as $u \rightarrow \infty$). It is thus natural to assume that $c = \lambda\mu(1 + \theta)$ where $\theta > 0$ is the relative security loading.

Exact Results

To determine $\phi(u)$, a standard approach is to show that $\phi(u)$ satisfies an integro-differential equation, and this yields the following integral equation:

$$\begin{aligned} \phi(u) &= \phi(0) + \frac{\lambda}{c} \int_0^u \phi(v)[1 - F(u - v)] dv, \\ u &\geq 0 \end{aligned} \quad (3)$$

One gets from equation (3) that when $u = 0$ (no initial reserve),

$$\phi(0) = \frac{1 - \lambda\mu}{c (= \theta/(1 + \theta))} \quad (4)$$

When $u > 0$, equation (3) is called a defective renewal equation (it corresponds to a Volterra-type integral equation of the second kind). Several algorithms developed in numerical analysis permit one to obtain the numerical solution to such an equation (see, e.g., Appendix D in [6]). An alternative method is to derive from equation (3), the Laplace transform of $\phi(u)$ and then to invert it numerically (e.g. [8]).

From equation (3) one can also prove that $\phi(u)$ is the distribution function of a compound geometric distribution, i.e.,

$$\phi(u) = \left(\frac{1 - \lambda\mu}{c} \right) \sum_{i=0}^{\infty} \left(\frac{\lambda\mu}{c} \right)^i F_e^{*i}(u), \quad u \geq 0 \quad (5)$$

where $F_e^{*i}(u)$ is the i -fold convolution of a distribution function $F_e(u)$ which corresponds to the equilibrium distribution associated with $F(\cdot)$, i.e., $F_e(u) = (1/\mu) \int_0^u [1 - F(x)] dx$, $u \geq 0$. This representation of $\phi(u)$ is called the Pollaczek–Khinchine (or Beekman) formula. The result is elegant, but does not lend itself easily to numerical evaluations because of the infinite series of convolution powers in equation (5). It can simplify, however, for particular claim distributions. For example, if claims are exponentially distributed with $F(x) = 1 - e^{-x/\mu}$, $x > 0$, then $\phi(u) = 1 - (\lambda\mu/c) e^{-(1/\mu - \lambda/c)u}$, $u \geq 0$. A large class of claim laws for which formula (5) simplifies is the class of phase-type distributions. Formula (5) can also be used for estimating ruin probabilities by Monte Carlo simulation (e.g. [9]). Finally, this formula is of interest when addressing problems of a theoretical nature (such as the approximation of $\phi(u)$ for large u).

Approximations

An efficient method introduced in [10] is to first approximate the claim amount law by a discrete distribution. The successive claims are so approximated by independent discrete random variables $\{X_i\}$ with common probability mass function g_n , $n = 1, 2, \dots$. Several techniques permit arithmetizing distributions and provide, in general, a high degree of accuracy

(see, e.g., Section 6.15 in [6]). One can then prove that for discrete claims, the series given by formula (5) for $\phi(u)$ may be transformed into a simple finite sum. For clarity, choose u a nonnegative integer. Following [11], the formula is written as

$$\phi(u) = \left(\frac{1 - \lambda\mu}{c} \right) \sum_{j=0}^u h_j \left(\frac{j - u}{c} \right) \quad (6)$$

where the quantities $h_j(t)$ of argument $t \leq 0$ are pseudo-probability mass functions that can be computed using the recurrence equation (10) given below.

Approximation by moment fitting is a standard approach in statistics. The method presented in [12] consists of replacing the surplus process by another one with exponential claims of parameter $\hat{\beta}$, Poisson parameter $\hat{\lambda}$, and premium rate $\hat{c}$. The three new parameters are chosen to get the first three cumulants match, and it then suffices to apply the simple formula known for the exponential case. Although purely empirical, this procedure gives very good results in practice. An approximation method of similar nature is given in [13].

Bounds and Asymptotics

The adjustment coefficient (or Lundberg's exponent), denoted by γ , plays a central role here. Assume that the moment generating function of X_1 , i.e., $m(s) = Ee^{sX_1}$, is finite for some $s > 0$ (this implies that $P(X_1 > x)$ decreases exponentially fast). Then, γ is defined as the unique positive solution to the equation $\lambda + cs = \lambda m(s)$, if such a root exists. This equation may be easily solved numerically by applying, for instance, the Newton–Raphson method.

A fundamental result is Lundberg's inequality: $\psi(u) \leq e^{-\gamma u}$, $u \geq 0$ (the ruin probability is bounded by a decreasing exponential function, at rate γ , of the initial surplus). This upper bound is often taken into account in place of the true ruin probability (the larger the value of γ , the safer the portfolio). It can be used to find the initial surplus u , or the security loading θ , needed to keep the ruin probability below a fixed level. A refined inequality exists, as well as a similar exponential lower bound. Furthermore, the asymptotic behavior of $\psi(u)$ for large u is given by the celebrated Cramér–Lundberg approximation:

$$\psi(u) \approx Ke^{-\gamma u} \quad \text{as } u \rightarrow \infty, \\ \text{with } K = \frac{(c - \lambda\mu)}{[\lambda m'(\gamma) - c]} \quad (7)$$

where $m'(s)$ is the derivative of $m(s)$. The approximation works rather well with moderate values of u ; it becomes exact for exponentially distributed claims.

We underline that the basic assumption above is the existence of the coefficient γ . This is true with light-tailed claim amount distributions like the exponential or gamma. It is not the case with heavy-tailed distributions, like the lognormal or Pareto, which are useful laws for certain actuarial applications. Asymptotic approximations, however, can still be derived and exhibit a very different behavior of $\psi(u)$ (e.g. [14]).

Surpluses at Ruin

When ruin occurs ($T(u) < \infty$), $U[T(u)-]$ and $|U[T(u)]|$ measure the surplus just before ruin and the severity of ruin, respectively. These quantities give interesting information on what happens at ruin time. Let us suppose, for instance, that claim amounts are continuous variables with density function $p(x) = F'(x)$. We denote by $f(x, y|u)$ the defective joint density function of $U[T(u)-]$ and $|U[T(u)]|$, and by $f(x|u)$ the defective density function of $U[T(u)-]$, with $x, y > 0$.

When $u = 0$, one can prove the following key formula [15]:

$$f(x, y|0) = \left(\frac{\lambda}{c}\right) p(x+y), \quad x, y > 0 \quad (8)$$

Note that $f(x, y|0) = f(y, x|0)$, i.e., there is a symmetry between the laws of $U[T(u)-]$ and $|U[T(u)]|$. Integrating equation (8) over y , then gives $f(x|0) = (\lambda/c)[1 - F(x)]$, $x > 0$; obviously, it is also the density function of $|U[T(u)]|$ at point x .

When $u > 0$, it is easily seen that $f(x, y|u) = f(x|u) p(x+y)/[1 - F(x)]$, $x, y > 0$. One can then establish another key formula [16]:

$$f(x|u) = f(x|0) \frac{[\psi(u-x) - \psi(u)]}{[1 - \psi(0)]}, \quad x > 0 \quad (9)$$

where $\psi(\cdot)$ is the ultimate ruin probability, with the convention $\psi(u-x) = 1$ if $u < x$. All these results can be unified and extended through the study of an expected discounted penalty owing to ruin [17].

Finite-Time Ruin Probability

In this section, we consider again discrete claim amounts $\{X_i\}$ with probability mass function g_n , $n = 1, 2, \dots$. Continuous claims may also be discussed but, as already indicated, they are usually discretized for numerical calculations. For simplicity, u is a nonnegative integer.

We present two explicit formulae for $\phi(u, t)$. First, denote the probability mass function of $S(t)$ by $h_n(t)$, $n = 0, 1, 2, \dots$. Since $S(t)$ is of compound Poisson law, an easy way to compute $h_n(t)$ is to apply Panjer's recursion formula, which gives

$$h_0(t) = e^{-\lambda t}, \quad \text{and} \quad h_n(t) = \frac{\lambda t}{n} \sum_{j=1}^n j g_j h_{n-j}(t), \quad n = 1, 2, \dots \quad (10)$$

Note that the relation (10) may be written for any real t , positive or not. If $t < 0$, $\{h_n(t)\}$ corresponds to a pseudo-probability mass function (with no practical interpretation).

Now, when $u = 0$, a classical ballot theorem yields for $\phi(0, t)$ the Takàcs formula:

$$\phi(0, t) = \frac{1}{ct} \sum_{j=0}^{[ct]} (ct - j) h_j(t) \quad (11)$$

where $[x]$ is the integer part of x .

When $u > 0$, a simple probabilistic argument leads to the Seal formula for $\phi(u, t)$: if $ct \leq 1$, $\phi(u, t) = \sum_{j=0}^u h_j(t)$, and if $ct > 1$,

$$\begin{aligned} \phi(u, t) &= \sum_{j=0}^{u+[ct]} h_j(t) - \sum_{j=1}^{[ct]} h_{u+j} \left(\frac{j}{c}\right) \\ &\quad \times \phi\left(0, \frac{ct-j}{c}\right) \end{aligned} \quad (12)$$

An alternative approach proposed recently [11, 18] has recourse to the functions $h_n(t)$ defined for any real t . This permits one to obtain for $\phi(u, t)$ the Picard-Lefèvre formula:

$$\begin{aligned} \phi(u, t) &= \sum_{j=0}^u \left\{ h_j(t) + h_j \left(\frac{j-u}{c}\right) \sum_{n=u+1}^{u+[ct]} \right. \\ &\quad \times \left. \frac{u+ct-n}{u+ct-j} h_{n-j} \left(\frac{u+ct-j}{c}\right) \right\} \end{aligned} \quad (13)$$

Note that $(j - u)/c \leq 0$ in $h_j(\cdot)$ above. Formula (13) reduces to result (11) if $u = 0$ and becomes result (6) as $t \rightarrow \infty$. Numerical comparisons between formulae (12) and (13) show that formula (12) is well adapted when $u \gg ct$, while formula (13) is preferable when $u \ll ct$ [19].

Working with discrete claims is quite convenient because ruin can occur only at levels $u + 1, u + 2, \dots$, and the only possible ruin times are thus $1/c, 2/c, \dots$. Over a horizon $(0, t)$, it then suffices to examine the times j/c , $j = 1, \dots, [ct]$, and time t , which yields $\phi(u, t) = P\{S(j/c) < u + j, j = 1, \dots, [ct], S(t) < u + [ct] + 1\}$. When the premium income is arbitrary, adapting this reasoning permits the construction of a recursive method for $\phi(u, t)$, either directly [19] or by using the generalized Appell (or Sheffer) polynomials (e.g. [11]). Extensions are also possible to multirisks [20] and to interdependent claims (e.g. [21]).

Finally, we refer the reader to two approximation methods using a discrete-time scale [22, 23].

Related Problems and Models

The economic context can require incorporating additional factors in the model. Various extensions that can be taken into account are as follows: an arbitrary premium income, inhomogeneous claim amounts, premiums depending on the reserves, a fixed or random discount factor, a dividend barrier, interdependent claim amounts, dependence between arrival times and claim amounts, multiple risks or a Markovian environment.

The compound Poisson model itself can be questioned. Several variants often proposed are as follows: the compound binomial model (discrete-time analog of the model), the renewal Sparre–Andersen model (independent interclaim periods with arbitrary common law), and the compound Poisson model perturbed by a Brownian motion or an α -stable Lévy motion.

References

- [1] Asmussen, S. (2000). *Ruin Probabilities*, World Scientific, Singapore.
- [2] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, Heidelberg.
- [3] Gerber, H.U. (1979). *An Introduction to Mathematical Risk Theory*, S.S. Huebner Foundation Monograph, University of Philadelphia, Philadelphia.
- [4] Grandell, J. (1990). *Aspects of Risk Theory*, Springer-Verlag, New York.
- [5] Kaas, R., Goovaerts, M.J., Dhaene, J. & Denuit, M. (2001). *Modern Actuarial Risk Theory*, Kluwer Academic Publishers, Dordrecht.
- [6] Panjer, H.H. & Willmot, G.E. (1992). *Insurance Risk Models*, Society of Actuaries, Schaumburg.
- [7] Rolski, T., Schmidli, H., Schmidt, V. & Teugels, J. (1999). *Stochastic Processes for Insurance and Finance*, John Wiley & Sons, Chichester.
- [8] Embrechts, P., Grübel, R. & Pitts, S.M. (1993). Some applications of the fast Fourier transform in insurance mathematics, *Statistica Neerlandica* **47**, 59–75.
- [9] Asmussen, S. & Binswanger, K. (1997). Simulation of ruin probabilities for subexponential claims, *ASTIN Bulletin* **27**, 297–318.
- [10] Shiu, E.S.W. (1988). Calculation of the probability of eventual ruin by Beekman's convolution series, *Insurance: Mathematics and Economics* **7**, 41–47.
- [11] Picard, P. & Lefèvre, C. (1997). The probability of ruin in finite time with discrete claim size distribution, *Scandinavian Actuarial Journal* **1**, 58–69.
- [12] De Vylder, F.E. (1978). A practical solution to the problem of ultimate ruin probability, *Scandinavian Actuarial Journal* **2**, 114–119.
- [13] Beekman, J. (1969). A ruin function approximation, *Transactions of the Society of Actuaries* **21**, 41–48, 275–279.
- [14] Embrechts, P. & Veraverbeke, N. (1982). Estimates for the probability of ruin with special emphasis on the possibility of large claims, *Insurance: Mathematics and Economics* **1**, 55–72.
- [15] Dufresne, F. & Gerber, H.U. (1988). The surpluses immediately before and at ruin, and the amount of the claim causing ruin, *Insurance: Mathematics and Economics* **7**, 193–199.
- [16] Dickson, D.C.M. (1992). On the distribution of surplus prior to ruin, *Insurance: Mathematics and Economics* **11**, 191–207.
- [17] Gerber, H.U. & Shiu, E.S.W. (1998). On the time value of ruin, *North American Actuarial Journal* **2**, 48–72.
- [18] De Vylder, F.E. (1999). Numerical finite-time ruin probabilities by the Picard-Lefèvre formula, *Scandinavian Actuarial Journal* **2**, 97–105.
- [19] Rullière, D. & Loisel, S. (2004). Another look at the Picard-Lefèvre formula for finite-time ruin probabilities, *Insurance: Mathematics and Economics* **35**, 187–203.
- [20] Picard, P., Lefèvre, C. & Coulibaly, I. (2003). Multirisks model and finite-time ruin probabilities, *Methodology and Computing in Applied Probability* **5**, 337–353.

- [21] Ignatov, Z.G., Kaishev, V.K. & Krachunov, R.S. (2001). An improved finite-time ruin probability formula and its Mathematica implementation, *Insurance: Mathematics and Economics* **29**, 375–386.
- [22] De Vylder, F.E. & Goovaerts, M.J. (1988). Recursive calculation of finite-time ruin probabilities, *Insurance: Mathematics and Economics* **7**, 1–7.
- [23] Dickson, D.C.M. & Waters, H.R. (1991). Recursive calculation of survival probabilities, *ASTIN Bulletin* **21**, 199–221.

Related Articles

Individual Risk Models

Ruin Theory

Simulation in Risk Management

CLAUDE LEFÈVRE

Ruin Theory

Ruin theory examines the solvency of an insurer in a theoretical setting. The insurer's surplus or reserve for an insurance portfolio over time is modeled by a stochastic process. Of central importance in the evaluation of the surplus are the time of ruin at which the surplus becomes negative for the first time, the surplus immediately before the time of ruin, and the deficit at the time of ruin. The use of stochastic processes for modeling the surplus of an insurer can be dated back to 1909, when Lundberg presented his paper "On the theory of risk" at the Sixth International Congress of Actuaries (see [1]). In his paper, the surplus was described by a shifted compound Poisson process. The compound Poisson surplus process was further investigated by Cramér (see [2]) and is a topic of interest even today. This process has been extended and generalized in several directions. Interest rates are incorporated into surplus processes. The claim arrivals are assumed to follow a more general counting process than a Poisson process. Further, the fluctuation of the surplus is considered and modeled by a Brownian motion, and more generally by a diffusion process. Also considered are surplus processes with dividend policies are also considered.

The Compound Poisson Surplus Process

The compound Poisson surplus process can be described as follows: The Poisson process $N(t)$ with intensity λ representing the number of claims experienced by the insurance portfolio over time. More precisely, the number of claims over the time period $[0, t]$ is $N(t)$. The amount of the i th claim is represented by the random variable Y_i . Hence, the aggregate claims arising from the insurance portfolio over the time period $[0, t]$ is given by

$$S(t) = Y_1 + Y_2 + \cdots + Y_{N(t)} \quad (1)$$

with the convention that $S(t) = 0$ if $N(t) = 0$. It is assumed that these individual claim amounts $Y_1, Y_2, \dots$, are positive, independent and identically distributed with a common distribution function $P(x)$, and that they are independent of the number of claims processed $N(t)$. Thus, the aggregate claims

processed $\{S(t)\}$ is a compound Poisson process. As shown in *collective risk models* (see **Collective Risk Models**), the mean and the variance of $S(t)$ are $E\{S(t)\} = \lambda p_1 t$ and $\text{Var}\{S(t)\} = \lambda p_2 t$, where p_1 and p_2 are the first two moments about the origin of Y , a representative of the claim amounts Y_i 's. Suppose that the insurer has an initial surplus $u \geq 0$ and receives premiums continuously at the rate c per unit time. The insurer's surplus process is then given by

$$U(t) = u + ct - S(t) \quad (2)$$

The surplus process (2) is referred to as the *classical compound Poisson surplus process*. It is normally assumed that the insurer charges a risk premium and in this case, the premium rate c will exceed the expected aggregate claims per unit time, i.e., $c > E\{S(1)\} = \lambda p_1$. Let $\theta = [c - E\{S(1)\}]/E\{S(1)\} > 0$. Then one has $c = \lambda p_1(1 + \theta)$. The parameter θ is called the *relative security loading* (see **Ruin Probabilities: Computational Aspects**).

The time of ruin is given by $T = \inf\{t; U(t) < 0\}$ the first time the surplus becomes negative. The first quantity under consideration is the probability that ruin occurs in finite time, i.e., $\psi(u) = \Pr\{T < \infty\}$, the probability of ruin. With the independent increments property of the Poisson process, it can be shown that the probability of ruin $\psi(u)$ satisfies the defective renewal equation (see **Reliability Growth Testing**)

$$\begin{aligned} \psi(u) = & \frac{1}{1 + \theta} \int_0^u \psi(u - x) dP_1(x) \\ & + \frac{1}{1 + \theta} [1 - P_1(u)] \end{aligned} \quad (3)$$

where $P_1(x)$ is the equilibrium distribution function of $P(x)$ and is given by $P_1'(x) = [1 - P(x)]/p_1$. Since the probability of ruin $\psi(u)$ is the solution of the defective renewal equation (3), methods for renewal equations are applicable. The key renewal theorem (see [3]) implies that

$$\psi(u) \sim \frac{\theta p_1}{E\{Xe^{\kappa X}\} - (1 + \theta)p_1} e^{-\kappa u}, \quad \text{as } u \rightarrow \infty \quad (4)$$

where the parameter κ is the smallest positive solution of

$$\int_0^\infty e^{\kappa x} dP_1(x) = 1 + \theta \quad (5)$$

and $a(u) \sim b(u)$ means $\lim_{u \rightarrow \infty} a(u)/b(u) = 1$. The parameter κ is called the *Lundberg adjustment coefficient* in ruin theory and it plays a central role in analysis of not only the probability of ruin but also other important quantities in ruin theory. With the Lundberg adjustment coefficient κ , we can also obtain a simple exponential upper bound, called the *Lundberg bound*, for the probability of ruin:

$$\psi(u) \leq e^{-\kappa u}, \quad u \geq 0 \quad (6)$$

The defective renewal equation (3) also implies that $\psi(u)$ is the survival probability of a compound geometric distribution, where the geometric distribution has the probability mass function

$$\frac{\theta}{1+\theta} \left(\frac{1}{1+\theta} \right)^n, \quad n = 0, 1, \dots$$

and the secondary distribution is $P_1(x)$. This compound geometric representation enables us to obtain an analytical solution for $\psi(u)$. It is easy to see that

$$\psi(u) = \sum_{n=1}^{\infty} \frac{\theta}{1+\theta} \left(\frac{1}{1+\theta} \right)^n [1 - P_1^{*n}(u)], \quad u \geq 0 \quad (7)$$

where $P_1^{*n}(u)$ is the n -fold convolution of P_1 with itself. The formula (7) is often referred to as *Beekman's convolution formula* in actuarial science (see [4]) and the use of the formula can be found in [5, 6]. A disadvantage of the formula is that it involves an infinite number of convolutions, which can be difficult to compute in practice.

In some situations, a closed form solution for $\psi(u)$ can be obtained. When the individual claim amount distribution $P(x)$ is a combination of exponential distributions, i.e.,

$$P'(x) = \sum_{j=1}^m q_j \mu_j e^{-\mu_j x} \quad (8)$$

where $0 < \mu_1 < \mu_2 < \dots < \mu_m$ and $\sum_{j=1}^m q_j = 1$, then

$$\psi(u) = \sum_{j=1}^m C_j e^{-R_j u} \quad (9)$$

where R_j 's are the positive solutions of the Lundberg equation (5) and

$$C_j = \frac{\left\{ \sum_{i=1}^m \frac{q_i}{\mu_i - R_j} \right\}}{\left\{ \sum_{i=1}^m \frac{q_i \mu_i}{(\mu_i - R_j)^2} \right\}} \quad (10)$$

For further details see [7] and [8].

When the individual claim amount distribution $P(x)$ is a mixture of Erlang distributions, i.e.,

$$P'(x) = \sum_{j=1}^m q_j \frac{\mu^j x^{j-1} e^{-\mu x}}{(j-1)!} \quad (11)$$

where $\mu > 0$, $q_j \geq 0$, $j = 1, 2, \dots, m$ and $\sum_{j=1}^m q_j = 1$,

$$\psi(u) = e^{-\mu u} \sum_{j=0}^{\infty} C_j \frac{[\mu u]^j}{j!} \quad (12)$$

where the coefficients C_j 's are obtained by a recursive formula that can be found in [9]. An analytical expression is also available for more general phase-type individual claim amount distributions. However, its calculation is not as straightforward and requires the use of the matrix-analytic method (see [10] for details).

The classical results discussed above may be found in many books on risk theory, including [11–15].

The Expected Discounted Penalty Function

Further investigation of the surplus process involves the surplus immediately before the time of ruin $U(T-)$ and the deficit at the time of ruin $|U(T)|$. Dickson studied the joint distribution of $U(T-)$ and $|U(T)|$ in [16]. The expected discounted penalty function introduced by Gerber and Shiu in the seminal paper [17] is very useful for analysis of T , $U(T-)$, and $|U(T)|$ in a unified manner. The expected discounted penalty function is defined as

$$\phi(u) = E \left\{ e^{-\delta T} w(U(T-), |U(T)|) I(T < \infty) \right\} \quad (13)$$

where $w(x, y)$ is the penalty when the surplus immediately before the time of ruin is x and the deficit at

the time of ruin is y . Note that $I(A)$ is the indicator function of event A , and the parameter δ is the interest rate compounded continuously. Many quantities of interest may be obtained by choosing a proper penalty function $w(x, y)$. For example, when $w(x, y) = 1$, we obtain the Laplace transform of T where δ is the variable of the transform. If $w(x, y) = I(x \leq u, y \leq v)$ for fixed u and v , and $\delta = 0$, we obtain the joint distribution function of $U(T-)$ and $|U(T)|$. Furthermore, if $w(x, y) = x^k y^l$ and $\delta = 0$, the joint moments of $U(T-)$ and $|U(T)|$ are obtained.

It is shown in [17] that the expected discounted penalty function $\phi(u)$ satisfies the defective renewal equation

$$\begin{aligned} \phi(u) = & \int_0^u \phi(u-x)g(x) dx \\ & + \frac{\lambda}{c} e^{\rho u} \int_u^\infty e^{-\rho x} \int_x^\infty w(x, y-x) \\ & \times dP(y) dx \end{aligned} \quad (14)$$

where ρ is the unique nonnegative solution to the Lundberg equation

$$c\xi + \lambda\tilde{p}(\xi) - (\lambda + \delta) = 0 \quad (15)$$

and the kernel density $g(x)$ is given by

$$g(x) = \frac{\lambda}{c} \int_x^\infty e^{-\rho(y-x)} dP(y) \quad (16)$$

This result is of great significance as it allows for the utilization of the theory of renewal equations to analyze the expected discounted penalty function $\phi(u)$. For example, it can be shown that the Laplace transform of T is a survival probability of a compound geometric distribution. Moreover, the general expected discounted penalty function can be expressed in terms of the Laplace transform. As a result, many analytical properties of the expected discounted penalty function may be obtained (see [9, 17–19]).

The Compound Poisson Surplus Process with Interest Rate

The classical compound Poisson surplus process assumes that the surplus receives no interest over time. The compound Poisson surplus process with

constant interest rate was first studied in [20] and the model was further considered in [21–25]. They mainly focus on the probability of ruin. It is assumed that the insurer receives interest on the surplus at constant rate δ , compounded continuously. Thus, the surplus at time t can be expressed as

$$U(t) = ue^{\delta t} + c \frac{e^{\delta t} - 1}{\delta} - \int_0^t e^{\delta(t-s)} dS(s) \quad (17)$$

Sundt and Teugels [24] show that the probability of ruin $\psi(u)$ satisfies the integral equation

$$\begin{aligned} [c + \delta u]\psi(u) = & \int_0^u \psi(u-x) \\ & \times [\delta + \lambda(1 - P(x))] dx \\ & + \lambda \int_u^\infty (1 - P(x)) dx - \frac{\theta - \theta'}{1 + \theta'} \lambda p_1 \end{aligned} \quad (18)$$

where $\theta' < \theta$ is given in [24]. The equation (18) is nonlinear and hence is, in general, not solvable. However, an analytical solution exists for the exponential individual claims (see again [24]).

Recently, the expected discounted penalty function under the above model and with stochastic interest rates have been considered (see [26–28] and references therein).

The Compound Poisson Surplus Process with Diffusion

Dufresne and Gerber [29] first considered the compound Poisson surplus process perturbed by diffusion and assumed the surplus process is given by

$$U(t) = u + ct - S(t) + \sigma W(t) \quad (19)$$

where $\{W(t)\}$ is the standard Brownian motion and $\sigma > 0$. The diffusion (*see Role of Alternative Assets in Portfolio Construction; Default Risk*) term $\sigma W(t)$ may be interpreted as the additional uncertainty of the aggregate claims, the uncertainty of the premium income, or the fluctuation of the investment of the surplus, where σ is the volatility. In this

situation, the expected discounted penalty function needs to be modified. Let

$$\phi_d(u) = E \left\{ e^{-\delta T} I(T < \infty, U(T) = 0) \right\} \quad (20)$$

and

$$\begin{aligned} \phi_s(u) = E \left\{ e^{-\delta T} w(U(T-), |U(T)|) \right. \\ \left. I(T < \infty, U(T) < 0) \right\} \end{aligned} \quad (21)$$

The expected discounted penalty function is defined as

$$\phi(u) = w_0 \phi_d(u) + \phi_s(u) \quad (22)$$

The first term represents the penalty caused by diffusion and the penalty is w_0 , and the second term represents the penalty caused by an insurance claim. It is shown in [30] that the expected discounted penalties $\phi_d(u)$ and $\phi_s(u)$ satisfy the second-order integro-differential equations:

$$\begin{aligned} D\phi_d''(u) + c\phi_d'(u) + \lambda \int_0^u \phi_d(u-y) dP(y) \\ - (\lambda + \delta)\phi_d(u) = 0 \end{aligned} \quad (23)$$

$$\begin{aligned} D\phi_s''(u) + c\phi_s'(u) + \lambda \int_0^u \phi_s(u-y) dP(y) \\ + \lambda \int_u^\infty w(y, y-u) dP(y) \\ - (\lambda + \delta)\phi_s(u) = 0 \end{aligned} \quad (24)$$

where $D = \sigma^2/2$. Similar to the classical compound Poisson case, it has the Lundberg equation

$$D\xi^2 + c\xi + \lambda\tilde{p}(\xi) - (\lambda + \delta) = 0 \quad (25)$$

which has a nonnegative root ρ . The expected discounted penalty function $\phi(u)$ satisfies a defective renewal equation with the kernel density

$$g(x) = \frac{\lambda}{D} \int_0^x e^{-d(x-v)} \int_v^\infty e^{-\rho(y-v)} dP(y) \quad (26)$$

and $d = c/D + \rho$. An interesting finding is that the probability of ruin, as a special case, is a convolution of an exponential density with the tail of a compound geometric distribution (see [29]).

The compound Poisson process with diffusion equation (19) is a special case of Lévy processes and hence it is natural to consider a class of Lévy

processes for the surplus process that have only negative jumps. A number of papers discuss the probability of ruin for this class of surplus processes (see [31–33]). More recently, the expected discounted penalty function for general Lévy processes have been investigated in [34]. Generally, the results for Lévy surplus processes are very similar to those for the compound Poisson process with diffusion.

The Sparre Andersen Surplus Process

The compound Poisson surplus process implicitly assumes that the interclaim times are independently and identically exponentially distributed. A more general model involves the assumption that the interclaim times are independent and identically distributed, but not necessarily exponential. The resulting surplus process is referred to as a *renewal risk process*, or as a *Sparre Andersen process* after its usage was proposed in [35].

Analysis of the Sparre Andersen surplus process is more difficult than the compound Poisson special case, but some progress has been made. Because the process is regenerative at claim instants, it can be shown that the ruin probability is still a compound geometric tail. In fact, the expected discounted penalty function defined by equation (13) still satisfies a defective renewal equation, and the Laplace transform of the ruin time T is still a compound geometric tail. The defective renewal equation structure implies that Lundberg asymptotics and bounds hold in this more general situation. For further discussion of these issues, see [36–38] and references therein.

The difficulty with the use of this model, in general, is the fact that it is difficult to identify the geometric parameter and the ladder height (or geometric secondary) distribution. This is true even for the special case involving the ruin probabilities. Such identification normally requires that further assumptions be made about the interclaim time distribution and/or the claim size distribution. In [39], Dickson and Hipp considered Erlang(2) interclaim times, whereas Gerber and Shiu [36] and Li and Garrido [40] considered the Erlang(n) case. Also, Li and Garrido [37] considered the much more general case involving Coxian interclaims. Conversely, parametric assumptions about claim sizes rather than interclaim times were considered by Willmot [38]. The situation with phase-type assumptions is discussed in [10].

A closely related model is the delayed Sparre Andersen model where the time until the first claim is independent, but not distributed as the subsequent interclaim times. This model attempts to address the criticism that the Sparre Andersen model implicitly assumes that a claim occurs at time zero. Other classes of surplus processes that do not have stationary and independent increments are the nonhomogeneous Poisson process, the mixed Poisson process, and more generally, Cox processes. Similar to the renewal surplus process case, analysis of these processes is also difficult. For further details on the Sparre Andersen and related processes, see [10, 15, 41, 42] and references therein.

Dividend Strategies

As mentioned earlier, an insurer normally charges a risk premium on its policies, which implies that the premium incomes are greater than the expected aggregate claims. In this case, the probability of ruin is less than one. Hence, there is a positive probability that the surplus will grow indefinitely, an obvious shortcoming of the aforementioned surplus processes. To overcome this shortcoming, De Finetti in [43] suggested that a constant dividend barrier be imposed so that the overflow of the premium incomes is paid as dividends (the constant dividend barrier was originally applied to a binomial surplus process in his paper). This dividend strategy is usually referred to as the *constant barrier dividend Strategy*. Early studies on this strategy can be found in [11, 13, 44, 45]. A generalization of the constant barrier dividend strategy is the threshold dividend strategy under which a fixed proportion of the premiums is paid as dividends when the surplus is above the constant barrier. The threshold dividend strategy can also be viewed as a two-step premium as described in [10].

The research on surplus processes with dividend strategies is primarily concerned with (a) calculation of the expected discounted total dividends and identification of optimal dividend strategies and (b) the expected discounted penalty function. When the surplus process is compound Poisson, the expected discounted total dividends satisfies a homogeneous piecewise first-order integro-differential equation, and hence is solvable. The result also holds for the

Lévy surplus process except that the equation is of second order. When the objective is to maximize the expected discounted total dividends with no constraints, the optimal dividend strategy for the compound Poisson surplus process is the constant barrier dividend strategy if the individual claims follow an exponential distribution and the so-called band strategy for an arbitrary individual claim distribution. If an upper bound is imposed for the dividend rate, then the threshold dividend strategy is optimal for the compound Poisson surplus process with the exponential individual claim distribution. It is generally believed that these results also hold for general Lévy processes, but they have not been proved; see [11, 46] and references therein. Other objectives may be used for choosing an optimal constant barrier. Dickson and Waters [47] consider the maximization of the difference between the expected discounted total dividends and the expected discounted deficit at the time of ruin. It was extended to maximize the difference between the expected discounted total dividends and the expected discounted penalty in [48].

There has been renewed interest in studying the time of ruin and related quantities for surplus processes with dividend policies in recent years, especially since the introduction of the expected discounted penalty function in [17]. The probability of ruin is given in [10] for the compound Poisson surplus. It is shown in [49, 50] that the expected discounted penalty function satisfies a nonhomogeneous piecewise first-order integro-differential equation, and hence is solvable in many situations. Some related papers on this problem include [51, 52]. An interesting result among others (see [50]) is the dividends-penalty identity that states that the increment in the expected discounted penalty owing to the constant barrier strategy is proportional to the expected discounted total dividends. This result has been extended to the stationary Markov surplus process in [48]. Other extensions include nonconstant barriers and multiple constant barriers (see [53–57] and references therein). There are a number of papers considering the renewal surplus process with a constant dividend barrier; see [58], for example. However, a constant dividend barrier is less interesting as it can not provide an optimal dividend strategy owing to the nonstationarity of the claim arrivals.

References

- [1] Lundberg, F. (1909). Über die theorie der rickversicherung, *Ber VI Intern Kong Versich Wissens* **1**, 877–948.
- [2] Cramér, H. (1930). *On the Mathematical Theory of Risk*, Skandia Jubilee Volume, Stockholm.
- [3] Ross, S.S. (1995). *Stochastic Processes*, 2nd Edition, John Wiley & Sons, New York.
- [4] Beekman, J.A. (1968). Collective risk results, *Transactions of Society of Actuaries* **20**, 182–199.
- [5] Shiu, E.S.W. (1988). Calculation of the probability of eventual ruin by Beekman's convolution series, *Insurance: Mathematics and Economics* **7**, 41–47.
- [6] Willmot, G.E. (1988). Further use of Shiu's approach to the evaluation of ultimate ruin probabilities, *Insurance: Mathematics and Economics* **7**, 275–282.
- [7] Dufresne, F. & Gerber, H. (1988). The probability and severity of ruin for combinations of exponential claim amount distributions and their translations, *Insurance: Mathematics and Economics* **7**, 75–80.
- [8] Gerber, H.U., Goovaerts, M. & Kaas, R. (1987). On the probability and severity of ruin, *ASTIN Bulletin* **17**, 151–163.
- [9] Lin, X. & Willmot, G.E. (1999). Analysis of a defective renewal equation arising in ruin theory, *Insurance: Mathematics and Economics* **25**, 63–84.
- [10] Asmussen, S. (2000). *Ruin Probabilities*, World Scientific, Singapore.
- [11] Bühlmann, H. (1970). *Mathematical Methods in Risk Theory*, Springer-Verlag, New York.
- [12] De Vylder, F.E. (1996). *Advanced Risk Theory: A Self-Contained Introduction*, Editions de L'Université de Bruxelles, Brussels.
- [13] Gerber, H.U. (1979). *An Introduction to Mathematical Risk Theory*, S.S. Huebner Foundation, University of Pennsylvania, Philadelphia.
- [14] Kaas, R., Goovaerts, M., Dhaene, J. & Denuit, M. (2001). *Modern Actuarial Risk Theory*, Kluwer Academic Publishers, Boston.
- [15] Rolski, T., Schmidli, H., Schmidt, V. & Teugels, J. (1998). *Stochastic Processes for Insurance and Finance*, John Wiley & Sons, Chichester.
- [16] Dickson, D.C.M. (1992). On the distribution of surplus prior to ruin, *Insurance: Mathematics and Economics* **11**, 191–207.
- [17] Gerber, H.U. & Shiu, E.S.W. (1998). On the time value of ruin, *North American Actuarial Journal* **2**(1), 48–72. Discussions, 72–78.
- [18] Lin, X. & Willmot, G.E. (2000). The moments of the time of ruin, the surplus before ruin and the deficit at ruin, *Insurance: Mathematics and Economics* **27**, 19–44.
- [19] Willmot, G.E. & Lin X.S. (2001). *Lundberg Approximations for Compound Distributions with Insurance Applications*, *Lecture Notes in Statistics* 156, Springer-Verlag, New York.
- [20] Segerdahl, C.O. (1954). A survey of results in collective risk theory, *Probability and Statistics*, John Wiley & Sons, Stockholm, pp. 276–299.
- [21] Boogaert, P. & Crijns, V. (1987). Upper bounds on ruin probabilities in case of negative loadings and positive interest rates, *Insurance: Mathematics and Economics* **6**, 221–232.
- [22] Delbaen, F. & Haezendonck, J. (1990). Classical risk theory in an economic environment, *Insurance: Mathematics and Economics* **6**, 85–116.
- [23] Harrison, J.M. (1977). Ruin problems with compounding assets, *Stochastic Processes and their Applications* **5**, 67–79.
- [24] Sundt, B. & Teugels, J. (1995). Ruin estimates under interest force, *Insurance: Mathematics and Economics* **16**, 7–22.
- [25] Sundt, B. & Teugels, J. (1997). The adjustment coefficient in ruin estimates under interest force, *Insurance: Mathematics and Economics* **19**, 85–94.
- [26] Cai, J. & Dickson, D.C.M. (2002). On the expected discounted penalty function at ruin of a surplus process with interest, *Insurance: Mathematics and Economics* **30**, 389–404.
- [27] Kalashnikov, V. & Norberg, R. (2002). Power tailed ruin probabilities in the presence of risky investments, *Stochastic Processes and their Applications* **98**, 211–228.
- [28] Wu, R., Wang, G. & Zhang, C. (2005). On a joint distribution for the risk process with constant interest force, *Insurance: Mathematics and Economics* **36**, 365–374.
- [29] Dufresne, F. & Gerber, H. (1991). Risk theory for the compound Poisson process that is perturbed by diffusion, *Insurance: Mathematics and Economics* **12**, 9–22.
- [30] Gerber, H.U. & Landry, B. (1998). On the discounted penalty at ruin in a jump-diffusion and the perpetual put option, *Insurance: Mathematics and Economics* **22**, 263–276.
- [31] Huzak, M., Perman, M., Sikic, H. & Vondracek, Z. (2004). Ruin probabilities and decompositions for general perturbed risk processes, *Annals of Applied Probability* **14**, 1378–1397.
- [32] Klüppelberg, C., Kyprianou, A.E. & Maller, R.A. (2004). Ruin probabilities and overshoots for general Levy insurance risk processes, *Annals of Applied Probability* **14**, 1766–1801.
- [33] Yang, H.L. & Zhang, L. (2001). Spectrally negative Lévy processes with applications in risk theory, *Advances in Applied Probability* **33**, 281–291.
- [34] Garrido, J. & Morales, M. (2006). On the expected discounted penalty function for the Levy risk processes, *North American Actuarial Journal* **10**(4), 196–216.
- [35] Andersen, E.S. (1957). On the collective theory of risk in the case of contagion between claims, *Transactions XVth International Congress of Actuaries* **2**, 219–229.

- [36] Gerber, H.U. & Shiu, E.S.W. (2005). The time value of ruin in a Sparre Andersen model, *North American Actuarial Journal* **9**(2), 49–84.
- [37] Li, S. & Garrido, J. (2005). On a general class of renewal risk process: analysis of the Gerber-Shiu function, *Advances in Applied Probability* **37**, 836–856.
- [38] Willmot, G.E. (2007). On the discounted penalty function in the renewal risk model with general inter-claim times, *Insurance: Mathematics and Economics* **41**, 17–31.
- [39] Dickson, D.C.M. & Hipp, C. (2001). On the time of ruin for Erlang(2) risk processes, *Insurance: Mathematics and Economics* **29**, 333–344.
- [40] Li, S. & Garrido, J. (2004). On ruin for the Erlang(n) risk process, *Insurance: Mathematics and Economics* **34**, 391–408.
- [41] Grandell, J. (1991). *Aspects of Risk Theory*, Springer-Verlag, New York.
- [42] Grandell, J. (1997). *Mixed Poisson Processes*, Chapman & Hall, London.
- [43] De Finetti, B. (1957). Su un'ipostazione alternativa della teoria collettiva del rischio, *Transactions of the XV International Congress of Actuaries* **2**, 433–443.
- [44] Gerber, H.U. (1973). Martingales in risk theory, *Mitteilungen der Vereinigung Schweizerischer Versicherungsmathematiker* **73**, 205–216.
- [45] Segerdahl, C.O. (1970). On some distributions in time connected with the collective theory of risk, *Scandinavian Actuarial Journal* **53**, 167–192.
- [46] Gerber, H.U. & Shiu, E.S.W. (2006). On optimal dividend strategies in the compound Poisson model, *North American Actuarial Journal* **10**(2), 76–93.
- [47] Dickson, D.C.M. & Waters, H.R. (2004). Some optimal dividend problems, *ASTIN Bulletin* **34**, 49–74.
- [48] Gerber, H.U., Lin, X.S. & Yang, H. (2006). A note on the dividends-penalty identity and the optimal dividend barrier, *ASTIN Bulletin* **36**, 489–503.
- [49] Lin, X.S. & Pavlova, K. (2006). The compound Poisson risk model with a threshold dividend strategy, *Insurance: Mathematics and Economics* **38**, 57–80.
- [50] Lin, X.S., Willmot, G.E. & Drekcic, S. (2003). The classical risk model with a constant dividend barrier: analysis of the Gerber-Shiu discounted penalty function, *Insurance: Mathematics and Economics* **33**, 551–566.
- [51] Yuen, K.C., Wang, G. & Li, W.K. (2007). The Gerber-Shiu expected discounted penalty function for risk processes with interest and a constant dividend barrier, *Insurance: Mathematics and Economics* **40**, 104–112.
- [52] Zhou, X. (2005). On a classical risk model with a constant dividend barrier, *North American Actuarial Journal* **9**(4), 95–108.
- [53] Albrecher, H. & Hartinger, J. (2007). A risk model with multi-layer dividend strategy, *North American Actuarial Journal* **11**(2), 43–64.
- [54] Albrecher, H. & Kainhofer, R. (2002). Risk theory with a nonlinear dividend barrier, *Computing* **68**, 289–311.
- [55] Albrecher, H., Hartinger, J. & Tichy, R. (2005). On the distribution of dividend payments and the discounted penalty function in a risk model with linear dividend barrier, *Scandinavian Actuarial Journal* (2), 103–126.
- [56] Gerber, H.U. (1981). On the probability of ruin in the presence of a linear dividend barrier, *Scandinavian Actuarial Journal* 105–115.
- [57] Lin, X.S. & Sendova, K. (2007). The compound Poisson risk model with multiple thresholds, *Insurance Mathematics and Economics*, in press.
- [58] Li, S. & Garrido, J. (2004). On a class of renewal risk models with a constant dividend barrier, *Insurance: Mathematics and Economics* **35**, 691–701.

X. SHELDON LIN AND GORDON E. WILLMOT

Safety

Patient safety is of primary importance; safety must be ensured throughout the clinical trials (*see* **Randomized Controlled Trials**) that are conducted during the process of developing a new drug (intervention) and maintained after the drug is approved for use. Once a drug is approved for marketing, the safety data that are available provide patients and physicians with the ability to assess safety risks relative to efficacy benefits.

The safety assessment takes on different forms that depend on the various phases of the development program (phases I–IV) and continues through surveillance of marketed drugs. In preclinical animal studies (*see* **Cancer Risk Evaluation from Animal Studies**) and in early Phase I studies, we examine the absorption, excretion, dose ranging, tolerance, and other pharmacokinetics measures in humans [1]. In phase I dose finding studies, dose limiting toxicities are identified and the maximum tolerated dose as well as the recommended phase II dose are identified on the basis of the dose escalation scheme and expected and unexpected toxicities that occur during the process. The safety assessments in phase II and III trials include ongoing monitoring of adverse events (AEs) and serious adverse events (SAEs). (*See* **Toxicity (Adverse Events)**.)

Safety is assessed at the individual patient level, the trial level, and across trials. AEs are defined, as such, if they arise for the first time during treatment, or, if they increase in severity during treatment (if they are present prior to the start of treatment). AEs that occur to an individual patient are recorded and reviewed; SAEs are reported to regulatory agencies and study sponsors; and the toxicity profile within a patient is reviewed. The patient level AE review focuses on severity, drug causality, drug relatedness, and resolution of the AE. At the trial level, this review includes summaries of the proportion of patients on the drug with specified AEs (including severity and/or drug relatedness). Summaries can be grouped by types of abnormalities in laboratory measurements, and by anatomy (body system) and/or pathophysiology (changes in body function caused by disease). Trends or patterns of AEs in subgroups of patients (e.g., gender, age, and ethnicity subgroups) can also be identified. The experimental group can be compared to the control group in phase III by

examining the time to the first occurrence of specified AEs (using Kaplan–Meier curves), and by estimating *relative risks* (*see* **Relative Risk**), *odds ratios* (*see* **Odds and Odds Ratio**), and *hazard ratios* (*see* **Hazard and Hazard Ratio**). Safety data can be combined across clinical trials using techniques of meta-analysis (*see* **Meta-Analysis in Clinical Risk Assessment**).

In some phase II trials, and in all phase III randomized clinical trials, an independent data safety and monitoring committee (DSMC) monitors the ongoing trial for safety (as described above), and, if planned, for efficacy. See [2, 3] for discussion of DSMC roles.

Once a drug is approved, postmarketing (phase IV) trials are conducted to study additional indications, new dosing regimens, or new formulations. These trials are similar to phase III trials both in conduct and monitoring. In addition to the safety data captured in phase IV trials, all serious, unexpected AEs that occur in the general population that uses the drug must be reported to the sponsor and to regulatory agencies. New or unexpected serious AEs, which are rare, can usually be detected only in postmarketing phase when relatively large numbers of patients use the drug compared to the numbers treated in clinical trials.

One should note, however, that assessment of safety when a drug is widely used is difficult because patient and physician AE reports are *ad hoc* (spontaneous), and the extent of drug usage in the population is usually unknown. Often, once a drug is available, anecdotal evidence and spontaneous reports lead to the perception that usage is associated with an unexpected safety problem. Such perception can lead to a reanalysis of the clinical trial data, additional clinical trials for safety, review of regulatory spontaneous reporting databases, or active surveillance programs.

Various statistical approaches to the identification of potential safety issues (risks) have been proposed using routine databases such as the US food and drug administration (FDA) spontaneous reporting system. For example, Bayesian data mining has been proposed to identify compounds or classes of compounds with safety risks [4–7].

One should note that active postmarketing surveillance can be planned *a priori* or when potential safety issues are identified during drug development. For example, active surveillance is often planned for vaccines that are generally given to healthy subjects, and in many cases to children [8].

References

- [1] Chow, S.-C. & Liu, J.-P. (2004). Safety assessment, in *Design and Analysis of Clinical Trials: Concepts and Methodologies*, 2nd Edition, Wiley-Interscience, New York, p. 562–601.
- [2] Ellenberg, S.S., Fleming, T.R. & DeMets, D.L. (2003). *Data monitoring Committees in Clinical Trials: A Practical Perspective*, John Wiley & Sons, Chichester.
- [3] DeMets, D.L., Furberg, C.D. & Friedman, L.M. (2005). *Data Monitoring in Clinical Trials: A Case Studies Approach*, Springer, New York.
- [4] DuMouchel, W. (1999). Bayesian data mining in large frequency tables with an application to the FDA spontaneous reporting system, *The American Statistician* **53**, 177–190.
- [5] O'Neill, R.T. & Szarfman, A. (1999). Bayesian data mining in large frequency tables with an application to the FDA spontaneous reporting system: discussion, *The American Statistician* **53**, 190–196.
- [6] Louis, T. & Shen, W. (1999). Bayesian data mining in large frequency tables with an application to the FDA spontaneous reporting system: discussion, *The American Statistician* **53**, 196–198.
- [7] Madigan, D. (1999). Bayesian data mining in large frequency tables with an application to the FDA spontaneous reporting system: discussion, *The American Statistician* **53**, 198–200.
- [8] Iskander, J.K., Miller, E.R., Pless, R.P. & Chen, R.T. (2004). *Vaccine Safety Postmarketing Surveillance: the Vaccine Adverse Event Reporting System*, Centers for Disease Control and Prevention, Department of Health and Human Services. Document 130012.

JUDITH D. GOLDBERG AND BENJAMIN
LEVINSON

Sampling and Inspection for Monitoring Threats to Homeland Security

Since the events of September 11, 2001, there has been increased emphasis on monitoring and surveillance to detect and prevent further terrorist attacks. While significant resources have been devoted to the *mechanics* of screening (e.g., improved X-ray equipment, chemical “sniffers”), considerably less attention has been paid to quantifying the efficacy of these surveillance programs.

Given the high volumes of passengers, containers, mail items, and various other inter and intracontinental “movements” an important consideration is the identification of a level of inspection that balances cost, inconvenience, detection success, and probability of “false triggering”. From a statistical perspective, the answer is partly provided by the theory and methods of statistical process control (SPC) and elementary probability theory. However, bio-surveillance, syndromic surveillance (*see Syndromic Surveillance*), and counterterrorism surveillance (*see Managing Infrastructure Reliability, Safety, and Security; Game Theoretic Methods; Public Health Surveillance*) are fundamentally different from monitoring activities for quality assurance in an industrial manufacturing process. Unlike the industrial setting, where there is generally good information on the performance of the manufacturing process (e.g., percent defective, proportion of nonconforming or “out-of-spec” items), monitoring in the context of bio/homeland security is characterized by extreme uncertainty. For example, because of the nefarious activities of terrorist organizations, security and intelligence organizations do not always know what it is they are looking for. As terrorists become increasingly sophisticated in their modes of attack, our uncertainty in both the likelihood of an attack and which sentinels to monitor increases. Furthermore, considerations of cost and logistics mean that we could never hope to reduce the likelihood or probabilities of such events to zero.

Given both these constraints and uncertainties, one is thus faced with the problem of how best to sample and inspect objects that potentially represent a threat to homeland security ([1] and references therein).

These objects could be, for example, containers entering shipping ports with concealed threats such as biological agents, radioactive material or weapons [2]; water reservoirs with threats of contamination; or aircraft with bombs and/or terrorists on board.

Various techniques can be used to model risks and uncertainty [3–5] associated with homeland security, including classical max–min theories, worst-case scenarios (*see Premium Calculation and Insurance Pricing; Risk Measures and Economic Capital for (Re)insurers*), expert opinion (*see Risk in Credit Granting and Lending Decisions: Credit Scoring; Operational Risk Modeling; Reliability Demonstration*), and Bayesian [6] methods (*see Repair, Inspection, and Replacement Models; Imprecise Reliability; Lifetime Models and Risk Assessment; Bayesian Statistics in Quantitative Risk Assessment*). More recently, Information Gap decision theory has been applied to monitoring for homeland security [1, 7, 8].

We present here a simple model for sampling and inspecting generic “objects” to detect threats to homeland security. The model yields rules of thumb that could be used by decision makers to set (upper) limits on the number of objects to be inspected for a given or desired level of risk, or to estimate levels or limitations to risk as a function of the number of objects inspected. The approach can be generalized in various ways, for example, to take account of imperfect detection rates [9] and several monitoring protocols [10] when the underlying probabilities of threats are subject to info-gap uncertainty.

A Model for Sampling and Inspection

Suppose that there are N objects entering an inspection point (e.g., a container port) over a time period of t units (e.g., t weeks), and that n of these objects are sampled and inspected for a particular threat T .

Define p to be the probability that an object contains a T and θ to be the probability that a T is detected if one is present. Assuming that there are no clustering effects or serial correlations, the probability that precisely k of those objects inspected contain a T and are detected, and that the remaining $N - k$ objects contain no T is $(p\theta)^k(1 - p)^{N-k}$. Multiplying this expression by $\binom{n}{k}$, the number of ways of choosing k from n , and summing over $k = 0, 1, \dots, n$ gives an expression for the probability that

no T passes undetected over some time period t . The compliment of this probability, *viz.*,

$$P_N(n, t) = 1 - [1 - p(1 - \theta)]^n [1 - p]^{N-n} \quad (1)$$

is then the probability that at least one T passes through the inspection point undetected in time t .

Ideally, one would like to minimize $P_N(n, t)$, which is easily seen from equation (1) to occur when $n = N$, i.e., when every object is inspected – which is obvious, although not achievable in practice. Furthermore, since N and n are increasing functions of time t , it follows from equation (1) that $P_N(n, t)$ approaches unity for large t except when $\theta = 1$ and $n = N$. In other words, regardless of the precise values of p , θ , n , and N (except $\theta = 1$ and $n = N$), at least one threat T will eventually pass through the inspection point undetected. To limit risks to homeland security, for example, by imposing a (small) upper bound on $P_N(n, t)$, it is thus clear that one must impose limits on import volumes (N) over specified time periods (t). For example, if we require

$$P_N(n, t) \leq \pi_c \ll 1 \quad (2)$$

with π_c some specified small number, we deduce from equations (1) and (2), by rearranging equation (1), and taking logarithms that

$$n \log \left(1 + \frac{p\theta}{1-p} \right) + N \log(1-p) \geq \log(1-\pi_c) \quad (3)$$

Assuming $p \ll 1$ (and $\pi_c \ll 1$), we can use the linear approximation $\log(1+x) \approx x$ ($|x| \ll 1$) to rewrite equation (3) as

$$N \leq \frac{\pi_c}{p(1-f\theta)} \quad (4)$$

where

$$f = \frac{n}{N} \quad (5)$$

is the inspection fraction which we assume, for illustrative purposes, to be a constant.

It follows from equation (4) that the import volume has a finite upper bound except, as noted, when $f = \theta = 1$. In practice, of course, $\theta < 1$ so that regardless of the precise values of π_c , p , θ , and f , equation (4) implies that limits must be imposed on import volumes if one wishes to reduce the chance

of threats passing inspection points undetected [2]. It is also clear from equation (4) that if one wishes to increase the upper limit on N for given π_c , or equivalently, decrease the lower limit on π_c for given N , one should choose the largest possible value of the inspection fraction f in equation (5) that is consistent with economic and other constraints. The problem, of course, is that precise values for these limits depend on p and θ which are themselves unknown and uncertain. Nevertheless, equation (4) provides a simple rule of thumb that decision makers could use in practice, using, for example, what-if or worst-case scenario estimates for p and θ .

As an example, suppose that economic constraints limit inspection fractions to 75%, and worst-case estimates are $p = 10^{-4}$ and $\theta = 2/3$. For this scenario, equation (4) gives

$$N \leq \sim 2\pi_c \times 10^{+4} \quad (6)$$

Thus, if we require $\pi_c = 10^{-2}$, say, then equation (6) implies $N \leq 200$; this presumably would impose severe restrictions on the (unit) time period over which one satisfies the likelihood constraint of equation (2). Similarly, if say, 1000 objects pass through an inspection point every month, the condition imposed by equation (2) is only achievable over a 2-month period ($N = 2000$) when $\pi_c \geq 1/10$. In other words, in this scenario over a 2-month time frame, one can only be assured that there is not more than a 1 in 10 chance of at least one threat passing undetected.

Discussion

In this article, we have presented a simple model for sampling and inspection of objects for threats to homeland security. An important general conclusion is that unless one inspects every object entering an inspection point, and unless one has perfect detection rates, at least one threat will eventually pass through the inspection point undetected. We thus need to impose limits on import volumes and periods to limit the chances of threats passing undetected. The simple model presented here can be generalized and extended in various ways by using info-gap decision theory [7] to model uncertainties in the underlying probabilities of threats being present and detected in import objects [1]. In info-gap terminology, the inequality represented by equation (2) is only one

of many possible performance requirements that decision makers may consider in arriving at so-called robust-optimal solutions [1] in which there are trade-offs between desired levels of performance and immunity to uncertainty [7]. Generalized info-gap models for homeland security are the subject of ongoing research [9, 10].

References

- [1] Moffitt, L.J., Stranlund J.K. & Field, B.C. (2005). Inspections to avert terrorism: robustness under severe uncertainty, *Journal of Homeland Security and Emergency Management* 2(3), 1–17 Art 3, <http://www.bepress.com/jhsem/vol2/iss3/3>.
- [2] Harrold, J.R., Stephens, H.W. & van Dorp, J.P. (2004). A framework for sustainable port security, *Journal of Homeland Security and Emergency Management* 1(2), 1–21 Art 12, <http://www.bepress.com/jhsem/vol1/iss2/12>.
- [3] French, S. (1986). *Decision Theory: An Introduction to the Mathematics of Rationality*, Ellis Horwood, Chichester.
- [4] Render, B., Stair Jr, R.M. & Hann, M.E. (2003). *Quantitative Analysis for Management*, Prentice Hall, Englewood Cliffs.
- [5] Burgman, M.A. (2005). *Risks and Decisions for Conservation and Environmental Management*, Cambridge University Press.
- [6] Ouchi, F.A. (2004). *Literature Review on the Use of Expert Opinion in Probabilistic Risk Analysis*, World Bank Policy Research Working Paper 3201, World Bank, World Bank Institute.
- [7] Ben-Haim, Y. (2006). *Information-Gap Decision Theory: Decisions Under Severe Uncertainty*, 2nd Edition, Academic Press, San Diego.
- [8] Yoffe, A. & Ben-Haim, Y. (2006). An Info-Gap approach to policy selection for bio-terror response, *IEEE International Conference on Intelligence and Security Informatics, ISI 2006*, San Diego, pp. 554–559, May.
- [9] Fox, D.R. & Thompson, C.J. (2007). *Risk and Uncertainty in Bio-Surveillance Having Imperfect Detection Rates*, Proceeding 56th session International Statistical Institute, Lisbon Portugal, pp. 22–29.
- [10] Thompson, C.J. & Fox, D.R. (2006). Robust monitoring and resource allocation among anti-terrorist protocols, In preparation.

COLIN THOMPSON AND DAVID R. FOX

Scenario Simulation Method for Risk Management

In financial risk management, two types of risk measurements are commonly used. The first type measures the sensitivities of portfolio value to some particular market variables. Usually, a portfolio's risk profile can be described by a large number of those sensitivities. Examples include delta, gamma, and vega (*see Risk-Neutral Pricing: Importance and Relevance; Weather Derivatives; Default Risk; Statistical Arbitrage*) in options portfolios, or duration and convexity (*see Asset-Liability Management for Life Insurers*) in bond portfolios.^a The second type is more comprehensive as it calculates the probability distribution of the portfolio value at a given horizon. This then provides common risk measures that summarize the portfolio risk, such as the widely used value at risk (VaR) (*see Risk Measures and Economic Capital for (Re)insurers; Credit Scoring via Altman Z-Score; Compliance with Treatment Allocation*), defined as the maximum loss from an adverse market movement with a specified probability over a period of time.

Among the commonly applied methods to estimate VaR, the simplest is "delta approximation" (*see Simulation in Risk Management*). The method, however, critically depends on two assumptions: the normality assumption of portfolio value, and the linearity assumption of the relationship between transactions' prices and market variables. For most portfolios, especially for portfolios with options and/or embedded options (*see Options and Guarantees in Life Insurance; From Basel II to Solvency II – Risk Management in the Insurance Sector*), Monte Carlo simulation (*see Simulation in Risk Management; Structured Products and Hybrid Securities; Reliability Optimization; Uncertainty Analysis and Dependence Modeling*) is a more appropriate method. The difficulty with the Monte Carlo approach is its computational burden. To obtain a reliable estimate, the sample size has to be large. Since each sample requires a repricing of the entire portfolio, the required large

sample size often makes the Monte Carlo approach impractical.

The scenario simulation method (*see Scenario-Based Risk Management and Simulation Optimization*) described here is a computationally efficient alternative to conventional Monte Carlo for multicurrency fixed-income portfolios. Following [1], the model approximates a multidimensional lognormal distribution of interest rates and exchange rates by a multinomial distribution of key factors. While it allows very large samples of correlated yield-curve (*see Structured Products and Hybrid Securities; Credit Migration Matrices*) joint scenarios, the number of scenarios in each currency is limited, which implies that the number of portfolio evaluations is limited.

Scenario simulation provides the entire distribution of future portfolio returns (see Figures 6 and 7). From this, not only VaR, but standard deviation and other measures of risk, such as the "coherent measures of risk" of Artzner *et al.* [2], can be computed. The scenario simulation model can be applied in estimating a portfolio's overall risk profile of joint market and credit risks. In addition, scenario simulation is better suited than traditional Monte Carlo to implementing stress tests.

We present the model in two stages: single currency scenario simulation and multicurrency scenario simulation, and then discuss its applications in risk management.

Single Currency Scenario Simulation

Similar to the traditional Monte Carlo method, we use a vector of key zero-coupon rates to describe a yield curve:

$$\{r_1, \dots, r_i, \dots, r_n\} \quad (1)$$

We assume that these rates are correlated, and the stochastic process of each key rate is described by the lognormal diffusion equation^b

$$\frac{dr_i}{r_i} = \mu_i(t) dt + \sigma_i dz_i \quad (2)$$

where z_i are correlated Brownian motions (*see Numerical Schemes for Stochastic Differential Equation Models*) with a correlation matrix $\mathbf{R}$. Using the current yield curve and the covariance matrix of $\{r_1, \dots, r_i, \dots, r_n\}$, the traditional Monte

Carlo method would directly generate a large number of yield curves. As we pointed out, the portfolio repricing would consume significant computational resources. Moreover, unless the sample size is very large, there are possibilities that the Monte Carlo sample may not cover adequately the tails of the distribution, which are of significance in evaluating a portfolio's risk exposure.

The scenario simulation method takes a different approach. First, one applies principal component (PC) analysis to reduce the dimensionality.^c Most empirical evidence supports the result that the first three PCs can explain almost the entire variation of yield-curve movements. By using the first three PCs, we derive the following approximation for the yield-curve movements:

$$\frac{dr_i(t)}{r_i(t)} \approx \mu_i(t) dt + \delta_{i1} dw_1 + \delta_{i2} dw_2 + \delta_{i3} dw_3 \quad (3)$$

where

$$\delta_{ij} = \sigma_i \beta_{ij} \quad (4)$$

and β_{ij} is the i th component of the j th PC of the correlation matrix $\mathbf{R}$. We note that the principal factors $\{w_i, i = 1, 2, \dots, n\}$ are independent Brownian motions.

In the second step, the problem is further simplified by using the multinomial distribution to discretize the multinormal distribution. In other words, instead of generating a large number of yield curves, each with the same probability, we generate a limited number of scenarios, each with a different probability. For example, we can select seven scenarios for the first PC, five scenarios for the second PC, and three scenarios for the third PC. This yields 105 yield-curve scenarios with different assigned probabilities, thus providing the distribution of yield-curve movements. It has been numerically demonstrated that for risk management purposes, this set of scenario curves is as good as or better than the Monte Carlo method with a large sample size (see [1]). By design, the set of scenarios always includes extreme cases with very small probabilities. Figure 1 summarizes the methodology of the single currency scenario simulation.

Example 1 Japanese Yen (JPY) Scenario Curves. Using RiskMetrics data as of 04/07/1999, we obtain the first three PCs (see **Fault Detection and Diagnosis**) for the correlation matrices of Japanese Yen yield-curve movements as shown in Table 1.

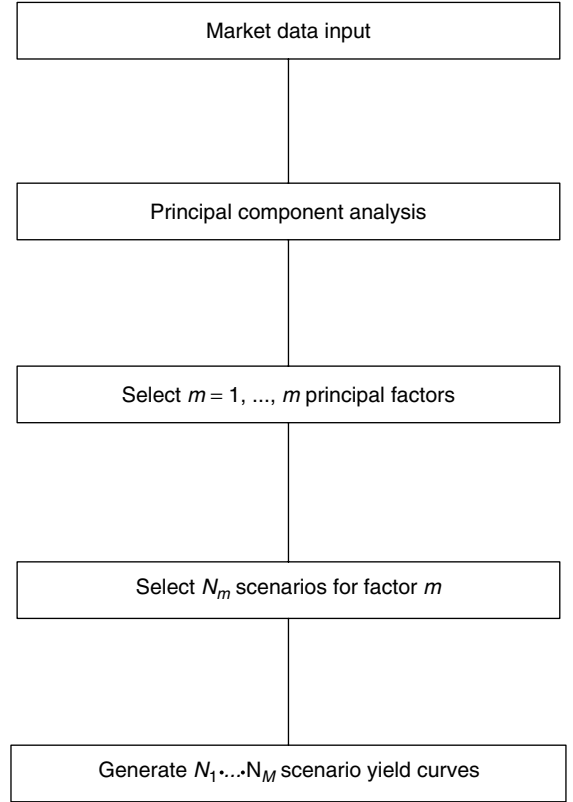


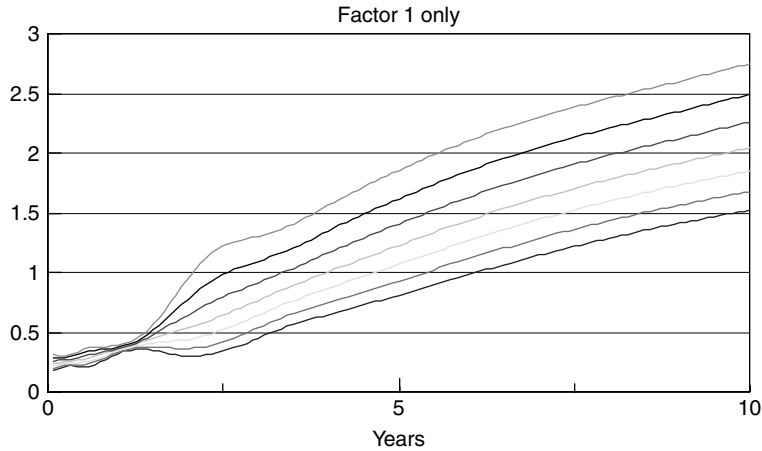
Figure 1 Diagram of single currency scenario simulation model

The columns correspond to zero-coupon rates with respective maturities. The numbers in the first row are all of the same sign, thus the corresponding principal factor is often explained as a “parallel shift” factor. The second factor can be called a *twist* factor and the third one, a *butterfly* factor.

In the scenario simulation model, each principal factor w_i is discretized by a binomial distribution with N_i successive outcomes one standard deviation apart. Since the first principal factor is most important, seven scenarios ($N_1 = 7$) are selected; for the second principal factor five scenarios ($N_2 = 5$) are selected; and three scenarios are selected for the third factor ($N_3 = 3$).^d Altogether, there are ($n_1 = 0, \dots, 6; n_2 = 0, \dots, 4; n_3 = 0, \dots, 2$) possible scenarios, and the total number of scenarios is $7 \times 5 \times 3 = 105$. Because all principal factors are independent of each other, the probabilities of yield-curve scenarios can be calculated easily. As an example,

Table 1 The first three principal components of JPY yield curve

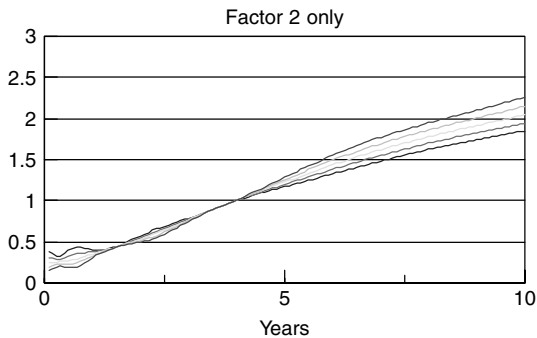
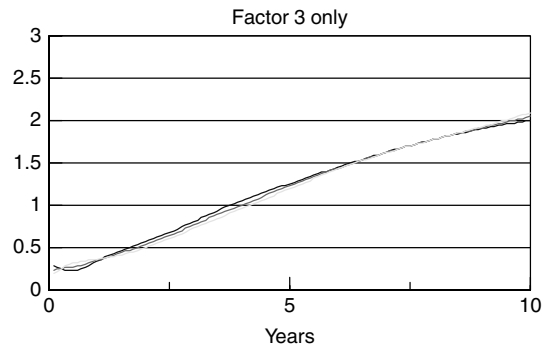
	1 month	3 months	6 months	1 year	2 years	3 years	4 years	5 years	7 years	9 years	10 years	20 years
First PC	0.340	0.440	0.407	0.423	0.892	0.945	0.962	0.965	0.938	0.913	0.883	0.731
Second PC	-0.750	-0.810	-0.716	-0.819	-0.141	-0.043	0.019	0.141	0.297	0.343	0.376	0.461
Third PC	-0.419	0.014	0.420	0.298	-0.170	-0.166	-0.176	-0.094	-0.008	0.039	0.127	0.365

**Figure 2** JPY yield curve scenarios

the probability of scenario ($n_1 = 4, n_2 = 2, n_3 = 1$) is equal to

$$\frac{15}{64} \times \frac{3}{8} \times \frac{1}{2} \approx 4.395\% \quad (5)$$

Figures 2–4 show the Japanese Yen scenario yield curves over a 1-year horizon.^c All curves are zero-coupon yield curves, and the middle curve is the current yield curve 1 year forward. The seven curves

**Figure 3** JPY yield curve scenario**Figure 4** JPY yield curve scenario

in Figure 2 are generated using only the first principal factor. It shows that the changes in the yield levels resemble parallel shifts. Figures 3 and 4 demonstrate that the second factor and the third factor reflect the twist and the butterfly movements of the yield curve, respectively. Each curve is associated with a probability of the corresponding scenario. For example, the third curve from the top of Figure 2 represents scenario (4, 2, 1) whose probability is 4.395%.

Multicurrency Scenario Simulation

The advantage of the scenario simulation method becomes more evident when the portfolio consists of derivative transactions in multiple currencies. In the multicurrency scenario simulation, we use the empirical correlation between principal factors of yield curves of different currencies to Monte Carlo simulate correlated yield-curve scenarios jointly in multiple currencies. This is done by the Gaussian copula method.^f Indeed, we know each discretized principal factor is binomially distributed, so (centering each at zero) we can construct a unique joint distribution of the principal factors of all currencies using these (centered) binomially distributed marginals in addition to using the Gaussian copula with the correlation matrix of the principal factors. This then easily provides a random joint-curve scenario by simply generating a multivariately normally distributed random number.

In greater detail, let

$$\begin{aligned} a_{i+1} &= \Psi^{-1}(F(i)) \quad i = 0, \dots, m; \\ (a_0 &= -\infty, a_{m+1} = +\infty) \end{aligned} \quad (6)$$

where Ψ represents the standard normal cumulative distribution function of the normal distribution, and F represent the binomial cumulative distribution function. Thus, a_i ($i = 0, \dots, m+1$) are the points on the real line such that the area under the normal curve on the segment $[a_i, a_{i+1}]$ equals the binomial probability of state i . For example, if $m = 6$, we have

$$\begin{aligned} a_1 &= -2.15387, a_2 = -1.22986, a_3 = -0.40225, \\ a_4 &= 0.40225, a_5 = 1.22986, a_6 = 2.15387 \end{aligned} \quad (7)$$

The area under the normal curve on $(a_0, a_1]$ is equal to $1/64$, the probability of the first state in this seven-state binomial distribution, and so on.

Define the function $B^{(m)}$ on the real line with values in the set $\{0, \dots, m\}$:

$$B^{(m)}(z) = i \quad \text{if } a_i \leq z < a_{i+1} \quad (8)$$

Clearly, if Z is a normally distributed with mean zero and variance one, then $B^{(m)}(Z)$ is binomially distributed with $m+1$ states.

Now, if X is a k -dimensional normal variate with correlation matrix $\mathbf{Q}$,

$$X = (X_1, \dots, X_k) \sim N(0, \mathbf{Q}) \quad (9)$$

each X_i having mean zero and variance one, define its discretization

$$B^{(m)} = (B^{(m)}(X_1), \dots, B^{(m)}(X_k)) \quad (10)$$

Each coordinate of $B^{(m)}$ is binomially distributed and equals the discretization of the corresponding component of X . As such, this multivariate discretization X preserves the univariate discretization of each component of X . It can be shown that $B^{(m)}$ converges to X with probability one as m increases.

To generate correlated multicurrency yield-curve scenarios requires random samples of $B^{(m)}$. First, we generate a multivariate $N(0, \mathbf{Q})$ normal deviate $x = (x_1, \dots, x_k)$, then simply find between which a_i and a_{i+1} each x_j lies, and thus form $(B^{(m)}(x_1), \dots, B^{(m)}(x_k))$. Recall each $B^{(m)}(x_i)$ represents a currency yield-curve scenario. The above procedure thus generates a joint yield-curve scenario, that is, a yield curve for each currency.

The joint probability density of $B^{(m)}$ can be written as

$$\begin{aligned} \text{prob}[B^{(m)} = (i_1, \dots, i_k)] \\ = \int_{a_{i_1}}^{a_{i_1+1}} \dots \int_{a_{i_k}}^{a_{i_k+1}} \\ \times p(x_1, \dots, x_k) dx_1 \dots dx_k \end{aligned} \quad (11)$$

where $p(x) = p(x_1, \dots, x_k)$ is the multivariate normal density function of X with correlation matrix $\mathbf{Q}$.

Figure 5 illustrates the methodology. First, scenario simulation is performed in each currency. If three factors and a total of 105 scenarios for each currency curve are used, and seven scenarios for each exchange rate are selected, each currency portfolio will have an output with $105 \times 7 = 735$ joint interest rate and exchange rate scenarios, except for the base currency portfolio which has 105 scenarios. Then, to obtain the risk exposure of the whole portfolio in the base currency, joint market scenarios based on the correlation matrix of principal factors and exchange rates are generated using Monte Carlo simulation as described. For a given joint multicurrency market scenario s , the portfolio value $V(s)$ is simply

$$V(s) = \sum_{k=1}^K X_k(s) \cdot V_k(s) \quad (12)$$

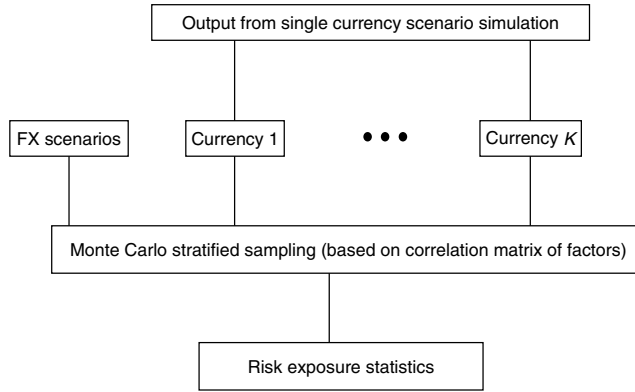


Figure 5 Diagram of multicurrency scenario simulation

where X_k is the exchange rate and V_k is the portfolio value of currency k . Using this method, even when a million joint scenarios are generated, only 105 yield curves are produced for each currency, and therefore each transaction has to be priced only 105 times rather than a million times.

In summary, the scenario simulation method makes the simulation computationally practical for very large multicurrency portfolios of fixed-income securities and derivatives.

Applications to Value-at-Risk Estimation and Stress Testing

The scenario simulation model provides an effective alternative to the delta approximation method and the traditional Monte Carlo method. We can directly apply the model by valuing each transaction say, 105 times, and obtain the portfolio's profit and loss distribution, and with it, the portfolio risk exposure statistics. The portfolio's VaR is simply one of the portfolio risk statistics, namely,

$$VaR = \max\{v: \text{Probability}(-\Delta V \leq v) \leq \alpha\} \quad (13)$$

where α is the confidence level and ΔV is the random change in portfolio value. For example, if α is 99%, then with 99% probability the amount of loss $(-\Delta V)$ will be less than VaR.

Usually, portfolio risk parameters, such as delta and gamma, are readily available to risk managers. In this case, the calculation can be further simplified for short horizons. Suppose that δ and γ are the delta

and gamma of the k th currency portfolio. Delta is a vector whose i th component is defined as the first-order derivative of the market value V_k with respect to y_i , the i th bucket forward rate. Gamma is the matrix of the second-order derivatives of V_k with respect to y_i . For a given yield-curve scenario s , its portfolio value change $\Delta V_k(s)$ can be approximated by

$$\Delta V_k(s) \approx \delta \cdot \Delta y(s) + \frac{1}{2} \Delta y(s)^T \cdot \gamma \cdot \Delta y(s) \quad (14)$$

where $\Delta y(s)$ represents the vector of forward rates changes from the current yield curve to the given scenario curve. In contrast to the full simulation method, which requires each transaction be valued in all scenarios, with this method, the delta and gamma of each transaction are calculated only once, using the current yield curves. Once delta and gamma are calculated, the above approximation yields for each scenario s the change $\Delta V_k(s)$ without the need to evaluate the portfolio at the scenario s . Once $\Delta V_k(s)$ is thus calculated, the portfolio's profit and loss distribution and VaRs at various confidence levels can be obtained by the stratified sampling and equation (13) as before.

Example 2 Estimating VaR by Scenario Simulation. Table 2 shows VaR calculation results using scenario simulation method for three simple derivative transactions, all denominated in US Dollars (USD) and with a \$100 million notional. The first transaction is a 5-year receiver swap (*see Securitization/Life; Structured Products and Hybrid Securities*) (party A receives fixed rate and pays 6-month LIBOR); the second transaction party A sells

Table 2 VaR using scenario simulation method

	Swap receiver	Cap	Payer swaption	Portfolio
Full simulation	\$2 722 285	\$2 045 630	\$1 226 831	\$2 975 627
Delta only	2 825 615	1 960 242	1 407 511	3 188 506
Delta and gamma	2 774 731	2 121 828	1 205 577	3 007 058

a 5-year 6% cap, and the third is a 1 year payer swaption into 5-year swap with fixed rate at 7%. Two-week VaRs are calculated using 99% confidence level. The method using both delta and gamma gives closer approximation to the full simulation results than the delta only method. The last column of Table 2 shows VaRs of the portfolio with these three transactions.

The scenario simulation model can also be applied to perform stress testing. VaR analysis is usually supplemented by stress testing, using the data from extraordinary historical market events. It is also important to stress test the assumptions of VaR models such as volatilities and correlations. Because of its computational efficiency, scenario simulation method is particularly suitable for such testing. In particular, the method allows risk managers to examine a portfolio's risk exposures within the tail of a given distribution and to identify specific stress scenarios under which the portfolio may become vulnerable.

Applications to Portfolio's Credit Risk Exposure Estimation

There are three basic ingredients that go into estimating a portfolio's potential credit risk:

1. market value of bonds and transactions for each issuer/counterparty;
2. issuer or counterparty credit quality and default or rating migration probabilities;
3. the recovery rate distribution.

A portfolio will suffer a credit loss if an issuer or a counterparty defaults, the market value of the portfolio with respect to the issuer or the counterparty is positive, and the recovery rate is less than 100%. We have the following basic credit risk model:

$$\text{Credit loss} = \max(0, \text{mark-to-market value}) \times (1 - \text{recovery rate}) \times d \quad (15)$$

where $d = 1$ if counterparty defaults, and 0 otherwise. Clearly estimating credit risk is related to, but more complicated than, market risk estimation. All the above three ingredients are random, and they are usually correlated with each other. Unlike market risk management where the horizon is usually short, say one day or a few days, credit risk management looks at much longer horizons, often the entire life of a transaction.

The scenario simulation model can be applied to estimate a portfolio's credit risk exposure as well as its joint market and credit risk exposure. For a given market scenario (s), we define the credit risk exposure with respect to a counterparty at time t as

$$V_t^+(s) = \max\{V_t(s), 0\} \quad (16)$$

Using the scenario simulation model, $V_t^+(s)$ for all selected scenarios and for various horizons t are readily available, and credit risk parameters such as maximum credit exposure with a given confidence level (credit-VaR) can be calculated. In addition, the correlation of defaults among different issuers/counterparties can be incorporated into the scenario simulation model. For example, in the Gaussian copula (see **Credit Risk Models; Simulation in Risk Management; Default Correlation; Copulas and Other Measures of Dependency**) approach, the joint default distribution $\{d_1, d_2, \dots, d_n\}$ can be mapped into a Gaussian distribution $\Psi(x_1, x_2, \dots, x_n; \Sigma)$, where Σ is the correlation matrix.⁸ The default scenarios can then be generated to compute the potential credit losses under various market scenarios.

Example 3 Market and Credit Exposures of a Multicurrency Portfolio. Figures 6–8 illustrate the simulation results for a multicurrency swap portfolio. The portfolio contains more than 300 interest rate and currency swaps in five different currencies. Figure 6 shows the portfolio's market exposure over a 1-month period. Not surprisingly the simulated distribution

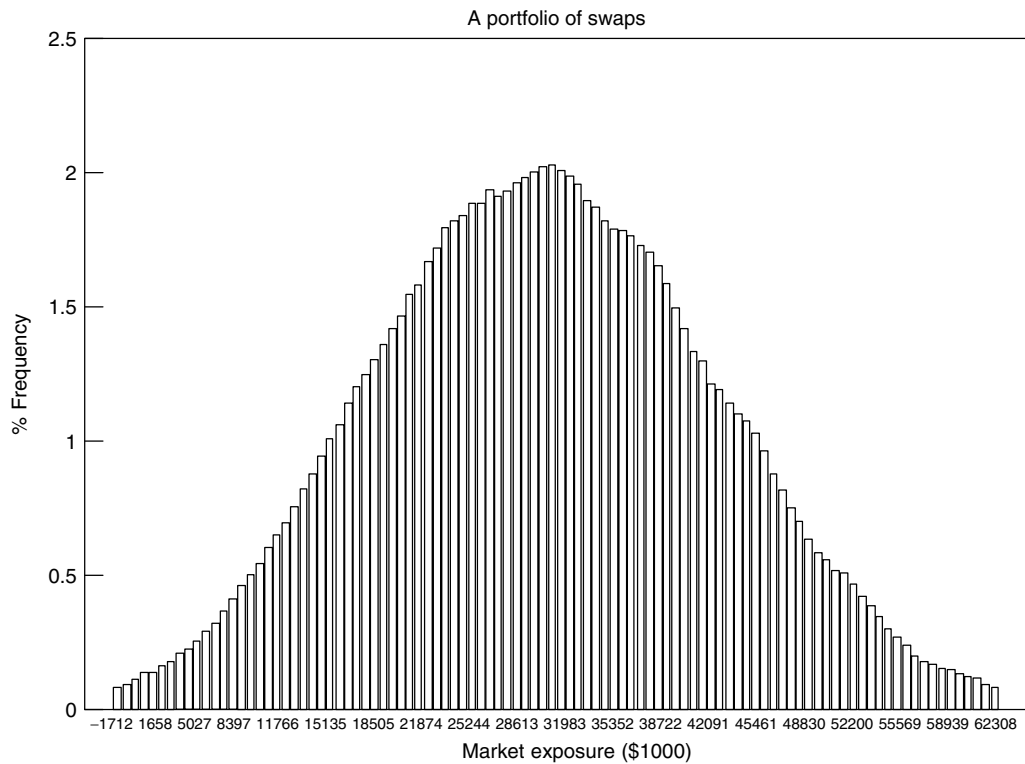


Figure 6 Market exposure distribution

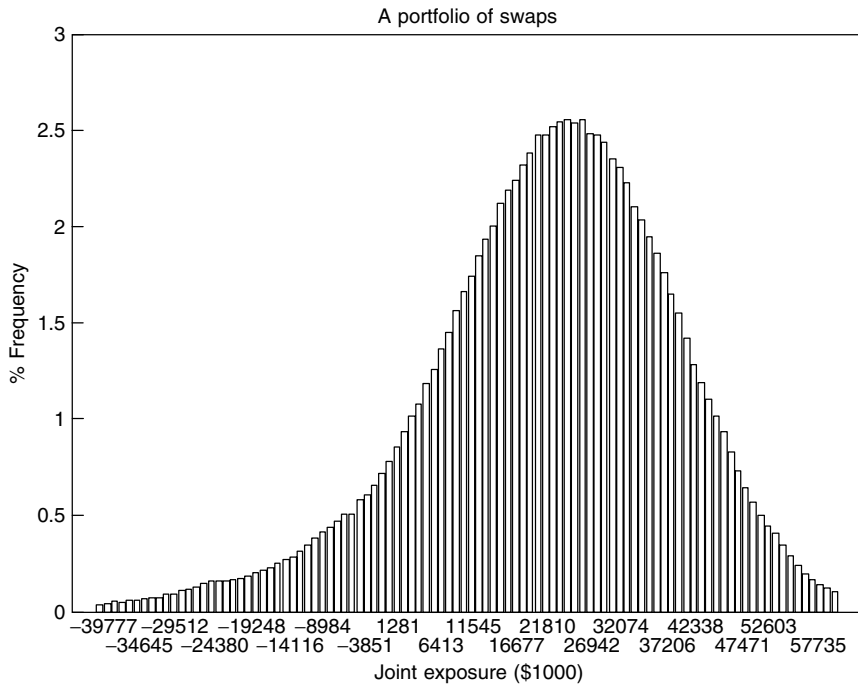


Figure 7 Joint exposure distribution

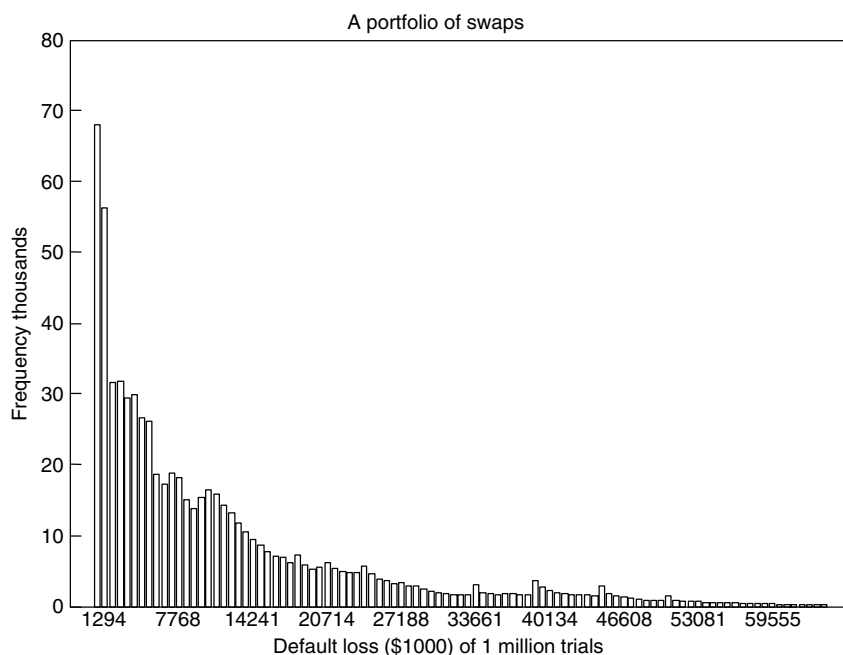


Figure 8 Default loss distribution

and credit exposure is no longer symmetric as shown in Figure 7. The portfolio's potential loss distribution due to counterparty default is illustrated in Figure 8.

End Notes

- ^a. Quite often, these risk parameters are vectors.
- ^b. When yield-curve scenarios are simulated over long time horizons such as in the case of estimating credit exposures, we incorporate mean reversions into the processes.
- ^c. Principal component analysis is a multivariate statistical method (see [3]). A working group of Bank for International Settlement (BIS) studied various methodologies of measuring aggregate market risk, including scenario generating using principal component analysis. See [4].
- ^d. The actual number of factors and the number of scenarios for each factor should be determined empirically.
- ^e. The scenario curves are based on Japanese Yen swap yield curve as of 05/03/1999.
- ^f. The Gaussian copula method is way of creating a multivariate distribution function given the marginals

and a correlation matrix. See [5] for an in-depth exposition.

^g. Copula functions are utilized by Li [6] to describe correlated defaults.

References

- [1] Jamshidian, F. & Zhu, Y. (1997). Scenario simulation: theory and methodology, *Finance and Stochastics* **1**, 43–67.
- [2] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1997). Thinking coherently, *Risk* **10**, 68–71.
- [3] Dillon William, R. & Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications*, John Wiley & Sons.
- [4] Loretan, M. (1997). "Generating Market Risk Scenarios Using Principal Components Analysis: Methodological and Practical Considerations", BIS Research Papers.
- [5] Nelsen, R. (1999). *An Introduction to Copulas*, Springer-Verlag, New York.
- [6] Li, D. (2000). On default correlation: a copula approach, *Journal of Fixed Income* **9**, 43–54.

FARSHID JAMSHIDIAN AND YU ZHU

Scenario-Based Risk Management and Simulation Optimization

Without uncertainty there is no risk. However, very few – if any – real-world situations are risk free. In the world of deterministic optimization, we often choose to “ignore” uncertainty in a given situation in order to come up with a unique and objective solution to a problem. But in situations where uncertainty is at the core of the problem, as it is in those that involve risk management (*see Role of Risk Communication in a Comprehensive Risk Management Approach*), a different strategy is required.

In the field of optimization, there are various approaches designed to cope with uncertainty. In this context, the exact values that the parameters (e.g., the data) of the optimization problem will have are not known with absolute certainty, but may vary to a larger or lesser extent depending on the nature of the factors they represent. In other words, there may be many possible “realizations” of the parameters, each of which is called a *scenario*.

Sections titled “Scenario Optimization” and “Robust Optimization” briefly describe traditional scenario-based approaches to optimization, such as *scenario optimization* and *robust optimization*. The main section titled “Simulation Optimization” then illustrates an important real world application of a recently emerging approach that is having a profound impact on risk management called *simulation optimization*.

Scenario Optimization

In [1] Dembo offers an approach to solving stochastic programs based on a method for solving deterministic scenario subproblems and combining the optimal scenario solutions into a single feasible decision.

Imagine a situation in which we want to minimize the cost of producing a set J of finished goods. Each good j ($j = 1, \dots, n$) has a per unit production cost c_j associated with it, as well as an associated utilization rate a_{ij} of resources for each finished good. In addition, the plant that produces the goods has a limited amount of each resource i ($i = 1, \dots, m$),

denoted by b_i . We can formulate a deterministic mathematical program for a single scenario s (the scenario subproblem, or SP) as follows:

SP :

$$z^s = \min \sum_{j=1}^n c_j^s x_j \quad (1)$$

$$\text{Subject to: } \sum_{j=1}^n a_{ij}^s x_j = b_i^s \quad \text{for } i = 1, \dots, m \quad (2)$$

$$x_j \geq 0 \quad \text{for } j = 1, \dots, n \quad (3)$$

where c^s , a^s , and b^s respectively represent the realization of the cost coefficient, the resource utilization, and the resource availability data under scenario s . Consider, for example, a company that manufactures a certain type of maple door. Depending on the weather in the region where the wood for the doors is obtained, the cost of raw materials and transportation will vary. The company is also considering whether to expand production capacity at the facility where doors are manufactured, so a total of six scenarios must be considered. The six possible scenarios and associated parameters for maple doors are shown in Table 1.

Therefore, model SP needs to be solved once for each of the six scenarios.

The scenario optimization approach can be summarized in two steps:

1. Compute the optimal solution to each deterministic scenario subproblem SP .
2. Solve a tracking model to find a single, feasible decision for all scenarios.

The key aspect of scenario optimization is the function of the tracking model in step 2. For illustration we introduce a simple form of tracking model. Let p^s denote the estimated probability for the occurrence of scenario s . Then, a simple tracking model for our problem can be formulated as follows:

$$\begin{aligned} \text{Minimize } & \sum_s p^s \left(\sum_j c_j^s x_j - z^s \right)^2 \\ & + \sum_s p^s \left(\sum_{ij} a_{ij}^s x_j - b_i^s \right)^2 \end{aligned} \quad (4)$$

$$\text{Subject to: } x_j \geq 0 \quad \text{for } j = 1, \dots, n \quad (5)$$

Table 1 Possible scenarios for maple doors

Scenario	Capacity	$P(C)(\%)$	Weather	$P(W)(\%)$	$P(\text{scenario})$	Cost c_j	Utilization a_{ij}	Availability b_j
1	Current	50	Dry	33	1/6	Low	High	Low
2			Normal	33	1/6	Medium	Low	Low
3			Wet	33	1/6	High	Low	Low
4	Expanded	50	Dry	33	1/6	Low	High	High
5			Normal	33	1/6	Medium	Low	High
6			Wet	33	1/6	High	Low	High

The purpose of this tracking model is to find a solution that is feasible under all the scenarios, and it penalizes solutions that differ greatly from the optimal solution under each scenario. The two terms in the objective function are squared to ensure nonnegativity. More sophisticated tracking models can be used for various different purposes. In risk management, for instance, we may select a tracking model that is designed to penalize performance below a certain target level.

Robust Optimization

Robust optimization may be used when the parameters of the optimization problem are known only within a finite set of values. The robust optimization framework gets its name because it seeks to identify a robust decision – i.e., a solution that performs well across many possible scenarios.

In order to measure the robustness of a given solution, different criteria may be used. Kouvelis and Yu identify three criteria: (a) absolute robustness, (b) robust deviation, and (c) relative robustness. We illustrate the meaning and relevance of these criteria by describing the robust optimization approach described in [2].

Consider an optimization problem where the objective is to minimize a certain performance measure such as cost. Let S denote the set of possible data scenarios over the planning horizon of interest. Also, let X denote the set of decision variables and P denote the set of input parameters of our decision model. Correspondingly, we let P^s identify the value of the parameters belonging to scenario s , and let F^s identify the set of feasible solutions to scenario s . The optimal solution to a specific scenario s is then

$$z^s = f(X_s^*, P^s) = \min_{X \in F^s} f(X, P^s) \quad (6)$$

We assume here that f is convex. The first criterion, absolute robustness, also known as *worst-case optimization*, seeks to find a solution that is feasible for all possible scenarios and optimal for the worst possible scenario. In other words, in a situation where the goal is to minimize the cost, the optimization procedure will seek the robust solution, z^R , that minimizes the cost of the maximum-cost scenario. We can formulate this as an objective function of the form

$$z^R = \min \left\{ \max_{s \in S} f(X, P^s) \right\} \quad (7)$$

The second criterion, robust deviation, is the one that minimizes the largest deviation from optimality. In other words, we seek to minimize the worst-case deviation that separates the robust solution from the optimal solution to each scenario. The third criterion, relative robustness, is similar to the second, but instead of a direct deviation measure, we use a relative deviation measure with respect to optimality. Suppose for convenience we reformulate the scenario subproblem introduced in the previous section as follows (based on [2, pp. 26–29]):

$$z^s = \min c^s x \quad (8)$$

$$\text{Subject to: } A^s x = b^s \quad (9)$$

$$x \geq 0 \quad (10)$$

Then, the robust deviation problem would be formulated as follows:

$$\text{Minimize } y \quad (11)$$

$$\text{Subject to: } c^s x \leq y + z^s \quad (12)$$

$$A^s x = b^s \quad (13)$$

$$x \geq 0 \quad (14)$$

Variations to this basic framework have been proposed in [3] to capture the risk-averse nature

of decision makers by introducing higher moments of the distribution of z^s in the optimization model and implementing weights as penalty factors for infeasibility of the robust solution with respect to certain scenarios.

Simulation Optimization

Until quite recently, the methods available for finding optimal decisions have been unable to effectively handle the complexities and uncertainties posed by many real-world situations of the form treated by simulation. The areas of scenario optimization and robust optimization described in the previous sections attempt to deal with some of these, but the modeling framework limits the range of situations that can be tackled with such technology.

The complexities and uncertainties that render real-world systems mathematically intractable are the primary reason that simulation is often chosen as a basis for handling decision problems associated with those systems. On one hand, the set of possible scenarios is often unknown *a priori*. On the other hand, the possible combinations of parameters is too numerous to be handled efficiently by methods such as those previously described. Consequently, decision makers must deal with the dilemma that many important types of optimization problems can only be treated by the use of simulation models, but once these problems are submitted to simulation there are no optimization methods that can adequately cope with them.

Recent developments are changing this picture. The field of metaheuristics – the domain of optimization that augments traditional mathematics with artificial intelligence and methods derived from physical, biological, or evolutionary analogs – has witnessed advances yielding effective engines for guiding a series of complex evaluations in the quest for optimal values for the decision variables, as described in [4–9].

Modern simulation optimization tools are designed to solve optimization problems of the following form:

$$\text{Minimize} \quad F(x) \quad (\text{objective function}) \quad (15)$$

$$\text{Subject to:} \quad Ax \leq b \quad (\text{constraints}) \quad (16)$$

$$g_l \leq G(x) \leq g_u \quad (\text{requirements}) \quad (17)$$

$$l \leq x \leq u \quad (\text{bounds}) \quad (18)$$

where the vector x of decision variables includes variables that range over continuous values and variables that only take on discrete values (both integer values and values with arbitrary step sizes).

The objective function $F(x)$ is typically highly complex; under the context of simulation optimization $F(x)$ represents a performance measure output by the simulation, and may be any mapping from a set of values x to a real value. The constraints represented by inequality $Ax \leq b$ are linear (given that nonlinearity in the model is embedded within the simulation itself), and the coefficient matrix A and the right-hand-side values corresponding to vector b must be known. The requirements $g_l \leq G(x) \leq g_u$ impose simple upper and/or lower bounds on a function that can be linear or nonlinear and is an output of the simulation. The values of the bounds g_l and g_u must be known constants. All the variables are bounded and some may be restricted to be discrete, as previously noted. Each evaluation of $F(x)$ and $G(x)$ requires a simulation of the system. By combining simulation and optimization, a powerful design tool results.

Optimizers designed for simulation embody the principle of separating the method from the model. In such a context, the optimization problem is defined outside the complex system. Therefore, the evaluator (i.e., the simulation model) can change and evolve to incorporate additional elements of the complex system, while the optimization routines remain the same. Hence, there is a complete separation between the model that represents the system and the procedure that is used to solve optimization problems defined within this model.

The optimization procedure uses the outputs from the system evaluator, which measures the merit of the inputs that were fed into the model. On the basis of both current and past evaluations, the method decides upon a new set of input values (see Figure 1).

Provided that a feasible solution exists, the optimization procedure ideally carries out a special search

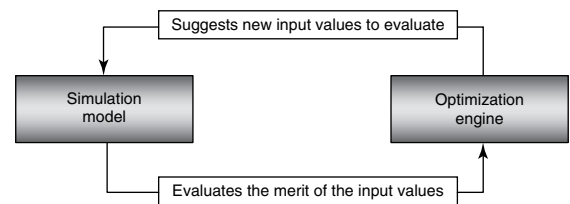


Figure 1 Coordination between optimizer and simulator

where the successively generated inputs produce varying evaluations, not all of them improving, but which over time provide a highly efficient trajectory to the globally best solutions. The process continues until an appropriate termination criterion is satisfied (usually based on the user's preference for the amount of time devoted to the search).

The simulation optimization framework is very flexible in terms of the performance measures the decision maker wishes to evaluate. In fact, the only limitation is not on the side of the optimization engine, but on the simulation model's ability to evaluate performance based on specified values for the decision variables. In order to provide in-depth insights into the use of simulation optimization in the context of risk management, we present a practical application through the use of an illustrative example in the context of optimal portfolio selection.

Example: Selecting Optimal Portfolios of Projects

The energy industry uses project portfolio optimization to manage investments in exploration and production, as well as power plant acquisitions [10, 11]. Decision makers typically wish to maximize the return on invested capital, while controlling the exposure of their portfolio of projects to various risk factors.

The following example involves a company that has 61 potential projects in its investment funnel. Each type of project requires an initial investment and a certain number of business development, engineering, and earth sciences personnel. The company has a budget limit for its investment opportunities, and a limited number of personnel in each skill category. Projects may start in different time periods, but there is a restricted window of opportunity of up to 3 years for each project. The company must select a set of projects to invest in that will best further its corporate goals.

Probably the best-known model for portfolio optimization is rooted in the work of Nobel laureate Harry Markowitz. The model, called the *mean-variance model* [12], is based on the assumption that the expected portfolio returns will be normally distributed. The model seeks to balance risk and return in a single objective function, as follows.

Given a vector of portfolio returns r , and a covariance matrix Q of returns, then we can formulate

the model as follows:

$$\text{Maximize} \quad r^T w - k w^T Q w \quad (19)$$

$$\text{Subject to:} \quad \sum_i c_i w_i = b \quad (20)$$

$$w \in \{0, 1\} \quad (21)$$

where k represents a coefficient of the firm's risk aversion, c_i represents the initial investment in project i , w_i is a binary variable representing the decision whether to invest in project i , and b is the available budget. We will use the mean-variance model as a base case for the purpose of comparing with other selected models of portfolio performance.

To facilitate our analysis, we make use of the OptFolio® software that combines simulation and optimization into a single system specifically designed for portfolio optimization [13–15].

We examine three cases to demonstrate the flexibility of this method to enable a variety of decision alternatives that significantly improve upon traditional mean-variance portfolio optimization, and illustrate the flexibility afforded by simulation optimization approaches in terms of controlling risk. The results also show the benefits of managing and efficiently allocating scarce resources like capital, personnel, and time. The weighted average cost of capital, or annual discount rate, used for all cases was 12%.

Case 1: Mean-Variance Approach

In this first case, we implement the mean-variance portfolio selection method of Markowitz described above. The decision is to determine participation levels (0 or 1) in each project with the objective of maximizing the expected net present value (NPV) of the portfolio while keeping the standard deviation of the NPV below a specified threshold of \$140 M. We denote the expected value of the NPV by μ_{NPV} , and the standard deviation of the NPV by σ_{NPV} . This case can be formulated as follows:

$$\text{Maximize} \quad \mu_{NPV} \quad (\text{objective function}) \quad (22)$$

$$\text{Subject to:} \quad \sigma_{NPV} < \$140 \text{ M} \quad (\text{requirement}) \quad (23)$$

$$\sum_i c_i x_i \leq b \quad (\text{budget constraint}) \quad (24)$$

$$\sum_i p_{ij} x_i \leq P_j \quad \forall j$$

(personnel constraints) (25)

All projects must start in year 1

$$x_i \in \{0, 1\} \quad \forall i \text{ (binary decisions)} \quad (26)$$

The optimal portfolio has the following performance metrics:

$$\begin{aligned} \mu_{NPV} &= \$394 \text{ M}, \\ \sigma_{NPV} &= \$107 \text{ M}, \\ P(5)_{NPV} &= \$176 \text{ M} \end{aligned} \quad (27)$$

where $P(5)_{NPV}$ denotes the 5th percentile of the resulting NPV probability distribution (i.e., the probability of the NPV being lower than the $P(5)$ value is 5%).

The bound imposed on the standard deviation in equation (23) does not seem binding. However, owing to the binary nature of the decision variables, no project additions are possible without violating the bound. Figure 2 shows a graph of the probability distribution of the NPV obtained for 1000 replications of this base model. The thin line represents the expected value.

Case 2: Risk Controlled by 5th Percentile

In the context of risk management, statistics such as variance or standard deviation of returns are not always easy to interpret and other measures may be more intuitive and useful. For example, it is clearer to say, “there is a 95% chance that the portfolio return is above some value X” than to say that the standard deviation is \$107M. This measure can be easily implemented in a simulation optimization procedure

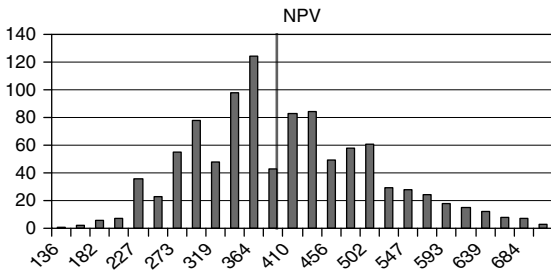


Figure 2 Mean-variance portfolio

by imposing a requirement on the 5th percentile of the resulting distribution of returns. In case 2, the decision is to determine participation levels (0, 1) in each project with the objective of maximizing the expected NPV of the portfolio, while keeping the 5th percentile of the NPV distribution above the value determined in case 1. This is achieved by imposing the requirement in equation (30) on the model below. In other words, we want to find the portfolio that produces the maximum average return, as long as no more than 5% of the observations fall below \$176M. Also, in this case we allow for delays in the start dates of projects, according to windows of opportunity defined for each project. In order to achieve this, we have created copies of a project that are shifted by one, two or three periods into the future (according to the windows of opportunity defined for each project). Mutual exclusivity clauses ensure that only one start date for each project is selected. For example, to represent the fact that Project A can start at time $t = 0, 1$ or 2, we include the following mutual exclusivity clause as a constraint:

$$\text{Project } A_0 + \text{Project } A_1 + \text{Project } A_2 \leq 1 \quad (28)$$

The subscript following the project name corresponds to the allowed start dates for the project, and the constraint only allows at most one of these to be chosen.

$$\text{Maximize} \quad \mu_{NPV} \quad (29)$$

$$\text{Subject to:} \quad P(5)_{NPV} > \$176 \text{ M} \quad (\text{requirement}) \quad (30)$$

$$\sum_i c_i x_i \leq b \quad (\text{budget constraint}) \quad (31)$$

$$\sum_i p_{ij} x_i \leq P_j \quad \forall j$$

(personnel constraints) (32)

$$\sum_{m_i \in M} x_i \leq 1 \quad \forall i$$

(mutual exclusivity) (33)

$$x_i \in \{0, 1\} \quad \forall i \quad (\text{binary decisions}) \quad (34)$$

where m_i denotes the set of mutually exclusive projects related to project i . In this case we have replaced the standard deviation with the 5th percentile

as a measure of risk containment. The resulting portfolio has the following attributes:

$$\begin{aligned}\mu_{NPV} &= \$438 \text{ M}, \\ \sigma_{NPV} &= \$140 \text{ M}, \\ P(5)_{NPV} &= \$241 \text{ M}\end{aligned}\quad (35)$$

By using the 5th percentile instead of the standard deviation as a measure of risk, we are able to obtain an outcome that shifts the distribution of returns to the right, compared to case 1, as shown in Figure 3.

This case clearly outperforms case 1. Not only do we obtain significantly better financial performance, but we also achieve a higher personnel utilization rate, and a more diverse portfolio.

Case 3: Probability Maximizing and Value at Risk

In case 3, the decision is to determine participation levels (0, 1) in each project with the objective of maximizing the probability of meeting or exceeding the mean NPV found in case 1. This objective is expressed in equation (36) of the following model.

$$\text{Maximize} \quad P(NPV \geq \$394 \text{ M}) \quad (36)$$

$$\text{Subject to:} \quad \sum_i c_i x_i \leq b \text{ (budget constraint)} \quad (37)$$

$$\sum_i p_{ij} x_i \leq P_j \quad \forall j$$

(personnel constraints) (38)

$$\sum_{m_i \in M} x_i \leq 1 \quad \forall i$$

(mutual exclusivity) (39)

$$x_i \in \{0, 1\} \quad \forall i \text{ (binary decisions)} \quad (40)$$

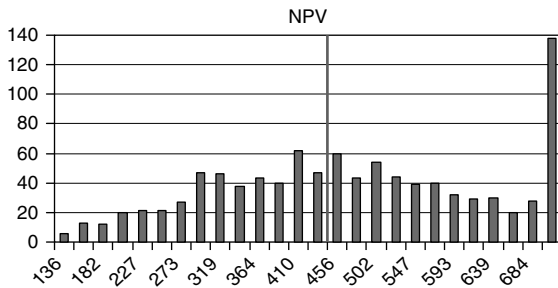


Figure 3 Portfolio risk controlled by $P(5)$

This case focuses on maximizing the chance of achieving a goal and essentially combines performance and risk containment into one metric. The probability in equation (36) is not known *a priori*, so we must rely on the simulation to obtain it. The resulting optimal solution yields a portfolio that has the following attributes:

$$\begin{aligned}\mu_{NPV} &= \$440 \text{ M}, \\ \sigma_{NPV} &= \$167 \text{ M}, \\ P(5) &= \$198 \text{ M}\end{aligned}\quad (41)$$

Although this portfolio has a performance similar to the one in case 2, it has a 70% chance of achieving or exceeding the NPV goal. As can be seen in the graph of Figure 4, we have succeeded in shifting the probability distribution even further to the right, therefore increasing our chances of exceeding the returns obtained with the traditional Markowitz approach. In addition, in cases 2 and 3, we need not make any assumption about the distribution of expected returns.

As a related corollary to this last case, we can conduct an interesting analysis that addresses value at risk (VaR). In traditional (securities) portfolio management, VaR is defined as the worst expected loss under normal market conditions over a specific time interval and at a given confidence level. In other words, VaR measures how much the holder of the portfolio can lose with probability equal to α over a certain time horizon [16]. In the case of project portfolios, we can define VaR as the probability that the NPV of the portfolio will fall below a specified value. Going back to our present case, the manager may want to limit the probability of incurring negative returns. In this example, we formulate the problem in a slightly different way: we

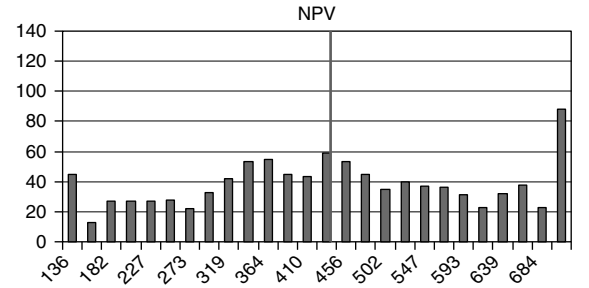


Figure 4 Probability-maximizing portfolio

still want to maximize the expected return, but we limit the probability that we incur a loss to $\alpha = 1\%$ by using the requirement shown in equation (43) as follows:

$$\text{Maximize} \quad \mu_{NPV} \quad (42)$$

$$\text{Subject to:} \quad P(NPV < 0) \leq 1\% \text{ (requirement)} \quad (43)$$

$$\sum_i c_i x_i \leq b \quad \text{(budget constraint)} \quad (44)$$

$$\sum_i p_{ij} x_i \leq P_j \quad \forall j \quad \text{(personnel constraints)} \quad (45)$$

$$\sum_{m_i \in M} x_i \leq 1 \quad \forall i \quad \text{(mutual exclusivity)} \quad (46)$$

$$x_i \in \{0, 1\} \quad \forall i \quad \text{(binary decisions)} \quad (47)$$

The portfolio performance under this scenario is

$$\mu_{NPV} = \$411 \text{ M}, \sigma_{NPV} = \$159 \text{ M}, P(5) = \$195 \text{ M} \quad (48)$$

These results from the VaR model turn out to be slightly inferior to the case where the probability was maximized. This is not a surprise, since the focus is on limiting the probability of downside risk, whereas before, the goal was to maximize the probability of obtaining a high expected return. However, this last analysis could prove valuable for a manager that wants to limit the VaR of the selected portfolio. As shown here, for this particular set of projects, a very good portfolio can be selected with that objective in mind.

Conclusions

Practically every real-world situation involves uncertainty and risk, creating a need for optimization methods that can handle uncertainty in model data and input parameters. We have briefly described two popular methods, *scenario optimization* and *robust optimization*, that seek to overcome limitations of classical optimization approaches for dealing with uncertainty, and which undertake to find high-quality

solutions that are feasible under as many scenarios as possible. However, these methods are unable to handle problems involving moderately large numbers of decision variables and constraints, or involving significant degrees of uncertainty and complexity. In these cases, simulation optimization is becoming the method of choice. The combination of simulation and optimization affords all the flexibility of the simulation engine in terms of defining a variety of performance measures as desired by the decision maker. In addition, as we demonstrate through a project portfolio selection example, modern optimization engines can enforce requirements on one or more outputs from the simulation, a feature that scenario-based methods cannot handle. Finally, simulation optimization produces results that can be conveyed and grasped in an intuitive manner, providing an especially useful tool for identifying improved business decisions under risk and uncertainty.

References

- [1] Dembo, R. (1991). Scenario optimization, *Annals of Operations Research* **30**, 63–80.
- [2] Kouvelis, P. & Yu, G. (1997). *Robust Discrete Optimization and Its Applications*, Kluwer, Dordrecht, pp. 8–29.
- [3] Mulvey, J.M., Vanderbei, R.J. & Zenios, S.A. (1995). Robust optimization of large-scale systems, *Operations Research* **43**(2), 264–281.
- [4] Campos, V., Laguna, M. & Martí, M. (1999). Scatter search for the linear ordering problem, *New Methods in Optimization*, McGraw-Hill, New York, pp. 331–339.
- [5] Glover, F. (1998). *A Template for Scatter Search and Path Relinking, Artificial Evolution, Lecture Notes in Computer Science 1363*, Springer-Verlag, New York, pp. 13–54.
- [6] Glover, F. & Laguna, M. (1997). *Tabu Search*, Kluwer, Norwell.
- [7] Glover, F., Laguna, M. & Martí, R. (2000). Fundamentals of scatter search and path relinking, *Control and Cybernetics* **29.3**, 653–684.
- [8] Glover, F., Laguna, M. & Martí, R. (2003). *Scatter Search, Advances in Evolutionary Computing: Theory and Applications*, Springer-Verlag, New York, pp. 519–537.
- [9] Laguna, M. & Martí, R. (2003). *Scatter Search: Methodology and Implementations in C*, Kluwer, Norwell, pp. 217–254.
- [10] Haskett, W.J. (2003). Optimal appraisal well location through efficient uncertainty reduction and value of information techniques, in *Proceedings of the Society of Petroleum Engineers Annual Technical Conference and Exhibition*, Denver.

- [11] Haskett, W.J., Better, M. & April, J. (2004). Practical optimization: dealing with the realities of decision management, in *Proceedings of the Society of Petroleum Engineers Annual Technical Conference and Exhibition*, Houston.
- [12] Markowitz, H. (1952). Portfolio selection, *Journal of Finance* **7**(1), 77–91.
- [13] April, J., Glover, F. & Kelly, J.P. (2002). Portfolio optimization for capital investment projects, *Proceedings of the 2002 Winter Simulation Conference*, December 8–11, San Diego, pp. 1546–1554.
- [14] April, J., Glover, F. & Kelly, J.P. (2003). Optfolio – a simulation optimization system for project portfolio planning, *Proceedings of the 2003 Winter Simulation Conference*, December 7–10, Washington DC, pp. 301–309.
- [15] Kelly, J.P. (2002). Simulation optimization is evolving, *INFORMS Journal of Computing* **14.3**, 223–225.
- [16] Benninga, S. & Wiener, Z. (1998). Value-at-risk (VaR), *Mathematica in Education and Research* **7**(4), 1–7.

Related Articles

Comparative Risk Assessment

MARCO BETTER AND FRED GLOVER

Scientific Uncertainty in Social Debates Around Risk

There are three primary audiences in science policy: scientists, policymakers, and the public [1]. Current issues in environment and health are often clouded with uncertainty. The public is becoming increasingly alarmed over potential adverse health effects stemming from chemical toxicants and industrial wastes being discharged into the environment at all levels (i.e., local, national, and global). They look to scientists to evaluate, assess, and identify potential risks. There is also an expectation that policy makers make informed decisions, on the basis of scientific evidence, to implement the appropriate solutions that will “solve” the environmental problems of the day. However, mismanagement of past risk issues, by both scientists and policy makers, have left the public distrustful (*see Stakeholder Participation in Risk Management Decision Making*).

When risk controversies arise, a power struggle may result between the competing interests, ideologies, values, and approaches employed by the actors involved: industry, scientists, policy makers, the media, and the public. Each group’s conception of what constitutes “safe” may differ. The typical response from scientists and industry during risk controversies is that the public is “irrational”, the media “twists” the scientific information, and policy makers rarely listen to “objective” scientific evidence. Although such sentiments may be superficially plausible, they reflect a misunderstanding of the uncertain nature of living in a technological society and the limitations of risk analysis [2].

Prior to the 1960s, analytical chemistry was based largely on the senses of sight and smell. Since then, considerable technological and disciplinary advances within analytical chemistry have enabled scientists to identify the presence of elements and chemical compounds as low as n parts per billion; *detection limits* (*see Detection Limits*) for some substances have reached as low as parts per quadrillion [3]. Such advances in measurement sensitivity have often led to a redefinition of toxicity in terms of trace elements and compounds. However, this dependence

on apparently precise numbers, which conveys an impression that science is accurate in its estimates, has done comparatively little to objectify the processes for determining when such compounds are toxic in terms of their risk to ecosystems or human health. Such decisions are, in part, the consequence of social and political processes taking place among scientists and policy makers [4]. Thus, what begins as a scientific measurement and analysis problem frequently turns into an arena of interpretation and social debate [3]. These debates typify how uncertainties in science are managed in a policy context.

Science as a Knowledge Producer

The process of scientific discovery is essential in the construction of scientific knowledge claims. These take place in established institutions: laboratories; working groups; disciplines; peer-review committees; journal publications; and scientific prizes, such as the Nobel prize [5]. Although these institutions provide avenues for the development of scientific facts, it is critical that their “consumption” in a social world be analyzed. In other words, it is important to examine what happens once scientific claims leave the laboratories and enter journals, professional meetings, reports [6], regulatory agencies [7], and the media [8]. It is through the consumption of scientific claims that experts seek to establish boundaries as to who can legitimately speak as an authority. “Authority” can only exist when the production of knowledge claims are conferred to some but denied to others [6]. Latour [9] uses the analogy of the two-faced Janus to depict stages of science: ready-made science (where facts are agreed upon) and science in action (where facts and evidence are not immutable and are subject to debate). It is during this latter stage that “credibility contests” occur [6] where the credibility of both the producer of the knowledge claims and of the challenger are equally “on the line” [8].

Although credibility contests may occur privately, (i.e., in an environment where there is a controlled membership, like a scientific meeting or a peer-reviewed journal), it is when these contests move into the public domain that scientists may present a unified position to protect the autonomy of science [6]. A public disagreement can occur once scientific facts are brought out into the open through the courts or in

the media. For example, in the area of climate change, scientists were instrumental in putting the issue on the policy agenda. Although some scientists raised disconfirming evidence, these alternative views were shut out [10].

Scientists seek to convince others of the importance of their work to secure funding, and consequently control the direction of the research agenda [5]. Issues of uncertainty are pivotal to this process. In a science-policy domain, there needs to be a balance in the level of uncertainty: too much and the problem seems daunting; too little and there is no motivational push for research funds dedicated to study the issue [11]. Consequently, “uncertainty” becomes both a social and cultural process used as a forum for competing ideologies and interests [12].

From a risk perspective, risks need to be first identified and measured by science before they can gain social recognition [2]. Hence, risks are “invisible” (i.e., not objective forms of knowledge) until they are measured. These knowledge claims then have to be translated within a policy domain, where only that knowledge that is made “meaningful” gains political attention. Salter [7] refers to these processes as “mandated science”; in other words, scientific evidence is used in policy making.

The Uptake of Science to Policy

There is insufficient understanding of how environmental health policy decisions are made in the face of scientific uncertainty (*see Environmental Health Risk*). It is felt that only through the development of more precise methodologies, the continuous replication of results, and increasingly rigorous studies science will be able to resolve much of the uncertainty, possibly making the task of the policy makers easier. Similarly, there is some debate as to whether better science will lead to greater certainty in decision making. The tendency in science is to reduce phenomena to examinable components in the search for specific cause and effect relationships. This is not an easy task in environmental health where there is a multitude of exposures, confounders, and effects. It becomes much more difficult to determine if certain health effects are the result of a dominant cause or if they are due to a combination of environmental assaults [13]. As such, social considerations of what constitute important exposures and acceptable

risk need to be incorporated into regulatory decisions. One way to understand the interrelationships between various actors in the policy subsystem is through an examination of “epistemic communities” [14].

Conceptually, “epistemic communities” evolved as a means to explain the translation of knowledge and evidence into policy action within a complex international environmental arena. Defined as a network of scientific experts with an authoritative claim to policy-relevant knowledge [14], an epistemic community provides consensual knowledge regarding issues of uncertainty, thereby facilitating policy intervention. Jasanoff and Wynne [15, p. 51] outline four defining characteristics of an epistemic community:

- shared normative and principled beliefs providing a value-based rationale for the community’s proposed social actions;
- shared causal beliefs, including professional judgments linking causal explanations to possible policy actions;
- shared notions of validity including intersubjective, internally defined criteria for establishing valid knowledge; and
- a shared policy endeavor based on a commonly recognized set of problems to be solved.

Common features of these characteristics involve shared values and ideas for preferred policy solutions to identified problems. In other words, “knowledge” and “facts” are viewed and accepted in specifically defined ways according to preset criteria for inclusion and exclusion (e.g., the scientific method). Epistemic communities are often able to define problems in a particular light making some solutions more attractive to decision makers [16]. In an epistemic community “membership” and “behavior” are peer controlled and monitored [9].

Critiques of the concept of “epistemic communities” argue that the role of “power” in the policy arena is typically ignored [17]. There is an assumption that it will only be “expert knowledge” that will sway policy decisions. Yet, policy decisions are usually made on the basis of a number of reasons with expert knowledge being the only one input. Moreover, scientific facts are often not disputed; it is the interpretation of those facts that lead to dissenting opinions and uncertainty. Jamieson argues that science is culturally viewed as a knowledge producer, and consequently, often it cannot bring public decisions to closure. Too

frequently, issues of uncertainty are paramount, often pitting experts against experts – hence the need to control the level of “uncertainty” in both science and policy. The other main critique is that, in categorizing a group of scientists as an epistemic community, shared beliefs that may transfer between communities in a policy subsystem may be masked. The assumption is that communities are homogenous, when in reality, there can be a number of differences within a community group and similarities with members in different epistemic communities [17]. However, by adopting more interpretive approaches, some of these differences can be revealed.

Policy analysis to date has largely been dominated by a rational and incrementalist approach. A rational approach to policy analysis assumes that the intended policy goals are clear and agreed upon, the means of evaluation are well defined, and information about the consequences is complete. Rational policy approaches have been criticized as reflecting an extremely naive view of reality. They make the assumption that decisions are made in a value-free, linear, and sequential manner by completely informed individuals [18]. An alternative to such approaches is incrementalism, where policy decisions are made in small cumulative steps. Incrementalism does not assume a value-free approach. Rather, it recognizes that values play a role in how options and alternatives are assessed. Moreover, there is an understanding that “perfect” information does not exist [19]. However, it is the interpretive approaches that have been gaining increasing attention in the policy literature [20]. Interpretive approaches focus on language and meaning as ways to examine the underlying social context of an issue. Interpretive approaches respond to researchers’ needs for understanding how a policy is “framed” or what it “means” [20]. They also involve an analysis of the role that members of a policy subsystem may play in problem definition, as well as how science and policy become coproducers of relevant knowledge [15].

A useful conceptual framework for understanding the construction and contest of scientific knowledge claims is “agenda-setting” – that is, how issues move on or off an agenda at any given time. Largely informed by the policy literature, agenda-setting analyses can assist in the examination of scientific discourses surrounding a public policy issue characterized by uncertainty. Kingdon [16] outlines three factors that influence the agenda-setting process; these have been modified for a scientific discourse

(see [11]). The first is problem recognition. In any given moment, there is competition of ideas within a policy arena. Those ideas that gain recognition move higher up the agenda to provide a focal point for different science and policy actors. Sometimes, problems are recognized by “focusing events”, which will draw attention to the issue through a key scientific study and its uptake by a regulatory body. Second, in a regulatory arena, scientific analyses are requested to address identified policy problems. However, not all analyses are given the same level of attention, nor do they adequately address the policy problem as it is in the process of being defined by others. If the scientific analysis addresses the identified problem, it will determine for how long that science remains relevant. New scientific knowledge may provide a better understanding of the problem, thereby “tip-ping” the issue into the public/political consciousness. This process is by no means simple. The science must resonate with the values and assumptions of the decision makers for it to remain relevant. The political process itself is the third factor that may influence the science agenda. The “national mood”, public opinion polls, election results, or changes in the administration will influence the nature and level of attention that an issue may receive on the agenda [16] as well as the funding that may be attached to some research agendas [21].

When Science and Policy Meet the Public

Throughout processes of scientific uncertainty and the uptake (or not) of scientific evidence in a policy domain, the role played by the public is often obscured. By and large, the public is viewed as innumerate and scientifically illiterate. However, the public has a greater understanding of the role of science in understanding risk than scientists have of how public attitudes and beliefs are formed within a political participatory system. Considerable research has been dedicated to the public understanding of science (*cf.* [15, 21, 22]), including an academic journal of the same name. This body of literature has shown that tensions lie when the public, operating in a “demosphere” [23], do not fully trust or accept potentially dangerous or uncertain scientific evidence arising from studies in the “technosphere” [15]. It is not that the public are irrational, rather that they incorporate multiple considerations – social,

cultural, political – beyond scientific evidence when making decisions about what is “risky” (see [24]). This has led some to call for the coproduction of shared knowledge between scientists and lay audiences [25]. These processes are not without their challenges and much more research is needed to identify optimum strategies to effect positive and meaningful public participation in the face of science-policy uncertainty.

Summary

When existing scientific understandings are contested or different scientific evidence are used within a policy arena, they enter the domain of mandated science [7], where objective rationality cannot be maintained. Policy makers can choose to use scientific uncertainty as a means to justify a decision that runs counter to the evidence, or equally, uncertainty can be used to justify that no policy decision is needed. For the public, lay understandings of a risk situation, based on personal experiences, oral histories, or the media, may factor into their decision making process. Hence, although there is often a recognition among the public that scientific evidence is a necessary condition, it is not sufficient. As Leiss [26] contends, presently there is a divide between quantitative risk assessments – which reflect a presentation of risk information – and public understanding of those risks, a divide that needs to be addressed through the development of good risk communication practice. Part of this process involves knowledge translation of scientific evidence in plain language; sensitivity to the different sets of “knowledge” (social, cultural, and local) that the public uses to understand risk situations; and concerted efforts to work together as partners to develop shared understanding – both expert and lay.

References

- [1] Throgmorton, J.A. (1991). The rhetorics of policy analysis, *Policy Sciences* **24**, 153–179.
- [2] Beck, U. (1992). *Risk Society*, Sage Publications, Thousand Oaks.
- [3] Harris, W.E. (1992). Analyses, risks, and authoritative misinformation, *Analytical Chemistry* **64**(13), 665A–671A.
- [4] Kunreuther, H. & Slovic, P. (1996). Science, values, and risk, *The Annals of the American Academy of Political and Social Science* **545**, 116–125.
- [5] Schneider, A.L. & Ingram, H. (1997). *Policy Design for Democracy*, University of Kansas Press, Lawrence.
- [6] Gieryn, T. (1999). *Cultural Boundaries of Science: Credibility on the Line*, University of Chicago, Chicago.
- [7] Salter, L. (1988). *Mandated Science*, Kluwer Academic Publishers, Norwell.
- [8] Hilgartner, S. (2000). *Science on Stage: Expert Advice as Public Drama*, Stanford University Press, Stanford.
- [9] Latour, B. (1987). *Science in Action*, Harvard University Press, Cambridge.
- [10] Garvin, T. & Eyles, J. (1997). The sun safety metanarrative: translating science into public health discourse, *Policy Sciences* **30**, 47–70.
- [11] Driedger, S.M., Eyels, J., Elliott, S.D. & Cole, D.C. (2002). Constructing scientific authorities: issue framing of chlorinated disinfection byproducts in public health, *Risk Analysis* **22**(4), 789–802.
- [12] Aronson, N. (1984). Science as a claims-making activity: implications for social problems research, in *Studies in the Sociology of Social Problems*, J.A. Schneider & J. Kitsuse, eds, Ablex Publications, Norwood, pp. 1–30.
- [13] Bates, D. (1994). *Environmental Health Risk and Public Policy: Decision-Making in Free Societies*, University of British Columbia Press, Vancouver.
- [14] Haas, P. (1992). Introduction: epistemic communities and international policy coordination, *International Organization* **46**, 1–35.
- [15] Jasanoff, S. & Wynne, B. (1998). Science and decision-making, in *Human Choice and Climate Change, Volume 1: The Societal Framework*, S. Rayner & E.L. Malone, eds, Battelle Press, Columbus.
- [16] Kingdon, J.W. (2003). *Agendas, Alternatives, and Public Policies*, Little, Brown, Toronto.
- [17] Litfin, K.T. (1994). *Ozone Discourses: Science and Politics in Global Environmental Cooperation*, Columbia University Press, New York.
- [18] Hogwood, B. & Gunn, L. (1984). *Policy Analysis for the Real World*, Oxford University Press, Oxford.
- [19] Lindblom, C.E. (1959). The science of ‘muddling through’, *Public Administration Review* **19**, 79–88.
- [20] Yanow, D. (1992). Silences in public policy discourse: organizational and policy myths, *Journal of Public Administration Research and Theory* **2**, 399–423.
- [21] Irwin, A. & Wynne, B.E. (1996). *Misunderstanding Science? The Public Reconstruction of Science and Technology*, Cambridge University Press, Cambridge.
- [22] Cunningham-Burley, S. (2006). Public knowledge and public trust, *Community Genetics* **9**(3), 204–210.
- [23] Krinsky, S. & Alonzo, P. (1988). *Environmental Hazards: Communicating Risks as a Social Process*, Auburn House, Westport.
- [24] Garvin, T. (2001). Analytical paradigms: the epistemological distances between scientists, policy makers and the public, *Risk Analysis* **21**(3), 443–455.

- [25] Corburn, J. (2007). Community knowledge in environmental health science: co-producing policy expertise, *Environmental Science and Policy* **10**, 150–161.
- [26] Leiss, B. (2004). Effective risk communication practice, *Toxicology Letters* **149**, 399–404.

Risk and the Media

Role of Risk Communication in a Comprehensive Risk Management Approach

S. MICHELLE DRIEDGER

Related Articles

Considerations in Planning for Successful Risk Communication

Securitization/Life

Securitization can be defined as the isolation of a pool of assets (or liabilities or the rights to a set of cash flows) and the repackaging of those assets (or liabilities or cash flow rights) into securities that are traded in capital markets (*see Asset–Liability Management for Life Insurers*). Insurance companies have traditionally been seen as risk warehouses – they would take on insurance risks and bear them themselves – and securitization allows them to move toward a model in which they become risk intermediaries instead – that is, they originate risks, but repackage them and sell them on. The risk-intermediary model of an insurance company has many attractions over the warehouse model, because it is often more efficient for many types of traditional insurance risks to be traded in capital markets rather than held on-balance sheet in risk warehouses, with consequential capital or reserving requirements.

Securitization began in the 1970s when banks in the United States began to sell off pools of mortgage-backed loans. In the 1980s, they then started to sell off other assets, such as automobile, credit card and home equity loans. The amounts involved in these asset-backed security (ABS) arrangements grew rapidly, and by 2002 some \$450 billion dollars worth of assets had been securitized [1]. By comparison, insurance companies have been very slow to catch on to securitization. The first insurance company securitizations related to property and casualty, most notably the catastrophe (“cat”) bonds developed in the 1990s: these securitized extreme property losses arising from natural disasters such as earthquakes and hurricanes. The first life securitizations took place in the late nineties, and took the form of sales of rights to future profits from blocks of life policies and annuities. The pace of innovation in life securitization then started to pick up, and recent developments include the launch of reserve funding (or capital arbitrage) securitizations, mortality and longevity bonds, and, most recently, pension bulk buyout securitizations. However, the amounts involved in life securitizations are still very small compared to those involved in bank securitizations.

The Mechanics of Securitization

The basic idea behind securitization is to move an asset off-balance sheet. (We focus on asset

securitizations for convenience, but similar principles apply to liability securitizations.) This is usually done by transferring the asset (or liability) to a special purpose vehicle (SPV). This SPV is a passive entity that is specially created to house the asset and issue securities using the asset as collateral. The SPV then issues securities to investors who contribute funds to the SPV. The SPV also remits some or all of these funds to the originating institution, which in return transfers the asset or the rights to the cash flows generated by the asset to the SPV. Many of these ABS arrangements involve tranching – the SPV issues different classes of security with different claims and different seniority rights to the underlying cash flows. The tranches themselves can be designed to appeal to different classes of investors and mitigate asymmetric information problems, and thus increase the overall value of the securitization [2]. In many cases, the SPV also enters into a swap: for example, if the underlying cash flows from the asset are not interest sensitive, but if it is perceived that the investors want interest-sensitive coupon payments, then an SPV might swap the fixed payments from the underlying asset into floating payments and issue securities with floating coupon payments. The swap counterparty then becomes another participant in the securitization arrangement.

Most of these arrangements also involve credit enhancement features to protect one or more participating parties against default risk. These features include over-collateralization (where the value of the assets transferred to the SPV exceeds the value of the securities it issues), subordination (where the SPV issues securities with varying levels of seniority), and external guarantees such as parental guarantees, letters of credit, credit insurance, and reinsurance. Many SPVs also include an arrangement by which the originating life institution continues to service the original customers. This is especially important in life securitizations where there is a need to ensure that policyholders do not allow their policies to lapse.

Benefits of Securitization

Life securitizations offer many potential benefits to both the companies that engage in them and to outside investors. In general terms, securitization enables companies to unbundle the different functions associated with traditional insurance – and in particular, to separate out the policy origination function from

the investment management, policy servicing, and risk-bearing functions associated with insurance. This enables insurance companies to exploit the benefits of specialization and make more efficient use of capital. Securitization also gives insurance companies access to the capital available in the capital markets, which vastly exceeds the capital available through traditional channels such as *reinsurance* (see **Reinsurance**). Securitization also offers insurance companies new hedging tools and the potential for better risk management. In addition, it helps make insurance-related positions more marketable and therefore more liquid, and improves the transparency of on-balance sheet positions that have traditionally been very opaque.

Securitization also offers insurance companies more specific benefits. It allows companies to unlock embedded value, monetize the present value of future cash flows from a block of business, and release capital for other uses. Insurance companies can use securitization to lay off unwanted risks – for example, they might use securitization to exit a particular line of business or hedge against extreme mortality risks – which are then passed on to the capital markets. It can also help firms exploit arbitrage opportunities and, in particular, it can relieve the capital strain imposed by conservative capital requirements on certain lines of business. Other firms with large pension funds can also use life securitizations to lay off the *longevity risks* (see **Longevity Risk and Life Annuities**) in their pension funds, as evidenced by the emerging pension fund buyout market in the United Kingdom; such securitizations are potentially very useful because they can relieve firms of their pension fund risks and leave them free to concentrate on their core businesses.

For their part, investors can benefit from securitization in various ways. The most obvious benefit is that it offers them the opportunity to invest in life-related risks. These provide not only new investment outlets involving mortality and longevity risks that were previously inaccessible to them but also particularly attractive outlets because life risks have little or no correlation with most other risks; life securitizations therefore have low or zero betas relative to traditional financial market risks, and this means that they add little extra risk to a diversified portfolio. Investors can therefore improve the risk-expected return efficiency of their portfolios by investing in life securitizations. Suitably designed life securitizations

can also provide investors with investment opportunities that are tailor-made for their risk appetites and expertise. For example, investors who want exposure to life risks but are not particularly knowledgeable about them might wish to invest in tranches with high seniority, whereas more informed investors might wish to invest in riskier tranches where their superior knowledge gives them the opportunity to earn higher returns.

“Block of Business” Securitizations

The earliest securitizations are “block of business” securitizations. (For a more extensive treatment of these securitizations, see [1].) These have been used to capitalize expected future profits from a block of business, recover embedded values, and exit from a geographical line of business. The last of these motivations is obvious, and the first two arise from the fact that the cost of writing new life policies is usually incurred in the first year of the policy and then amortized over the remainder of its term. This means that writing new business puts pressure on a company’s capital. Securitization helps to relieve this pressure by allowing the company immediate access to its expected future profits, and securitization is an especially attractive option when the company is experiencing rapid growth in a particular line of business. Some examples of these securitizations are a set of 13 transactions carried out by American Skandia Life Assurance Company (ASLAC) over 1996–2000; these were designed to capitalize the embedded values in blocks of variable annuity contracts issued by ASLAC in order to fund continued expansion in the variable annuity business.

Block of business securitizations have also been issued in conjunction with demutualizations. When an insurance company demutualizes, it creates a situation in which the new shareholders have a conflict of interest with the prior participating policyholders (see, e.g. [3]). A natural solution to this conflict of interest is to put prior participating policies into closed blocks and then securitize them.

Regulatory Reserving (XXX) Securitizations

Another form of life securitization is a regulatory reserving securitization, sometimes also known

as *reserving funding* or *XXX securitizations*. These arrangements are designed to give life insurers relief from excessively conservative regulatory reserving or capital requirements, and are used to release capital that can be used to finance new business or reduce the cost of capital. An early example of this type of securitization was a \$300 million deal arranged by First Colony Life Insurance Company through an SPV known as *River Lake Insurance Company* to obtain capital relief under regulation XXX. This regulation imposes excessively conservative assumptions in the calculation of the regulatory reserve requirement on some types of life policies with long-term premium guarantees.

Some traditional solutions to this type of problem are letters of credit or reinsurance, but these can be expensive and sometimes difficult to obtain. Where securitization is used, the investments made into the SPV provide collateral for the term policies written by the insurance company, and thus reduce its XXX regulatory reserves. In return, the insurer would pay premiums into the SPV, and the SPV would enter into a fixed-for-floating swap to exchange the fixed premiums paid by the insurer into the floating payments typically owed to the note-holding investors. Typically, payments would be credit enhanced and the notes would amortize over time as obligations on the underlying policies are gradually paid off. If no adverse events occur, the investors would receive the premiums paid out by the insurer; and if adverse events did occur, then the SPV would release funds to cover the shortfall.

Mortality and Longevity Bonds

Other recent securitizations are mortality and longevity bonds. The former are bonds with at least one payment tied to a mortality index, and their purpose is to protect life insurers and annuity providers against the adverse consequences of an unanticipated extreme deterioration in mortality. The latter are bonds with one or more payments tied to a longevity-related variable such as the survivorship rate of a specified cohort, and are designed to protect insurance companies or annuity providers against an unanticipated increase in longevity. More details on these instruments are provided by Blake *et al.* [4].

The first mortality bond was the Swiss Re 3-year mortality catastrophe bond, which was issued in

December 2003 and matured on 1 January 2007. The bond was designed to reduce Swiss Re's exposure to a catastrophic mortality deterioration (e.g., such as that associated with a repeat of the 1918 Spanish Flu pandemic). The issue size was \$400 million, and investors received quarterly coupons set at 3-month US dollar London Interbank Offered Rate (LIBOR) + 135 basis points. However, the principal repayment was contingent on a specified index of mortality rates across the United States, United Kingdom, France, Italy, and Switzerland. The bond was issued *via* an SPV called *Vita Capital* (VC), which invested the \$400 million principal in high-quality bonds and swapped the income stream on these for a LIBOR-linked cash flow.

The bond helped Swiss Re to unload some of its extreme mortality risk, but it is also likely that Swiss Re was mindful of its credit rating and wanted to reassure rating agencies about its mortality risk management. Furthermore, by issuing the bond themselves, Swiss Re were not dependent on the creditworthiness of other counterparties should an extreme mortality event occur. The bond therefore gave Swiss Re some protection against extreme mortality risk without requiring that the company acquire any credit risk exposure in the process. The Swiss Re bond issue offered a considerably higher return than similarly rated floating-rate securities and was fully subscribed. Investors in the bond also included a number of pension funds. These would have been attracted partly by the higher coupons, and also by the hedging opportunities that it offered, given that a mortality event that would have triggered a reduction in the repayment of the bond would presumably have reduced their own pension liabilities as well.

In April 2005, Swiss Re announced the issue of a second life catastrophe bond with a principal of \$362 million, using a new SPV called *Vita Capital II*. The maturity date is 2010 and the bond was issued in three tranches: Class B (\$62 million), Class C (\$200 million), and Class D (\$100 million). The principal is at risk if, for any two consecutive years before maturity, the combined mortality index exceeds specified percentages of the expected mortality level (120% for Class B, 115% for Class C, and 110% for Class D). This bond was also fully subscribed.

The first longevity bond was announced in November 2004. This was to have been issued by the European Investment Bank (EIB), with BNP Paribas

as the designer and originator and Partner Re as the longevity risk reinsurer. The face value of the issue was to be £540 million and the maturity was 25 years. The bond was to have been an annuity (or amortizing) bond with floating coupon payments, and its innovative feature was to link the coupon payments to a cohort survivor index based on the realized mortality rates of English and Welsh males aged 65 in 2002. The EIB/BNP Paribas/Partner Re Longevity Bond was designed to help pension plans hedge their exposure to longevity risk. However, the bond did not attract sufficient investor interest to be launched, and was subsequently withdrawn. One possible reason for its slow take-up was the likelihood that using the bond as a hedge might have involved a considerable amount of basis risk. Another possibility was that the amount of capital required was high relative to the reduction in risk exposure, and would have made the bond capital-expensive as a risk-management tool.

At the time of writing, it is known that a number of other investment banks are planning to launch new forms of mortality and longevity bond, and we anticipate rapid growth in this area over the near future.

Pension Bulk Buyout Securitizations

A final and very recent class of life securitizations are pension bulk buyouts. The most active market is in the United Kingdom, where a number of buyout funds, such as Paternoster (backed by Deutsche Bank) and Synesis Life (backed by JPMorgan) were set up in 2006 for this purpose.

Annuity books and their matching assets contain an array of risks, which bring an assortment of investment opportunities. The main matching assets for annuity liabilities are corporate and government bonds, the coupons and principal on which are used to make the payments on the annuities. The main risks on the asset side are therefore credit risk and interest rate risk. The key risks facing those paying annuities and pensions are inflation risk (since most annuities and pensions have some inflation uprating) and longevity risk. The key, as with all securitizations, is to sell off unwanted risks and retain those that the company feels most comfortable managing.

The idea behind pension bulk buyout securitizations is, therefore, for the company concerned to pay

a counterparty to take on some or all of the risks in its pension fund. One way to do so is by means of a leveraged buyout (LBO), also known as a *structured* buyout, which has been used for pension schemes in deficit where the sponsor wishes to wind up. To see how they work, consider the following example of Company ABC based in the United Kingdom:

- Value of scheme assets (A) = 100
- Scheme actuary's valuation of liabilities (L) = 120
- Deficit = 20

The company then approaches a buyout fund XYZ, which has been set up as an insurer. Suppose we have the following assessment by XYZ of ABC's pension scheme:

- XYZ's valuation of ABC's liabilities = 160
- XYZ's valuation of ABC's deficit = 60

So, XYZ's more conservative valuation of the deficit is even worse than the scheme actuary's valuation of the deficit of 20.

In the most common version of the LBO bond arrangement, XYZ offers to take on both the assets and liabilities at XYZ's valuation, subject to due diligence. XYZ also lends 60 to ABC to cover the deficit. This loan is typically at bank base rate + 1, 2, or 3% (depending on the credit rating of ABC) and must be paid off over 10 years. The assets are exchanged into a combination of fixed income and index-linked government bonds, using a combination of interest rate, duration, and inflation swaps.

As far as the company is concerned, this arrangement has the advantages of taking its pension liabilities completely off its balance sheet and replacing them by a standard fixed-term loan that everyone understands. The company also escapes the impact of FRS17 volatility on its profits (which would feed through to its reported Profit & Loss, P&L), Pension Protection Fund (PPF) levies, and asset management fees on pension assets.

For its part, the buyout fund gets a good return on its loan and, due to its size in the new issues market for government bonds, can buy the matching bonds at a discount relative to other investors. It also gains the difference of 40 in the valuation of the liabilities. In addition, by pooling the liabilities of different pension funds and buying in the relevant expertise, it can become a better manager of the mortality pool than,

say, a small annuity provider or a small company pension scheme that is unable to reap the benefits of specialized mortality expertise.

Instead of a full buyout, it is also possible to undertake many different forms of partial buyout. For example, a pension fund might use liability-driven investing (LDI) to manage liabilities out, say, 15 years but buyout liabilities above 15 years. Alternatively, it might buy out all members over 70, buy out spouses' pensions, buy out deferred pensions, or buy out level pensions in payment.

References

- [1] Cowley, A. & Cummins, J.D. (2006). Securitization of life insurance assets and liabilities, *The Journal of Risk and Insurance* **72**, 193–226.
- [2] Firla-Cuchra, M. & Jenkinson, T. (2005). *Why are Securitization Issues Tranche?* Oxford Financial Research Centre Working Paper 2005 feb04.
- [3] Cummins, J.D., Weiss, M.A. & Zi, H. (1999). Organizational form and efficiency: an analysis of stock and mutual property-liability insurers, *Management Science* **45**, 1254–1269.
- [4] Blake, D., Cairns, A.J.G. & Dowd, K. (2006). Living with mortality: longevity bonds and other mortality-linked securities, *British Actuarial Journal* **12**, 153–197.

Further Reading

- Dowd, K., Cairns, A.J.G. & Blake, D. (2006). Mortality-dependent financial risk measures, *Insurance: Mathematics and Economics* **38**, 427–440.
- Krutov, A. (2006). Insurance-linked securities: an emerging class of financial instruments, *Financial Engineering News* **48**, 7, 16.
- Lin, Y. & Cox, S.H. (2005). Securitization of mortality risks in life annuities, *Journal of Risk and Insurance* **72**, 227–252.

DAVID BLAKE AND KEVIN DOWD

Simulation in Risk Management

Risk is a complex concept, involving the fear of things going wrong and potential associated losses. It conjures notions of real and perceived disasters, singly or in combination, resulting in large losses and possibly default on debt or bankruptcy. The nature of the potential disaster differs, of course, from one enterprise to another, from the risk of hurricanes or plant shutdowns to unexpected turns in market direction or the threat of enormous losses at the hands of rogue traders (recall Barings Bank and Long-Term Capital).

There are two components of risk, the probability of combinations of (negative) events and the magnitude of the losses resulting from these events. Both probability and magnitude could differ depending on the perception of different individuals, and in some cases, the market perceived or “risk-neutral” probabilities (*see Options and Guarantees in Life Insurance; Weather Derivatives; Mathematical Models of Credit Risk*) of events are more important than the real-world probabilities, because they drive publicly traded asset prices in the immediate term.

The Basel committee (*see Fraud in Insurance; From Basel II to Solvency II – Risk Management in the Insurance Sector; Informational Value of Corporate Issuer Credit Ratings; Compliance with Treatment Allocation*) in [1] provides a framework for regulating minimum capital requirements for banks to cover losses incurred under five different types of risk: credit risk, market risk, operational risk, liquidity risk, and legal risk, and many of these categories carry over to different types of industry as well. For banks, the most important risk is credit risk, the risk of default of a counterparty to a transaction. Of course, each of these types of risk has contributing components, and the total risk exposure of an organization is an aggregate of all the components of all five types of risk.

What are the essential features of risk? Complex structures or organizations are exposed to many different risk factors or types of risk. The probability of one or more risk events is often very small and difficult to assess for lack of historical experience. There is a relationship among risk factors that may increase the probability of them occurring

in combination and this relationship is even more difficult to assess statistically from historical data. Because of the complexity, simulation methodology is quickly becoming the method of choice for evaluating and providing safeguards against the potential losses resulting from risk exposure. Monte Carlo simulation (*see Risk Measures and Economic Capital for (Re)insurers; Reliability Optimization; Uncertainty Analysis and Dependence Modeling; Digital Governance, Hotspot Detection, and Homeland Security*) is a cost-effective method to quantify the risk of a project or investment, a generally accepted tool throughout engineering, business, operations research, statistics, and finance.

General Simulation Techniques

The first step in quantifying risk through simulation is the construction of appropriate models that strike a balance between simplicity and a close approximation of the structure of the system being analyzed. Parameters associated with statistical distributions and failure rates should, wherever possible, be calibrated to historical data. For losses due to weather or natural disaster, or to a change in market prices of publicly traded assets such as a stocks or bonds, there is considerable useful data relevant to the estimation of the model parameters underlying marginal distributions. Unfortunately, estimation of the joint distributions is more difficult. Validation of the simulation model (*see Global Warming; Hazard and Hazard Ratio*) requires checking the reasonableness of the model output by comparison with the system being modeled. The actual statistical distributions used are somewhat less important than acknowledging and modeling the uncertainties in the system in the first place. This requires the use of real-world data wherever possible to calibrate distribution parameters such as the mean and the variance.

Let us suppose that we wish to generate a random variable X , a component of the loss, with given cumulative distribution function $F(x) = P(X \leq x)$. The most general method for simulating X , increasingly useful given the general computer implementation of inverse cumulative distribution functions, is *inverse transform* (*see Copulas and Other Measures of Dependency*): generate a uniform(0, 1) random variable U and then solve the equation $F(X) = U$ for X . When X is discrete, for example, it takes

only integer values, no such solution may exist, and in this case we define X to be the smallest value of x such that $F(x) \geq U$.

With rapid increases in computer processing speed, simulation error is increasingly less of a concern. In general, standard error decreases with the square root of the number of simulations (see **Bayesian Statistics in Quantitative Risk Assessment**): if a single stochastic simulation estimates expected loss $E(L)$ with variance $\text{var}(L) = \sigma^2$ and standard deviation σ , then the average $\frac{1}{n} \sum_{i=1}^n L_i$ of n simulated (independent) values of L_i has standard error $\sigma/\sqrt{n}$. If the standard error of a single simulated estimate of risk is 100% of the value of $E(L)$, then 1 million simulations, often possible in a matter of a few seconds, will provide an estimate of the risk with relative error only around 1/10th of 1%. There is little point in more precision than this, since the error in the model or the model parameters likely contribute more than this to the total error.

Simulating VaR and TVaR

Since risk is usually identified with large losses, its measurement requires quantitative measures of the size of the right tail of the loss distribution. Suppose we have observed or are able to model the distribution of the total loss or the loss in a single unit L over a certain time horizon such as 1 day or 2 weeks. Then the *value at risk* (see **Risk Measures and Economic Capital for (Re)insurers; Solvency; Credit Scoring via Altman Z-Score; Compliance with Treatment Allocation**) is the capital required to ensure solvency with a reasonably high probability, say 95%. For a continuous distribution and $0 < p < 1$ (usually 1 or 5%), VaR_p is defined as the quantile x_p such that $P(L > x_p) = p$. In order to estimate this quantity by simulation, we can randomly generate losses $L_1, \dots, L_n$ from the model, usually by random sampling of the risk factors, and then computing the fraction of losses, $\frac{1}{n} \sum_{i=1}^n I(L_i > x)$, that exceed each threshold x where $I(A) = 1$ if the event A occurs, and 0 otherwise. We then choose a threshold such that this is approximately equal to p , i.e., solve for $\hat{x}_p$

$$\frac{1}{n} \sum_{i=1}^n I(L_i > \hat{x}_p) \simeq p \quad (1)$$

The value x_p specifies the minimum capital requirements to ensure survival with probability

$1 - p$, but it does not measure the size of those losses that exceed x_p . If the loss distribution is specified by $\bar{F}(x) = P(L > x)$ then x_p solves $\bar{F}(x_p) = p$ and the conditional distribution of losses given that they exceed x_p is $P(L > x | L > x_p) = p^{-1} \bar{F}(x)$ for $x \geq x_p$. The expected size of these (large) losses is variously called the *tail conditional expectation*, *expected shortfall*, or *tail value at risk* (see **Nonlife Loss Reserving; Asset–Liability Management for Nonlife Insurers**).

$$\text{TVaR}_p = E[L | L > x_p] = \frac{1}{p} \int_{x_p}^{\infty} \bar{F}(x) dx \quad (2)$$

On the basis of the n losses $L_1, \dots, L_n$ we can simulate TVaR_p using the average of the losses that exceed $\hat{x}_p$, i.e.,

$$\widehat{\text{TVaR}}_p = \frac{\sum_{i=1}^n L_i I(L_i > \hat{x}_p)}{p} \quad (3)$$

Of course, it is difficult to fully summarize the distribution of losses using two numbers VaR_p and TVaR_p and other moments, such as $E[L^2 | L > x_p]$, might well be considered.

For various reasons, a simulation designed to estimate VaR_p and TVaR_p may be slow or inaccurate. The probability $P(L_i > x_p)$ is typically small and may require a large number of simulations in order to obtain a reasonable number for which $L_i > x_p$. It is often possible to improve the precision of this simulation by arranging for more informative simulations in the region of interest.

Variance Reduction Techniques

These are techniques designed to decrease the variance or improve the precision of a Monte Carlo estimator by design of the simulation (see [2] or [3]). Here we discuss only a few such methods (see **Asset–Liability Management for Life Insurers; Structural Reliability**).

Suppose we are interested in estimating the expected value of some function such as $E[h(L)]$. For example, we might take $h(L) = I(L > x)$ for fixed x since estimating this quantity allows finding x_p by solving (1). Suppose we know $E[h(Z)]$ for Z under a simple closely related distribution such as the normal

distribution and it is possible to generate Z using some of the most important uniform[0, 1] random numbers U that were used in the simulation of L so that Z is highly correlated with L . The idea here is that it is often easier to estimate the (hopefully small) difference $E[h(L) - h(Z)]$ than to estimate $E[h(L)]$ in isolation. We assume that Z has a distribution sufficiently simple that $E[h(Z)]$ is known. Then we can generate random variables Z_i with the distribution of Z but closely related to L_i (for example, we could use for $Z_i = \bar{G}^{-1}(U_i)$ where $\bar{G}(x) = P(Z > x)$ and U_i is a uniform input to L_i). We can then estimate $E[h(L)]$ using

$$E[h(Z)] + \frac{1}{n} \sum_{i=1}^n [h(L_i) - h(Z_i)] \quad (4)$$

The simplest Monte Carlo estimator is

$$\frac{1}{n} \sum_{i=1}^n h(L_i) \quad (5)$$

Then if the L_i are independent, equation (5) has variance $\text{var}(h(L_1))/n$ and equation (4) has variance $\text{var}(h(L_i) - h(Z_i))/n$. The latter is an improvement if $\text{var}(h(L_i) - h(Z_i)) < \text{var}(h(L_i))$, or roughly provided that the random variables $h(L_i)$ and $h(Z_i)$ are positively correlated. Equation (4) is referred to as a *control variate* method of variance reduction. For estimation of VaR , Glasserman *et al.* [4] used the first two derivatives of the portfolio value with respect to the underlying to obtain a *delta-gamma approximation* (see **Integration of Risk Types**), and then this approximation as a control variate.

There is a similar technique called *importance sampling* (see **Bayesian Statistics in Quantitative Risk Assessment**), useful in estimating rare events or extreme quantiles as is the case in estimates of VaR and TVaR . Typically too few of our crude simulations are around the value x_p and the sample is rather noninformative of this value.

Importance sampling increases the density of sampling in the region that is most informative. For example suppose that the loss distribution has probability density function $f(x) = -\bar{F}'(x)$. Consider a random variable Z that has probability density function $g(x)$ and suppose that $g(x) > 0$ whenever $f(x) > 0$. We are interested in estimating the expected value

$E[h(L)]$, and

$$\begin{aligned} E[h(L)] &= \int h(x) f(x) dx \\ &= \int h(x) \frac{f(x)}{g(x)} g(x) dx \\ &= E \left[h(Z) \frac{f(Z)}{g(Z)} \right] \end{aligned} \quad (6)$$

This latter expected value is estimated by generating random variables $Z_1, Z_2, \dots, Z_n$ from the probability density function $g(z)$ and estimating $E[h(L)]$ by a weighted average of the values of $h(Z_i)$ of the form $\frac{1}{n} \sum_{i=1}^n h(Z_i) w_i$ where the importance weights are

$$w_i = \frac{f(Z_i)}{g(Z_i)} \quad (7)$$

The variance of equation (7) is now given by $\frac{1}{n} \text{var} \left(h(Z_1) \frac{f(Z_1)}{g(Z_1)} \right)$. We try to find an importance density $g(z)$ roughly proportional to $h(z)f(z)$ because then the variance is small. One simple device for choosing an importance distribution $g(z)$ is the *exponential tilt* of the original distribution; for example, choose

$$g(x) = \frac{e^{x\theta} f(x)}{m(\theta)} \quad (8)$$

where $m(\theta) = \int e^{x\theta} f(x) dx$ is the moment generating function of f . We are free to choose the parameter θ in an attempt to reduce the variance of equation (7), namely $n^{-1} m(\theta) \text{var} \left(h(Z_1) e^{-Z_1\theta} \right)$. In the case of estimating VaR_p , this involves adjusting θ to increase the mean $E(L)$ until the expected value of the losses is around x_p . There is an application of this technique to efficient estimation of VaR in [4] in which the importance distribution is a multivariate normal with different mean and covariance matrix, and the choice of parameters is guided by a delta-gamma approximation to the portfolio.

Simulation and Credit Risk

The term *credit risk* refers to the potential for financial loss when a counterparty to a transaction defaults (i.e., is unable to fulfill) on its obligation. An example would be when a corporation is unable to make the coupon payments on its bonds. Credit risk is perhaps the most active area of research in quantitative

finance, with new papers and books appearing virtually every week. Bielecki and Rutkowski [5] provide a comprehensive technical introduction to the subject, and the website www.defaultrisk.com provides current working papers and book reviews.

Consider a financial institution with a portfolio of N different “credits”. This could be a bank with a large number of outstanding loans or a pension fund with a large bond portfolio. Of particular importance in the modeling and assessment of credit risk is the distribution of the random variable

$$L_T = \sum_{i=1}^N I_i \quad (9)$$

where I_i is an “indicator” random variable taking on the value 1 if the i th credit defaults at or before time T , and 0 otherwise. Thus L_T represents the number of defaults up to time T in the portfolio. As defaults of different obligors are almost never independent, understanding the distribution of L_T requires an understanding of the joint distribution of $(I_1, \dots, I_N)$. We now mention a few potential approaches to modeling such a distribution.

In the *Vasicek model* (see **Credit Risk Models**) [6], it is assumed that $I_i = I(X_i > b)$, where $(X_1, \dots, X_N)$ are equicorrelated multivariate normal random variables with mean of zero. The X_i are decomposed as $X_i = \sqrt{\rho}M + \sqrt{1-\rho}Y_i$ where $M, Y_1, \dots, Y_N$ are independent standard normal random variables. M represents a *systematic factor* that is common to all obligors, while Y_i represents an *idiosyncratic factor* that is firm-specific. In this model, the asymptotic distribution (as $N \rightarrow \infty$) of $\frac{1}{N} \sum_{i=1}^N I_i$ is available and quite easy to work with. Several authors have extended this model, using the same type of decomposition, but with different distributional assumptions for M and Y_i . As the asymptotic distributions are often quite tractable in this type of model, Monte Carlo simulation is generally not required.

Another popular class of models is the so-called copula models (see **Dependent Insurance Risks; Individual Risk Models; Default Correlation; Uncertainty Analysis and Dependence Modeling** [7]). This framework begins with nonnegative random variables $\tau_1, \dots, \tau_N$ representing the default times of the individual obligors. The default indicators then take the form $I_T^i = I(\tau_i \leq T)$. The marginal distributions of the τ_i are often inferred

from market data such as bond prices and credit default swaps, and the joint distribution is modeled by choosing a particular copula, such as the Gaussian or Student- t [8]. Efficient numerical methods for evaluating quantities such as $P(L_T = k)$ are often available for copula models, so as with asymptotic factor models above Monte Carlo methods are often not necessary (see [9]).

Structural models (see **Default Risk**) attempt to link default times with the underlying “value” of a firm. They are intuitively pleasing but can be intractable or produce unrealistic predictions. A structural model assumes the firm value V_t^i follows a specific stochastic process (such as geometric Brownian motion, possibly with jumps), and defines the default time as the first passage time of the value process to some time-dependent barrier $h(t)$, which would represent some lower threshold such that the firm must default on its outstanding obligations if the value of its assets falls below this threshold. As they involve first passage times of (often high-dimensional) multivariate correlated processes, structural models are mathematically complicated, even in the simplest case of correlated geometric Brownian motion. The remainder of this section is devoted to a discussion on how to simulate multivariate default indicators $I_i = I(\tau_i \leq T)$ in the situation where the τ_i are first passage times of correlated Brownian motion. The simplest structural model for credit risk assumes that the value of a firm’s assets at any time, V_t , follows a geometric Brownian motion. The firm defaults on its debt when the asset value falls below a deterministic barrier of the form $Ke^{\lambda t}$ for the first time. This is often referred to as the *Black – Cox model*, and a comprehensive treatment is available in [5]. If $dV_t = \mu V_t dt + \sigma V_t dW_t$, then $V_t = V_0 \exp((\mu - \sigma^2/2)t + \sigma W_t)$. Assuming $V_0 > K$ it is easy to see that τ is the first passage time of W_t to the linear barrier

$$\frac{\log(K/V_0)}{\sigma} + \frac{\lambda - \mu + \sigma^2/2}{\sigma} t \quad (10)$$

For simplicity we will assume that $\lambda = \mu - \sigma^2/2$ so that τ is the first passage time of W_t to the fixed level $\log(K/V_0)/\sigma < 0$. In this restricted structural model, simulating default indicators amounts to simulating indicators of the form $I(\tau \leq T)$ where τ is the first passage time of a standard Brownian motion to a fixed negative level. Before dealing with a multifirm model, we discuss a method for simulating

the indicators $I(\tau \leq T)$. The distribution of τ is easy to simulate from and we could simulate the indicators directly. Alternatively, we could divide the interval $[0, T]$ into M equal subintervals of length $\Delta = T/M$, and simulate the values of W at the endpoints of each interval. Let these simulated values be $w_1, \dots, w_M$. Now, if $w_j < b$ for any $j = 1, \dots, M$, then clearly $I(\tau \leq T) = 1$. If we choose to set $I = 0$ if $w_j \geq b$ for each j , then we would be underestimating the crossing probability since the path may still have crossed the barrier and returned inside one of the intervals. Thus, simply setting $I = 0$ if $w_j \geq b$ for each j will not produce an accurate simulation of the indicators (we will be getting too many observations that are 0). An accurate simulation requires us to “check” if the process crossed in any of the M intervals. Conditional on the values w_j (each of which is greater than b), the probability that W crossed the barrier and returned in the j th interval is

$$p_j = \exp\left(-\frac{2(w_j - b)(w_{j-1} - b)}{\Delta}\right) \quad (11)$$

and to check if such a crossing occurred in the j th interval we simply generate a random number U and compare it to p_j . We can check each interval recursively as follows: compute p_1 and generate a random number U . If $U \leq p_1$, then W crossed the barrier. We set $I = 1$ and the simulation is complete. Otherwise we repeat the procedure for each subsequent interval until we find one where the process crossed (in which case we set $I = 1$) or we have checked each interval (in which case we conclude that the process did not cross and set $I = 0$). This will provide an exact simulation of the indicators. The Black – Cox model can be easily extended to the case of many firms whose assets are correlated by assuming that the dynamics of the i th firm’s assets are given by

$$dV_t^i = \mu_i V_t^i dt + \sigma_i V_t^i dW_t^i \quad (12)$$

where W_t^i is a standard Brownian motion and the correlation between W_t^i and W_t^j is ρ_{ij} . We will assume that the default time of the i th firm is the first hitting time of W^i to a fixed level $b_i < 0$. Recall that we are interested in the distribution of

$$L_T = \sum_{i=1}^N I(\tau_i \leq T) \quad (13)$$

and that $P(L_T = k)$ is the probability that exactly k out of N firms default. For $N = 2$ a closed form for these probabilities is available and good approximations are available for $N = 3, 4, 5$. Unfortunately for $N \geq 6$ analytic expressions are presently out of the question and Monte Carlo simulation is required. We now provide a brief outline of how the simulation would proceed. We have an N -dimensional Brownian motion (see **Insurance Pricing/Nonlife; Nonparametric Calibration of Derivatives Models**).

$\mathbf{W}_t = (W_t^1, \dots, W_t^N)$, and our assumption regarding the correlation between the components means that the covariance matrix of $\mathbf{W}_t$ is Σt , where $\Sigma_{ij} = \rho_{ij}$. The increments of this process are independent, with the covariance matrix of the increment $\mathbf{W}_t - \mathbf{W}_s$ equal to $\Sigma(t - s)$.

- Divide the interval $[0, T]$ into M subintervals of length $\Delta = T/M$.
- Generate the increments $\mathbf{Z}_j = \mathbf{W}_{j\Delta} - \mathbf{W}_{(j-1)\Delta}$ for $j = 1, \dots, M$. Note that we assume $\mathbf{W}_0 = \mathbf{0}$. This simply amounts to generating M independent multivariate normal vectors $\mathbf{Z}_1, \dots, \mathbf{Z}_M$, each with zero mean and covariance matrix $\Delta\Sigma$. Setting $\mathbf{W}_{j\Delta} = \sum_{k=1}^j \mathbf{Z}_k$, we now have an exact simulation of a “skeleton path” for our multivariate correlated Brownian motion.
- Use the algorithm described earlier to generate the value of I^i for each i . When doing so we are only using the skeleton path for the i th component, and ignoring the simulated values of $I_1, \dots, I_{i-1}$.

This algorithm would be exact if, conditional on $\mathbf{W}_\Delta, \mathbf{W}_{2\Delta}, \dots, \mathbf{W}_T$, the events “ W^i crossed in the j th interval” and “ W^k crossed in the j th interval” were independent. Unfortunately this will not be true; however, they should be *nearly* independent if Δ is chosen to be small. The smaller the Δ , the more accurate the simulation will be, and a trade-off between accuracy and computational time is required.

Simulation and Market Risk

A firm may have assets such as securities or bonds that are accurately priced in a liquid market, and market data provides substantial experience concerning the range of potential prices and losses resulting from this portfolio. The risk associated with this portfolio is often measured with quantities such as *VaR* and *TVaR*

either from a model governing the distribution of the asset prices or from historical data. When the risk measure obtains from a model, it commonly uses a normal distribution, based loosely on the assumption that for a large number of independent losses, the total loss is approximately normally distributed according to the central limit theorem. Unfortunately, defaults or large losses are often driven largely by a small number of very large nonnormal components, and in these circumstances the central limit theorem does not apply. It is probably better, therefore, to conduct a simulation from a distribution that acknowledges that the tails of the distribution may be heavier than those of the normal. This could be achieved either by using a distribution in statistics with heavier tail than the normal such as the hyperbolic model (see [10]) or by using the statistical *bootstrap*, described below.

Suppose, for example, we wish to estimate $VaR_{0.05}$ over a time horizon of 2 weeks. Also suppose we have a reasonably large database of daily returns $R_1, R_2, \dots, R_N$ obtained, for example, over the past several years. The bootstrap assumes that future returns are sufficiently like past ones so that we can draw a representative sample of 10 returns, say $R_1^*, \dots, R_{10}^*$, from the set of values $\{R_1, R_2, \dots, R_N\}$ one at a time with replacement to simulate the daily returns over the next 2 weeks, and then assume that the return for the 2-week period is the sum of these 10 draws $R_1^* + \dots + R_{10}^*$. This means that the loss over the 2-week period is $L^* = -(R_1^* + \dots + R_{10}^*)$. This process can be repeated many times to get an idea of the distribution of biweekly losses. Provided the range of returns in the original database and the value of N are sufficiently large, this distribution will typically have larger tails than the normal. The estimate of VaR or $TVaR$ is obtained from a random sample of these values of L^* as in equations (3) and (1) but with the bootstrap (see **Combining Information; Nonlife Loss Reserving; Credit Migration Matrices**).

L^* replacing the simulated L_i . An advantage of the use of the bootstrap is that it does not assume any distribution for losses, but it only assumes that the future distribution of losses is similar to that in the past. However, since daily returns are randomly sampled in the bootstrap, the validity of this method depends heavily on the assumption that daily returns are independent. This independence is, of course, invalid if some change in the economy can have a persistent effect on returns or losses over a period

of a number of days. A partial solution simply samples adjacent 10-day returns, since this incorporates possible dependence among the observations, and only requires the assumption of stationarity, but we typically need more than a few years of daily data to implement the same. An alternative is to adopt assumptions governing how the observations scale to longer time periods (see [11, Section 2.3.4]).

Operational Risk

Operational risk (see **Nonlife Insurance Markets; Availability and Maintainability; Compliance with Treatment Allocation**) includes risks due to failure of internal operations, supervisory or management relations, fraud, etc. Many of the trading risks associated with hedge funds fit within this category, because given adequate supervision and constraints, the losses could have been avoided. For example, suppose that the compensation structure in an institution rewards risk-taking behavior by increasing the portfolio and compensation of traders based on their past returns. Suppose that there is an internal attempt to risk-adjust performance measures by constraining the VaR_p of traders. Since risk profile is essentially a whole distribution and VaR_p is a single number representing it, a savvy trader will have the ability to increase risk (e.g., $TVaR$ or the higher moments of the conditional tail behavior) while keeping VaR_p constant. This additional risk will often result in additional returns, leading to increased compensation, larger portfolios, and additional incentive to increase risk subject to whatever constraints have been imposed. Similarly, supervisors have an incentive to ease up controls on their most successful traders. Ultimately, when unrestrained, market forces are directed toward increasing risk and rewards. Increasingly, supervision will require sophisticated risk measures in order to avoid undue risk. Statistical models for operational risk are discussed at length in [12] and [11, Chapter 10] and there is a comprehensive discussion of extremal events in [13]. Although many of the statistical models for individual risk factors are reasonably tractable analytically, they will often require simulation to deal with the total risk of the operation.

Liquidity and Legal Risk

Liquidity risk (see **Enterprise Risk Management (ERM); Solvency**) is the risk of losses occurring

because an asset cannot be bought or sold in a timely manner. For example, the 1993 Metallgesellschaft crisis was caused by large margin calls and an inability to unwind positions taken on futures. Risk can be modeled to some extent by modeling the time between transactions, but this must include a possible adjustment for changing market conditions. See, for example [14]. Legal risk (*see Role of Risk Communication in a Comprehensive Risk Management Approach; Environmental Risk Regulation; Operational Risk Modeling*) is risk due to uncertainty in legal actions or applicability of contracts, laws, or regulations. Information such as the fraction of contracts that are deemed illegitimate by courts, the fraction of counterparties that do not have legal capacity to enter into a transaction, the fraction of firms susceptible to copyright challenges are relevant to the construction of a suitable model. One might assume a fraction of employees are susceptible to fraudulent behavior and a fraction of supervision less vigilant than it should be, and simulate the potential losses under various scenarios.

References

- [1] International Convergence of Capital Measurement and Capital Standards: A Revised Framework (2005). *Committee on Banking Supervision*, Bank for International Settlements, at <http://www.bis.org/publ/>.
- [2] McLeish, D.L. (2005). *Monte Carlo Simulation and Finance*, John Wiley & Sons, New York.
- [3] Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering, Applications of Mathematics*, 53, Springer, New York.
- [4] Glasserman, P., Heidelberger, P. & Shahabuddin, P. (2001). Efficient Monte-Carlo methods for value-at-risk, *Mastering Risk Volume 2: Applications*, Prentice Hall, London.
- [5] Bielecki, T.R. & Rutkowski, M. (2002). *Credit Risk: Modeling, Valuation and Hedging*, Springer-Verlag, Berlin.
- [6] Vasicek, O. (2002). Loan portfolio value, *Risk* 15, 160–162.
- [7] Li, D. (2000). On default correlation: a copula function approach, *The Journal of Fixed Income* 9, 43–54.
- [8] Schloegl, L. & O’Kane, D. (2005). A note on the large homogeneous portfolio approximation with the student-t copula, *Finance and Stochastics* 9, 577–584.
- [9] Hull, J. & White, A. (2004). Valuation of a CDO and nth to default CDS without Monte Carlo simulation, *Journal of Derivatives* 12, 8–23.
- [10] Eberlein, E. (2001). Recent advances in more realistic market risk management: the hyperbolic model, *Mastering Risk Volume 2: Applications*, Prentice Hall, London.
- [11] McNeil, A.J., Frey, R. & Paul Embrechts, E. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools*, Princeton University Press, Princeton.
- [12] Panjer, H.H. (2006). *Operational Risks, Modeling Analytics*, John Wiley & Sons, New York.
- [13] Embrechts, P., Kluppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.
- [14] Çetin, U., Jarrow, R.A. & Protter, P. (2004). Liquidity risk and arbitrage pricing theory, *Finance and Stochastics* 8, 311–341.

DON L. MCLEISH AND ADAM METZLER

Social Networks

Social network theory studies the formation and strength of ties among people. Early work focused on friendship ties or acquaintanceship ties, and this led to Milgram's famous rule that the average number of acquaintances linking two random people is about six (*cf.* [1]). This estimate was based on an experiment in which people tried to send mail to a stranger by relaying it through chains of acquaintances whom they believed might be "closer" to the target recipient and willing to forward it. This experiment is related to the recent *smallworld* theory, in which Milgram's relatively low number is explained in terms of most people having only local contacts, but a handful of people having very distant connections (*cf.* [2], for a review of this area).

More recently, people have become interested in networks of gene or protein interactions, or telecommunications networks, or Internet links (e.g. [3], use network methods to estimate the size of the world wide web). Scientific and commercial applications now dominate the social network field, and the traditional mathematical models (e.g., the p_1 and p^* models) do not necessarily provide very good explanation for these new situations. Additionally, much attention is now given to dynamic networks, in which new links form or disappear over time, and people join or leave the network.

The main driver for the recent renaissance in social network theory is the availability of data. Before the mid-1990s, only a few classic data sets were publicly available for analysis. However, the rise of computerized methods to automatically capture connection information, in many different contexts, as well as sophisticated new technology in bioinformatics (*see* **Microarray Analysis**) for studying relationships, have provided fresh datasets that are prompting new ideas. Additionally, the emergence of good graphic software for network visualization is an important enabler of research. GraphViz is one of the most popular freeware tools, but there are many others.

In the context of risk analysis, most of the interest in social network models pertains to either modeling the spread of contagious diseases or understanding the formation, structure, and evolution of terrorist cells. There is also some work in forensic social networks and in information networks that may find

unobvious links ("connecting the dots") in order to discover terrorist plots.

Social Network Models

Early ideas about the behavior and role of social networks were pioneered by Moreno [4], Festinger [5], and Katz [6–8]. However, modern statistical modeling was introduced by Holland and Leinhardt [9], with the p_1 model (*see* **Risk–Benefit Analysis for Environmental Applications**). The p_1 model assumes that the probability of a link *from* actor i and *to* actor j is

$$\text{logit}(p_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad (1)$$

where $\text{logit}(x) = \ln x/(1 - x)$, μ is the average rate with which actors in this network form links, α_i is the additional rate effect due to the relative extroversion or introversion of actor i , β_j is the effect due to the relative receptivity of actor j , and γ_{ij} is the effect due to joint properties of the actors (e.g., they work for the same company, or have opposing political views).

The initial applications used data from sociological surveys of small communities, such as the Sampson Monastery Data [10]. Such real-world work suggested various extensions of the p_1 framework. Feinberg *et al.* [11] placed it squarely in the framework of log-linear models (*see* **Dose–Response Analysis; Potency Estimation; Geographic Disease Risk**), and developed a version of it that applies when there are multiple kinds of edges describing different sorts of relationships (e.g., friend, family, and coworker). They emphasized that the p_1 model is an exponential family and thus many inferential tools apply automatically. They highlight the need for developing goodness-of-fit tests, which is still an open problem in general, although in some cases the null distribution of specific model diagnostics can be calculated.

Sociological theory, coupled with experimental data, indicate that human networks are shaped by transitivity, homophily, and clustering. Much of the modern work is an effort to build models that describe these effects, none of which are captured in the p_1 model. A recent paper that surveys these ideas in the context of modern latent space models is Handcock and Raftery [12].

It is worth emphasizing that transitivity, homophily, and clustering are absent from the recent

small world research, which is driven entirely by the distribution of the numbers of acquaintances (the in-degrees and out-degrees of the vertices in the graph). This implies that some important properties of networks can be understood from very basic principles. However, sociologists would assert that human social networks require more complex models that incorporate subtle, but completely reasonable, effects.

Transitivity

Transitivity refers to the fact that if Abelard is a friend of Balthazar and Balthazar is a friend of Clytemnestra, then Abelard and Clytemnestra are likely to become friends too (this is sometimes called *triad completion*). Frank and Strauss [13] developed a characterization of the most general exponential family model for network data with binary edges. Using the Hammersley–Clifford theorem, they show that all models of this kind can be written in terms of a joint distribution defined by sets of cliques. This allows models in which the probability of edge relations depend upon whether two edges share a common actor, permitting models that show transitivity (while retaining many desirable inferential properties). The popular p^* model developed by Wasserman and Pattison [14] is within this family and transitivity has a tendency to cause clique structures, which is a special case of the observed clustering that occurs in human networks.

Transitivity appears to play a remarkable role in enforcing behavioral norms. Krackhardt [15] did experiments in which pairs of people were invited to play “prisoner’s dilemma”, (see **Group Decision; Game Theoretic Methods**) a game in which there is a large individual reward for successful betrayal and a smaller joint reward for cooperation. He found that people who were strangers and people who were friends but had no mutual friends showed similar defection rates of about 70%, but that people who were friends and had mutual friends in common were much less likely to defect (the rate was about 30%).

Homophily

Homophily describes the tendency of people to form links with others like themselves. For most social networks, edges tend to join people of similar age, income, race, and education (but for sexual networks,

they tend to join people of opposite genders, an example of heterophily). Often these demographic traits are observable covariates, and could be included in the model. Data for fitting such models used to come from studies of small communities, but new data sources are now available from e-mail records and such.

In developing models for homophily, researchers want to build models that allow increased chance of connection between similar people. For example, $\text{logit}(p_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}$ is a trivial generalization of the p_1 model in which the vector $\mathbf{x}_{ij}$ contains covariate information on, say, the age and income of both actors, as well as indicators for whether or not the actors are of the same gender or work for the same company (as well as information on the extroversion of i and the receptivity of j). This use of covariate information allows models that boost the probability of ties between similar individuals. If the covariate information is categorical (e.g., gender or race), then one has the stochastic blockmodel proposed by Feinberg and Wasserman [16]. In general, it is hard to know when a given model is adequate; domain scientists use their best judgment about what kinds of similarities are most important.

Of course, in many applications, important covariates might not be measured. For example, although a study might record race, gender, and income, it is entirely possible that categorical variables such as political party or religious faith play an important role in the friendship structure, and these covariates might not be measured. This situation is addressed by a stochastic block model with latent (unobserved) categories, as pioneered by Anderson and Wasserman [17] and Snijders and Nowicki [18]. The model assumes that each actor has unknown category labels, and estimates these in the process of fitting the model (usually *via* Bayesian inference (see **Repairable Systems Reliability; Mathematics of Risk and Reliability: A Select History; Bayesian Statistics in Quantitative Risk Assessment; Subjective Probability**)).

Recently, latent categories have been generalized to positions in a latent space. Hoff *et al.* [19] proposed a model in which the probability of an edge from i to j is modeled as

$$\text{logit}(p_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \|\mathbf{z}_i - \mathbf{z}_j\| \quad (2)$$

where $\mathbf{x}$ is a vector of measured covariates on the actors i and j and the $\mathbf{z}_i$ and $\mathbf{z}_j$ are the positions

of actors i and j in a latent space endowed with a distance metric having norm $\|\cdot\|$. Often it is possible to estimate the relative locations of the actors in the latent space, and then interpret the dimensions of the space in meaningful ways, e.g., as degree of political liberalism or religiosity.

One issue in latent space models is the determination of the appropriate dimension; most people use two, simply because this is convenient to visualize. The norm is also up to the analyst, but the common choice is Euclidean distance. These models are typically fit from a Bayesian perspective using Markov chain Monte Carlo (*see* **Global Warming; Nonlife Loss Reserving; Reliability Demonstration; Expert Elicitation for Risk Assessment**) (*cf.* [20]).

Other Kinds of Clustering and Structure

Beyond transitivity and homophily, many networks exhibit additional cluster structure. This might be caused by the efforts of actors to balance their networks, so as to include people with specific abilities (this is reasonable for workplace friendships or terrorist cells in which piloting or bomb-making skills are needed). Also, there can be a tendency for people with many ties to accumulate other ties more quickly; the rate of increase might be proportional to the current value (this could describe a celebrity or politician). Sometimes there is special “betweenness” structure, in which one actor connects two relatively disjoint clusters (e.g., a bilingual person can connect two monolingual groups), and there is a theory that the number of social ties a person has cannot exceed about 150 (Dunbar’s number), this representing the size of a typical human tribe and the limit of cognitive ability to manage the pairwise relations among one’s connections (*cf.* [21]). This short list of examples does not exhaust the kinds of subtle effects that shape social networks; see Freeman [22] for a more detailed discussion.

However, there are two kinds of structures that merit special attention. The first concerns feedback loops; these are prominent in biological networks, but have not yet had much influence on social network models. Nonetheless, there are clearly mechanisms of this kind; often one actor will push another away if the relationship becomes too intense. Mathematical descriptions of such models will probably also need to incorporate long-term memory, which is not usually part of the biological thinking.

The second kind of structure reflects external forces. For example, in the workplace a great deal of the interaction, and much of the social network, is driven by the org-chart. People in the same division and at the same level are more likely to form ties, and more likely to compete directly. Blockmodeling allows one to partially capture some of this effect, but it would be better to have techniques that flexibly condition on specific social structures (such as org-charts or family relationships) and then use the previous techniques to study the relationships that are not already accounted for.

Counterterrorism

Social networks are becoming prominent in national security applications. From the perspective of risk analysis, people are interested in

- how social networks affect the progress of epidemics (natural or deliberate);
- how terrorist cells form, interlink, and possibly can be disconnected by strategic removal of key figures;
- how investigators can link suspects through forensic networks;
- how sets of “documents” relate to each other, and to specific people, events, or locations (this is the classic “connecting-the-dots” approach).

In the last application, text mining can be used to determine links.

However, it should be emphasized that social network methods may have been oversold. In none of these topics has there yet arisen a clear and compelling social network contribution that is of practical importance (at least in the open literature).

A second point is that many of these applications do not explicitly rely upon the mathematical models described in the previous section. Instead, they employ massive agent-based simulations. The actors are artificial, and they are assigned rules for their interactions (which may be probabilistic). The simulation then mimics the formation and dissolution of ties among them. Epstein and Axtell [23] give a provocative review of the potential scope for agent-based modeling.

Epidemic Spread

In order to assess the risk of a disease, one needs a method to estimate its consequences. For contagious disease, this implies developing models for its spread. Social network models are a tool for this purpose.

A standard example is the study of sexually transmitted diseases, such as HIV. If one understands the network of sex workers and heroin users in a community, then it becomes possible to forecast the magnitude of the public health problem and the rate at which the disease propagates through the network. In particular, it can help to gauge the impact of, say, a needle-exchange program, or find critical paths for spread, or more quickly detect the presence of a “super-spreader” who requires immediate medical treatment or counseling. A classic study of this kind examined HIV transmission among the gay and addict communities in Colorado Springs (*cf.* [24]).

Many epidemic studies use mathematical models for disease spread, such as the Reed–Frost or Kermack–McKendrick systems of coupled differential equations. However, these assume a freely mixing population, an assumption that quickly breaks down in the case of a pandemic or bioterrorist attack. One needs a procedure that accommodates self-isolation and other behavioral changes that occur. In this regard, the agent-based models are attractive; one can program the actors with rules that reflect realistic responses. This enables planners to explore the extent to which disease spread can be broken by early detection, or by closing schools and nonessential businesses, or other public health interventions. Carley *et al.* [25] is an example of agent-based modeling in the context of bioterrorism (*see* **Syndromic Surveillance; Public Health Surveillance; Digital Governance, Hotspot Detection, and Homeland Security**).

Terrorist Cells

After 9/11, social network people in federal and civilian research centers began to study the relationships among the hijackers using *post hoc* intelligence data, testimony, and background investigations. It has been suggested that this kind of work helped to identify a set of imams at mosques around the world who were recruiting disaffected young Muslims into Al-Qaeda cells. However, the evidence for his level of impact is spotty, and probably the social network research did

not add much to the connections that were discovered by intelligence analysts.

Nonetheless, the effort pointed out the kinds of things that social networks might be able to achieve someday. For example, one might be able to deduce the presence of an unknown terrorist from coordinated patterns of activity in apparently disconnected components of a network. Or one might be able to identify an actor who, if removed, would fragment the organization (*cf.* [26]). Perhaps the skill sets of the actors could indicate the kinds of terrorist activities that were planned (e.g., a genetics expert or nuclear scientist would look ominous) and from a legal standpoint, the absence of Zacarias Moussaoui from the social network of the 9/11 hijackers is intriguing.

These topics relate to inference on partially observed networks. This arises when one has

- an incomplete list of actors, so that the formation of links can be influenced by unseen relationships (e.g., a triad completion can occur without the researcher observing one of the members of the triad);
- an incomplete list of links, so that some actors might have a concealed relationship;
- one or more time periods when measurements did not occur.

Each of these three cases is a significant issue in counterterrorism. They also invoke the problem of creating models for dynamic networks, in which edges or actors change over time. So far, it appears that useful models for network dynamics have to be hand built for specific applications; general models are too vague.

An unclassified example of this kind of analysis concerns the Enron e-mail data. It was seized from the company under subpoena by the Department of Justice, and later obtained under the Freedom of Information Act for scientific research. Figure 1 shows a social network graph indicating who sent e-mail to whom. (It is too dense to read, but it illustrates the kind of visualization that can be used.)

Figure 1 is taken from Priebe *et al.* [27]. They apply the scan statistics to discover suspicious cliques, and the method extends to the time-stamped record of e-mail traffic, potentially showing how information propagates through the network within the company. Also, one could use text mining on the body of the e-mails to refine the network so that only

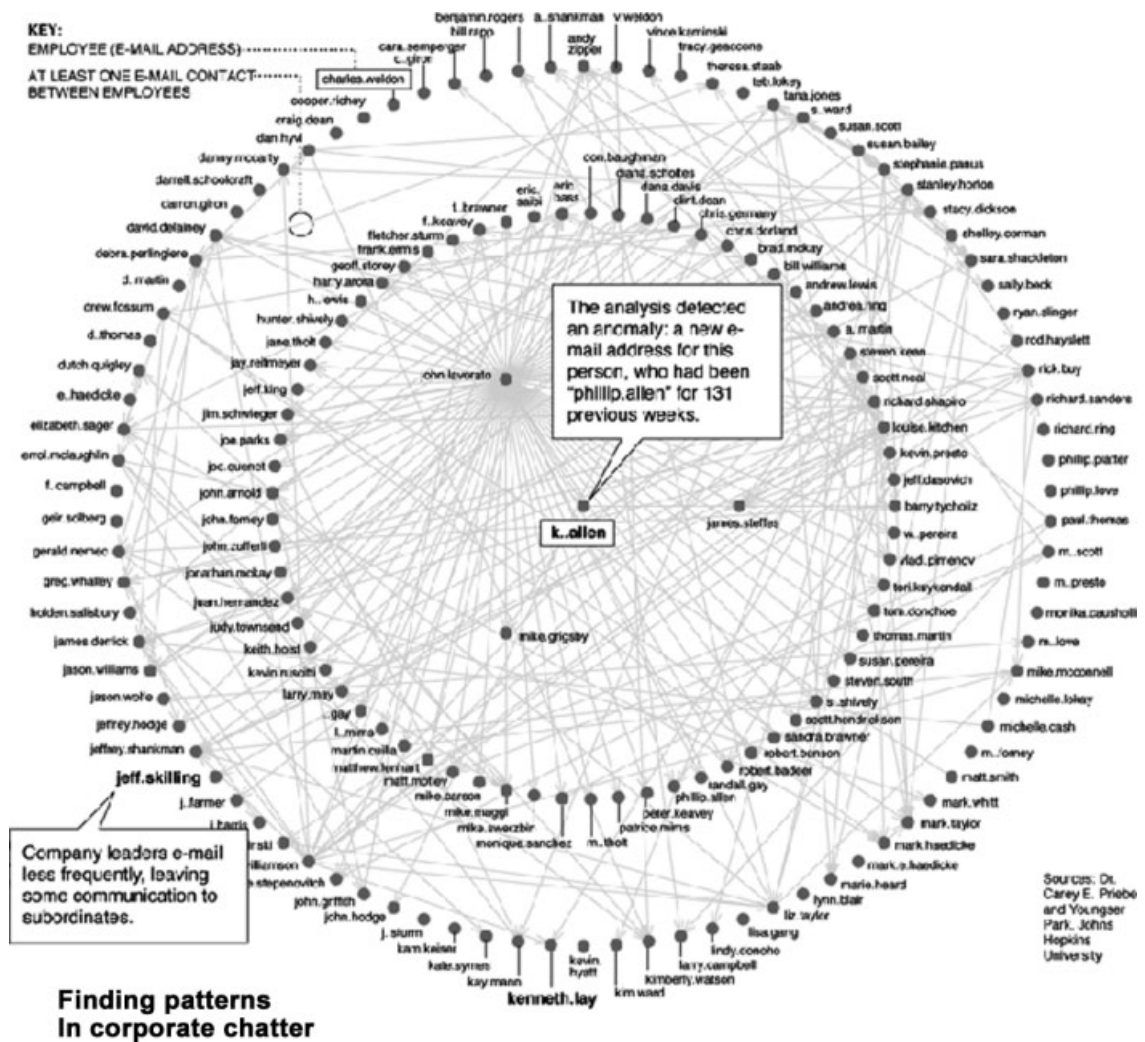


Figure 1 A visualization of the e-mail exchange network at Enron [Reproduced from [27]. © Springer, 2005.]

certain types of communication are displayed (e.g., exclude general all-hands memos, but include e-mails with the word "fraud").

However, the analyses of the Enron data have not discovered any fresh insights. There is a group of people who communicate to coordinate picking up their children, and an employee who changed his e-mail name, but there was nothing of legal consequence.

Forensic Networks and Connecting the Dots

A goal in network analysis is to use large databases to identify connections that can help in law enforcement

and counterterrorism. For example, one can imagine that police would someday be able to find the closest person in a murder victim's social network who owns a handgun, or who has a previous arrest for violent crime.

These applications would require software that enables investigators to troll through large, unstructured databases to build the social network, and then explore the links in an expanding neighborhood. Such large databases are already emerging; ChoicePoint and LexisNexis/Seisint are data marts that are currently used mostly for checking credit history, but they have subsumed many public datasets and

have the potential to support estimation of the social network surrounding an individual.

Similarly, the FBI now maintains DNA samples on a great many criminals. It is likely that with appropriate technology, a blood sample at a crime scene might be identifiable as coming from someone who is, say, a probable cousin of Tony Soprano and a matrilineal relative of Dr Melfi, even if the unknown perpetrator is not in the FBI database. A simple form of this technique was used to arrest Dennis Rader, the BTK strangler; his daughter's DNA showed a close, but not exact, match to archived evidence from the crime scenes.

These examples are special cases of the larger problem of connecting dots in seemingly disparate information sources. Abstractly, analysts would like the capability to text mine documents, phone call transcripts, commercial transaction records, and police reports in order to find threads of meaning that link clusters of records in sinister ways. The visionaries who propose this hope that social network methods may someday be supple enough to support automatic or semiautomatic research of this kind.

References

- [1] Travers, J. & Milgram, S. (1969). An experimental study of the small world problem, *Sociometry* **32**, 425–443.
- [2] Newman, M. (2003). The structure and function of complex networks, *SIAM Review* **45**, 167–256.
- [3] Dobra, A. & Feinberg, S.E. (2003). How large is the World Wide Web? in *Web Dynamics*, M. Levene & A. Poulouvassilis, eds, Springer-Verlag, New York, pp. 23–45.
- [4] Moreno, J.L. (1934). *Who Shall Survive?* Nervous and Mental Disease Publishing, Washington, DC.
- [5] Festinger, L. (1949). The analysis of sociograms using matrix algebra, *Human Relations* **2**, 153–158.
- [6] Katz, L. (1947). On the matrix analysis of sociometric data, *Sociometry* **10**, 233–241.
- [7] Katz, L. (1953). A new status index derived from sociometric analysis, *Psychometrika* **18**, 39–43.
- [8] Katz, L. (1955). Measurement of the tendency towards reciprocation of choice, *Sociometry and the Science of Man* **18**, 659–665.
- [9] Holland, P. & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs, *Journal of the American Statistical Association* **76**, 33–65.
- [10] Sampson, S.F. (1968). *A novitiate in a period of change: an experimental and case study of social relationships*, Ph.D. thesis, Cornell University.
- [11] Feinberg, S.E., Meyer, M.M. & Wasserman, S.S. (1985). Statistical analysis of multiple sociometric relations, *Journal of the American Statistical Association* **80**, 51–67.
- [12] Handcock, M. & Raftery, A. (2007). Model-based clustering for social networks, *Journal of the Royal Statistical Society, Series A* **170**, 1–22.
- [13] Frank, O. & Strauss, D. (1986). Markov graphs, *Journal of the American Statistical Association* **81**, 832–842.
- [14] Wasserman, S. & Pattison, P. (1996). Logit models and logistics regressions for social networks: I. An introduction to Markov random graphs and p^* , *Psychometrika* **60**, 401–425.
- [15] Krackhardt, D. (2006). Invited presentation at the Heider vs Simmel: comparing generative models of network formation, *Social Networks Satellite Meeting at the International Conference on Machine Learning*, June 25–29, Pittsburgh.
- [16] Feinberg, S.E. & Wasserman, S. (1981). Categorical data analysis of single sociometric relations, in *Sociological Methodology*, S. Leinhardt, ed, Jossey-Bass, San Francisco, pp. 156–192.
- [17] Anderson, C. & Wasserman, S. (1987). Stochastic a posteriori blockmodels: construction and assessment, *Social Networks* **9**, 1–36.
- [18] Snijders, T. & Nowicki, K. (1997). Estimation and prediction for stochastic blockstructures for graphs with latent block structure, *Journal of Classification* **14**, 75–100.
- [19] Hoff, P., Raftery, A. & Handcock, M. (2002). Latent space approaches to social network analysis, *Journal of the American Statistical Association* **97**, 1090–1098.
- [20] Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing*, Springer, New York.
- [21] Dunbar, R. (1993). Coevolution of neocortical size, group size and language in humans, *Behavioral and Brain Sciences* **16**, 681–735.
- [22] Freeman, L. (1979). Centrality in social networks I, *Social Networks* **1**, 215–239.
- [23] Epstein, J. & Axtell, R. (1996). *Growing Artificial Societies*, MIT Press, Cambridge.
- [24] Potterat, J., Phillips-Plummer, L., Muth, S., Rothenberg, R., Woodhouse, D., Maldonado-Long, T., Zimmerman, H. & Muth, J. (2002). Risk network structure in the early epidemic phase of HIV transmission in Colorado Springs, *Sexually Transmitted Infections* **78**, 159–163.
- [25] Carley, K., Altman, N., Casman, E., Fridsma, D., Yahja, A., Chen, L.-C., Kaminsky, B. & Nave, D. (2006). BioWar: scalable agent-based model of bioattacks, *IEEE Transactions on Systems, Man, and Cybernetics* **36**, 252–265.
- [26] Carley, K., Lee, J.-S. & Krackhardt, D. (2001). Destabilizing networks, *Connections* **24**(3), 31–34.
- [27] Priebe, C., Conroy, J., Marchette, D. & Park, Y. (2005). Scan statistics on Enron graphs, *Computational and Mathematical Organization Theory* **11**, 229–247.

DAVID L. BANKS AND NICHOLAS
HENGARTNER

Societal Decision Making

Decision and risk analyses are commonly used in the public sector to support major decisions relating to many issues including societal risks. If a potential development has significant risks to health, society in general, or the environment, not only is there a clear imperative to examine and weigh the issues in careful detail, but also an imperative to do so openly, as in the case of a broad range of western democracies. There is a need for auditable decision processes, both to show stakeholders that their concerns have been explicitly addressed even if not resolved to their entire satisfaction, and to demonstrate that good governance practices have been observed. For this reason, quantitative decision analyses have a long history in the public sector [1–3].

As we remarked in the article on decision analysis (*see* **Decision Modeling**), there is no commonly agreed methodology for the quantitative support of decision making. Here, we contrast the following two in relation to societal decision making: (Bayesian) decision analysis based upon the subjective expected utility model (*see* **Subjective Expected Utility**) and cost–benefit analysis (*see* **Cost-Effectiveness Analysis**), noting that other methodologies have also been used [4, 5]. We shall also note how they may be deployed in modern democracies, the role of regulators, and the growing role of public participation.

Cost–Benefit Analysis *versus* Decision Analysis

For many years, the most common type of analysis used to support societal decision making was cost–benefit analysis [1, 6, 7]. In fact, there is no single cost–benefit methodology but a family of methods sharing a common approach – namely, the value of a decision equals net expected benefits minus net expected costs, both expressed monetarily. The aims of cost–benefit approaches are laudable. They seek to present an analysis of potential courses of action and their monetary worth to a (large) group of people, often society in general; and by using money as a common measure, they seek to do so in an agreed objective fashion. However, the difficulty is that monetary worth is essentially a subjective judgment. Two individuals need not agree on the monetary worth of

an item or impact; and, indeed, they seldom do. Thus, in seeking to present a single monetary valuation of any course of action, cost–benefit analysis is ensuring that at its heart is an ill-defined concept. It begs controversy in assuming agreement and unanimity when it is virtually certain that there is none.

In contrast, decision analysis explicitly focuses on the individual and, in doing so, allows different individual perspectives to be compared and explored, supporting and articulating debate. Often, differences between individuals are reflected in the weights that they place upon different factors or in the likelihood of some event. By a careful use of sensitivity techniques [8], an analysis can identify which differences are key and thus focus debate on the issues that matter. It can often be shown that despite differences in judgments, many elements of society agree on the course of action, thus diffusing sterile debate. For this reason, decision analytic approaches have been gradually superseding cost–benefit ones, although some see this more as a convergence of paradigms than a shift from one to another [6, 9].

Democracy

In modern democracies, the context for such decision analyses is complex. An “Athenian” view of democracy might suggest that all members of society have an input into all societal decision making, but such direct or deliberative democracies are clearly infeasible in the modern world and did not really exist in classical times [10]. Also, as Arrow has shown, direct democracy is an ill-defined, unachievable concept [11, 12] (*see* **Group Decision**). Thus, modern democracies are constructed around representation, with politicians being elected to run the state, perhaps informed by occasional referenda. In many cases, they delegate societal decision making to regulators and planning agencies, e.g., the US food and drug administration (US FDA) or British health and safety executive (HSE). Modern societies are a complex weave of agencies and bodies with differing remits and powers of decision making overseen by national and local political assemblies and, perhaps, by an independent judiciary. And this, of course, is an ideal(!), in Politics politics are rife and decision making can sidestep some of the formal procedures, driven by other agendas and horse trading. Against this background, there is an overwhelming need for

analytic methodologies, which document explicitly the reasoning behind any decision, indicating the evidence and how it was weighed in reaching the decision.

Stakeholder Involvement and Public Participation

Societies are not static – they evolve. Modern Western democracies are suffering from a “democratic deficit” and growing public disillusionment with current, purely representational processes. Pressure groups are demanding a louder voice and a greater input for their particular stakeholder perspective into the shaping of society. This is leading to a move toward more public participation and a greater involvement of stakeholders [13–16] (*see Public Participation*). The use of mechanisms, such as stakeholder workshops, citizens’ juries, and on-line deliberation is emphasizing the need for analytical techniques to be transparent and available to a much wider audience than it was previously. If analysis is to inform debate – and surely that must be its purpose – then it must be easily understood by all the participants. Sensitivity analysis can help here, allowing different parties to the debate to see how their perspectives and judgment affect the conclusions and where they stand relative to other participants [8]. A careful use of sensitivity techniques can identify which differences are key and thus focus debate on the issues that matter. It can often be shown that despite differences in judgments, many elements of society agree on the course of action, thus diffusing sterile debate.

Discussion

A successful decision or risk analysis to support a societal decision must recognize the imperatives arising from the political context. It must recognize the different institutions involved and their statutory responsibilities, the different players involved (*see Players in a Decision*), and the need to communicate the analysis to each of these. It must seek to articulate and inform debate, explicitly demonstrating how different evidence and perspectives affect the conclusions. All this means that conducting risk and decision analyses requires many skills beyond the purely technical ones – process, communication, and

interaction are as important as the ability to model and calculate.

References

- [1] Layard, R. & Glaister, S. (1994). *Cost Benefit Analysis*, Cambridge University Press, Cambridge.
- [2] Carley, M. (1980). *Rational Techniques in Policy Analysis*, Heinemann Educational Books, London.
- [3] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Trade-offs*, John Wiley & Sons, New York.
- [4] Roy, B. (1996). *Multi-Criteria Modelling for Decision Aiding*, Kluwer Academic Publishers, Dordrecht.
- [5] French, S. (2003). Special issue: the challenges in extending the MCDA paradigm to e-democracy, *Journal of Multi-Criteria Decision Analysis* **12**, 61–233.
- [6] Bedford, T., French, S. & Atherton, E. (2005). Supporting ALARP decision-making by cost benefit analysis and multi-attribute utility theory, *Journal of Risk Research* **8**(3), 207–223.
- [7] Pearce, D.W. & Nash, C.A. (1981). *The Social Appraisal of Projects: a Text in Cost-Benefit Analysis*, Macmillan, London.
- [8] French, S. (2003). Modelling, making inferences and making decisions: the roles of sensitivity analysis, *TOP* **11**(2), 229–252.
- [9] Fischhoff, B. (1977). Cost benefit analysis and the art of motorcycle maintenance, *Policy Sciences* **8**, 177–202.
- [10] Crick, B. (2002). *Democracy: a Very Short Introduction*, Oxford University Press, Oxford.
- [11] Arrow, K.J. (1963). *Social Choice and Individual Values*, 2nd Edition, John Wiley & Sons, New York.
- [12] French, S. (2006). Web-enabled strategic GDSS, e-democracy and Arrow’s theorem: a Bayesian perspective, *Decision Support Systems* (in press) **43**, 1476–1484.
- [13] Beierle, T. & Cayford, J. (2002). *Democracy in Practice: Public Participation in Environmental Decisions*, Resources for the Future Press.
- [14] Renn, O. (1999). A model for an analytic-deliberative process in risk management, *Environmental Science and Technology* **33**(18), 3049–3055.
- [15] Renn, O., Webler, T., Rakel, H., Dienel, P. & Johnson, B. (1993). Public participation in decision making: a three step-procedure, *Policy Sciences* **26**(3), 189–214.
- [16] Gregory, R.S., Fischhoff, B. & McDaniels, T. (2005). Acceptable input: using decision analysis to guide public policy deliberations, *Decision Analysis* **2**(1), 4–16.

Related Articles

Behavioral Decision Studies
Expert Judgment
Supra Decision Maker

Software Testing and Reliability

Today computers and software are all pervasive and increasingly an important part of most systems; the latter includes embedded systems (as may be present in medical devices, airplanes, or cars), systems involving people working in harmony with computers (for instance, in air-traffic control), and so on. From the perspective of quantifying risk, it is often necessary to have some measure of the “goodness” of any associated software. This is typically captured in the concept of *software reliability*. See [1], for instance.

Of course, reliability is usually expressed as the probability of error in a particular situation and over a prescribed period of time. In some sense it is concerned with measuring the trust that can be placed in a given system. This is the basis for the definition of software reliability.

Software reliability is the probability that a piece of software continues to perform a particular function over a particular period of time and under stated conditions.

In software terms, reliability is a difficult concept. If a piece of software is executed in the same environment and with the same data then it will always behave in exactly the same way. So what is this concept of reliability? How should it be interpreted? The key is to consider software running with a variety of different sets of data and perhaps under different sets of conditions. The variations in input and environment serve to create a domain within which it makes sense to discuss reliability. It is then assumed that elements of this domain are selected at random and with equal probability. The measure of reliability of software is not necessarily static; in the context of computing and software systems, it can and often does change with time.

It is important to draw attention to related issues of safety, security, etc. In these high-integrity systems, reliability is the most important consideration. But reliability theory itself does not take into account the consequences of a failure. High-integrity considerations must take account of what happens in the event of software failing; reliability considerations, typically, do not address these. See [2] for further information.

Software Testing

In terms of achieving the desirable goal of being able to guarantee the reliability of software, the activity of testing that software is crucial. In general terms, the more effective this process is, the higher the likelihood that the software will achieve desirable levels of reliability.

The basic objective of software testing is to uncover faults in the software, typically by attempting to cause failure of the software or by showing that the software fails to meet its requirements. The faults should then be removed, leading hopefully to improved reliability of that software. For the purpose of testing the software, it is necessary to have a clear statement about the intended effect of a piece of code; both functional requirements and nonfunctional requirements need to be clearly stated. Testing will typically show the presence of errors but not their absence.

There are additional important objectives to testing. These generally relate to the considerations of the overall quality of software and software process. They emerge from considerations related to the careful planning and management of the testing process. For, in general, the testing process will require the selection of a suite of test cases and questions will naturally arise about the effectiveness of the particular choices. Thus, measuring the number and the type of errors provides measure of the efficiency of the testing process, as well as messages about the nature of defects and the effectiveness of the software development regime; monitoring the trends in terms of the failures over time can provide an indication of reliability of the software; monitoring the arrival of defect reports from clients provides an indication of the readiness of the software for release; and so on.

The scope of testing is considerable and there are many important aspects to it. Apart from checking that results are as expected, there is checking to obtain or ensure some measure of reliability, there is testing for security (e.g., of firewall software), testing for performance (e.g., timing), testing to ensure safety – which can be vital in safety-critical or other forms of high-integrity software, and so on. The choice of tests is often nontrivial and should be guided by a careful strategy incorporated within a test plan. For instance, ensuring some measure of independence between different tests may be an important feature to encourage efficiency in the testing process, retaining

tests so that they can be replayed quickly is likely to be important for large systems, and so on.

Within the range of techniques that can be applied for the purposes of testing a particular program, a range of checks can be performed:

Static checks are carried out on code without running or executing it. The checks include, for instance, whether the code exhibits syntax errors or type-checking errors, whether it compiles, etc. Checking on adherence to programming standards – an activity that may utilize information about metrics – also falls within the realms of static checks.

Dynamic checks necessitate running or executing the code. They include matters such as checking that the output is as expected, performance requirements are met with, run-time errors do not occur, etc.

The terms *verification* and *validation* are complementary and describe general approaches to dealing with aspects of the testing of software. Basically they refer to those activities that are used to ensure that the software product is being built correctly and corresponds to the needs and desires of the customer.

Formal verification is defined as the application of mathematical techniques and mathematical argument whose purpose is to demonstrate the consistency between a program and its specification.

Informal verification is defined as the process of determining whether the products of a given phase of the software development cycle fulfill the requirements established during the previous phase.

Validation is defined as the process of evaluating software at the end of the software development process to ensure compliance with software requirements.

Of the above, formal verification is somewhat mathematical (see [3, 4]) but it offers interesting insights into the nature and the structure of programs.

Testing, validation, and verification can typically be applied to all stages of the software life cycle. There is a general rule that the earlier an error is caught the better it is. This is important for productivity, cost, and quality, and as such impinges on the morale of software engineers. For success, typically, it is better to have a more formal approach; here “formal” is intended to convey a carefully thought out and disciplined approach that may or may not entail mathematics.

All validation and verification activities should, if possible, be carried out by a team independent of the one undertaking the development of the software.

In this way, problems arising from preconceived notions, ideas, or assumptions tend to be addressed. Verification has the capacity to overcome some of the limitations of testing.

Testing should not be an afterthought; it should be planned and addressed at all stages of the software life cycle:

- **Requirements stage**

Initial test cases must be generated from the specification and these need to be accompanied by the expected outcome; these can also be retained for final testing. The consequences of requirements (which ought to be testable) must be investigated. When this is being performed, the general testing strategy for the final product can be developed and this will include verification requirements, test schedule, milestones, and analyses.

- **Architectural design stage**

Plans for test configurations and integration testing are established. Verification requirements are expanded to include and highlight individual tests.

- **Detailed design stage**

Test data and procedures should be devised to exercise the functions introduced at this stage.

- **Construction stage**

Static analysis tools can be used for detecting code anomalies, identifying unreachable paths, collecting metrics, etc. Dynamic analysis tools are used for testing performance. Informal testing of individual modules may take place at this stage.

- **Operation and maintenance**

Fifty percent of the life cycle is typically involved in this activity. Careful management of tests and the results of tests will reduce the necessary effort here and enable subsequent tests to be carried out more efficiently. For example, after modification, only those modules that have been changed may need testing; regression testing (see below) involves performing some or all of the previous tests to ensure that no errors have been introduced; usually this process is automated with previous results having been retained for easy checking.

For certain phases of the life cycle – especially the earlier ones – technical reviews are often used. These typically take the form of walk-throughs or

inspections with a group of experts analyzing documents such as specifications, designs, etc., for quality attributes (e.g., completeness, consistency, etc.). For guidelines on the conduct of technical reviews, see [1].

Different Approaches

A wide variety of different approaches exist, some being relevant to particular stages of the life cycle. We highlight the main approaches.

Functional Testing

Functional testing (*see Reliability Data; Human Reliability Assessment; Fault Detection and Diagnosis*) – also called *black box testing* – is concerned with testing the developed software against its stated functional requirements. The technique is also used to validate the performance and behavioral requirements of the final, integrated system. Test data is derived from the specification without reference to the internal structure of the code. Implementation details are not considered. The specification will typically be described precisely in natural language, using a model-based notation such as Z or vienna development method (VDM), or using an algebraic or axiomatic notation. In all these cases, certain inputs can be chosen and the specification will identify the expected output. The types of errors typically found include identifying incorrect or missing functionality, interface errors, data manipulation errors, performance errors, and start-up and shut-down errors.

To improve the rigor and coverage of functional testing, several techniques that aim to ensure optimal usage of the testing effort can be employed. These include the following:

- **Equivalence partitioning**
Here the test data for the program is partitioned into a small number of equivalence classes on the basis of description or specification. Then test selected representative members of each class, and view these as representative of the class as a whole.
- **Boundary value analysis** (*see Potency Estimation*)
Often software fails at the boundaries, e.g., 0, 1, *nil*. Test data is typically selected on the basis of this (*see Potency Estimation*).

- **Cause–effect graphing**

The input/output domain is partitioned into various classes, each of which produces a particular effect. For instance, a user issues a particular command, a particular message is received by the system, a sensor provides certain information, and so on. This is typically used to derive test cases, each of which exercises a particular possibility.

- **Comparison testing**

When there are different implementations of a particular task, the same test can be applied to each version and then errors are indicated by different outputs from two or more versions. This is especially relevant in the context of N-version programming, for instance.

- **Grammar-based testing**

Here the input is described by means of a grammar and testing is then based on exercising all possible combinations of alternatives within the grammar. In such cases, automatic methods can be used to ensure coverage of the range of possibilities. This is particularly relevant for compilers and other syntax directed tools.

Structural Testing

Alternatively called *white box* testing, this approach takes the structure of code into account. Test cases are chosen so that, for instance,

- all independent paths through a program are tested; this leads to considerations about the effectiveness of testing in terms of path coverage;
- all statements in a program are executed, and hence leads to considerations about statement coverage.

Typically, logic and design errors are identified, often because they are on paths not accessed by functional testing; note that infeasible paths may exist, and it may not be possible to exercise certain statements. In addition, erroneous output, performance violation, exception raising, etc., are all addressed.

To allow the testing of partially complete software, devices such as stubs can be used; to prevent catastrophes from erroneous software, test harnesses can be used; and simulated environments often have a role to play.

Controlling Complexity

The complexity of the testing process is typically controlled by testing units individually; hence this is

called *unit testing*. *Integration testing* (see **Repeated Measures Analyses; No Fault Found**) is then performed by gradually combining units to form larger and larger entities and testing them. Eventually, an entire system is tested. These considerations give rise, naturally, to the following:

- **Integration testing**

This involves integrating and testing parts of, and eventually, the whole system: here, checks are made on matters such as functions being defined, interface compatibility, e.g., whether parameters are appropriate, and so on. Both bottom-up and top-down approaches are possible. This depends on careful structuring of the code with careful testing of modules, and then testing of interfaces between modules. In the particular case of object oriented programming, integration testing leads to testing of classes, then classes that tend to cooperate, and so on.

- **System testing**

Here test cases should be selected from an input population, which reflects the intended usage of the software; for safety-critical software, the software must be tested in situations where correct functioning is critical; reliability estimation is typically needed and this is often done using statistical testing techniques.

- **Acceptance testing**

This happens with the customer prior to the software being released to the customer.

- **Regression testing**

In many areas of software development it is desirable to retest software to ensure the absence of new errors and usually enhanced reliability. In particular, this should happen on completion of maintenance activity. Ideally, the whole process of retesting should be automated as far as possible, e.g., rerun various tests. Regression testing addresses this. To accomplish it, typically there is a need for careful management of all test information, for the use of stubs and drivers, for harnesses within which to carry out tests, and so on. All these entities ought to be brought under control of a configuration management and version control regime.

- **Performance testing**

Especially in areas such as real-time systems, safety systems, security systems, etc., there is a need to test the performance of software. Thus,

much of whether it meets stringent timing requirements even in situations of stress will depend on the criticality of the software, but considerations of worst possible scenarios are often used to guarantee levels of performance. Such matters, of course, must typically be considered when the software is being designed. For, an error in one piece of software can contaminate another piece of software and so isolating the software in particular machines can help in resolving certain difficult situations.

In some cases software systems can continue to operate in a degraded state, e.g., because of some failed component (software or hardware). In such cases it is important to carry out rigorous testing of the degraded state to ensure that it does perform to the expected standards.

Different Testing Scenarios

In the broad spectrum of testing, there are many levels of complexity. Testing simple programs that begin, execute, and halt is relatively straightforward. For testing programs that retain some memory from previous data, two possible approaches are apparent: choose input and internal state randomly; alternatively reinitialize each time and use a sequence of inputs for each test case. For testing real-time programs, each input is typically a stream of values; difficulties occur if long periods of operation are typical, in which case steps are taken to curtail testing time while remaining realistic. For safety-critical system testing, criteria remain the same as above but choice of inputs and therefore tests must cause the system to become active and critical.

It has been estimated that high-rate errors (i.e., those that occur most frequently when a system is being used) account for nearly two-thirds of software failures reported by users. By concentrating the testing effort on those parts of the software that are anticipated to be invoked most frequently during usage, these failures can be identified and removed first.

Statistical Testing. It is often the case that certain inputs to software can be expected to occur more often than others. In general, a frequency distribution or *operational profile* (see **Reliability Integrated Engineering Using Physics of Failure**) can be used to describe the expected frequencies with which certain inputs can occur, and these capture the patterns

of use. In practice, deriving an accurate operational profile is not always straightforward, yet can be accumulated over time by monitoring use. Alternatively, estimates can be derived from requirements, interviews with users, marketing studies, and other such user-centered sources.

Statistical testing involves testing software according to the way in which users intend to use it. It focuses on the external behavior of the entire integrated system and ignores internal structure. The intention is to uncover problems with the states that are likely to occur most frequently or to test more thoroughly those parts of the system that are likely to cause most damage. So an important purpose of statistical testing is to provide measures or estimates of the reliability of software. In the normal way, failure data can be collected and used to estimate the reliability of the software system.

Error Guessing. One of the major problems with software is that many faults do not become apparent until a certain combination of events results in an incorrect state during execution and a failure occurs. Owing to the complex nature of software systems, it is usually not viable to test every possible combination of events and therefore it is possible to miss such erroneous states. Yet, providing guesses of such combinations (as well as other such possibilities) can often be effective in terms of identifying fault situations.

Checklists. For certain kinds of systems, e.g., evaluating a human–computer interface, certain tests or checks reflecting accumulated best practice can be captured in checklists.

Testing and Safety-Critical Systems. In the case of safety-critical (*see Systems Reliability; Reliability Data; Human Reliability Assessment*) and other high-integrity software, guarantees typically need to be provided about measures of reliability (as well as safety, etc.). Modeling reliability growth and trends over time and during the continual evolution of the software is one approach. But effective approaches based on Bayesian statistics are emerging to demonstrate that the probability of failure is very low (*see* [5]). These studies provide additional pointers to choices of test cases.

Assessment of Testing Methods

Most methods are heuristic and lack a sound basis and intuitive ideas about the patterns of usage of

software (e.g., which functions will be invoked most frequently) are often inaccurate.

Functional and structural testing approaches are typically seen as complementary, rather than alternative, to each other. They uncover different types of errors. Likewise, performance testing is again complementary to these. So there is an onus on devising a cocktail of approaches to testing that in some sense is minimal and yet seeks to test and exercise the software in realistic and meaningful ways.

In employing the various testing methods, it is customary to develop a suite of tests. This might deal with separate classes in equivalence partitioning or separate routes through a program. When the various tests are viewed as a collection, it is possible to talk about the adequacy of that collection. For example, the issues could be what percentage of the possible equivalence classes and what percentage of the routes through a program have been tested. These considerations lead to measures of the *adequacy* of the various forms of testing. A good overview of the issues is contained in [6]. But it is not uncommon to seek 100% path coverage and around 80–90% statement coverage.

Mutation Testing. It is important to obtain some feeling for the number of faults in a program on the basis of an analysis of the outcome of testing.

Mutation analysis – seeding faults – can be used to assess the effectiveness of testing methods. This is usually accomplished by seeding or inserting faults into the software and basing analysis of the effectiveness of testing on the ability of the tests to discover the seeded faults. Typical examples are replacing + instead of –, or > instead of >=, for instance.

Mutation testing involves making changes of this kind (temporarily) to the program, running it with the same previously used test data, and then checking the final results.

If these results are the same as those before the errors were introduced, then some investigation is called for.

Test Planning

There are many ways in which software can be tested, the choice being dependent on characteristics such as the size, criticality, and complexity of the system as well as specific customer requirements. Careful test

planning is needed to maximize the coverage and reap the benefits of the testing activity.

In this it is important to remember the mutual dependencies and contributions of all the various components – hardware, system software, compilers and other tools, devices, etc. In all testing scenarios, the choice of inputs must be representative of the intended usage of the system.

Planning testing is typically important. It is desirable to be comprehensive on the one hand, and yet efficient and effective on the other. So tests should be chosen with care in such a way that they do not duplicate previous tests, but rather they explore some novel feature of the software.

In planning testing, much depends on the criticality of the application. But it is typically desirable to include black box testing and white box testing, and to proceed using unit testing, integration testing and system testing taking into account testing of any performance issues, testing usability and performance as well as testing of the user interface. In carrying out these considerations about the adequacy of test, coverage needs to be taken into account; ideally aim to have 100% coverage of equivalence classes, routes through a program, and so on. In addition, data verification and software performance modeling are frequently practiced.

Software Reliability Modeling

A model of reliability *versus* time can be obtained for a software product by measuring certain trends over the time during which the software is evolving. These can be plotted to produce a graph, for instance. Extrapolating is then one means to predict the future reliability of software. Failure data – possibly generated by tests, and followed by the use of statistical testing is relevant – are used to model current behavior and then statistical inference may be used to predict behavior at a specified time in the future. The problem is to decide which model provides a close enough picture of reality to give accurate results.

The purpose of reliability modeling is to predict reliability; current failure data is collected and used to infer future behavior. This is typically done by finding some mathematical means of capturing behavior and then using mathematical and statistical techniques to provide answers to common questions. Examples of the use of such predictions include the following:

determine at what point in time a particular level of reliability will be reached; or, determine what level of reliability will have been reached by a certain point in time.

Several models have been developed. However, these have tended to give often largely different results for the same data sets. An appropriate model must be selected according to the circumstances and the context, and a certain amount of judgment must be used to determine which model gives the best results.

The Logarithmic Poisson Model. In this model, initial modeling of failures is conducted with respect to execution time, that is, the actual processing time of the software on the computer. This can then be translated into calendar time, which humans experience. Calendar time is more useful to software engineers and managers in terms of predicting cost, schedule, and resources.

In certain applications of reliability theory it is customary to assume a constant failure rate. Then the reliability R can typically be expressed in the form

$$R = e^{-\lambda t} = e^{-t/T} \quad (1)$$

where T is mean time between failures, or λ is mean time to failure, whichever is most appropriate in the circumstances. Such a model, assuming the parameters are known, can then be used to predict reliability at some point in the future or conversely predict a time at which a certain level of reliability can be expected. In the context of software, the basic assumption of constant failure rate is unrealistic.

Jelinski–Moranda model

The basic Jelinski–Moranda model (see [7]) has been extended by Musa (see [8]). This is based on the notion that if a software contains n faults, then each of these faults contributes equally to failures; moreover the repairs done to remove faults are assumed to be perfect and, in particular, they do not introduce further faults. There is plenty of evidence to support the view that some of the assumptions here are quite unrealistic. Nevertheless, the model is fairly commonly used.

Littlewood Model. This model (see [9]) is based on the idea that the nature of the contribution from the individual faults to the reliability profile is random. A consequence of this is that the effect of removing

a fault will be unpredictable and dependent on the particular fault.

Again, since the effect of faults is unpredictable, there is something little unrealistic about the basic assumptions underlying this model.

Other Models. For a review of other models of reliability, see [10].

Static and Dynamic Analysis

Static Analysis Techniques

Static analysis is a process that involves analyzing software products without actually performing execution. Static analysis techniques have been applied mainly to source code, although it is possible to analyze other products (if these are expressed in formal or semiformal notation). Specifications, designs, and fault trees, for instance, all represent possible descriptions of a system and are hence valid for analysis. However, the analysis of source code is the most important target for analysis because this is ultimately what determines the behavior of the system.

In the course of this, data is collected and typically this relates to various aspects of the structure, complexity, size, and other internal attributes of the software. Since the code is not executed, issues involving the operational environment are not addressed.

Manual approaches to such analysis and data collection tend to be error prone, time consuming, mundane, and difficult to replicate. So automated software support is typically used. Such tools can greatly increase the efficacy of the static analysis process.

Static analysis can involve different levels of complexity and can be performed for different objectives. For example, it can be used to uncover faults in software, evaluate the quality of code, demonstrate that a module meets its specification, or prove the correctness of an algorithm. These activities may be performed at different stages of the life cycle.

One way in which a program can be shown to conform to its specification is known as *symbolic execution*. This involves deriving a symbolic expression for each variable from the code that affects it. The resulting expression can then be checked against the expected expression from the specification.

Some of the tasks that can be performed during static analysis include the following: detect anomalies, for instance, variables have been defined but

not used, variables have not been initialized, non-portable language features have been used, e.g., use of a particular language feature; apply certain techniques to derive *metrics* that can be used to draw conclusions about the quality of code; ensure coding standards have been adhered to (e.g., guidelines on module size, comment density, or the use of GOTO); check adherence to programming standards (e.g., the rules of structured programming); prove that software conforms to its specification (verification).

Metrics related to software structure are often defined on abstract models of the software rather than on the software itself. *Call graphs* and *control flow graphs* are examples of such models.

A *call graph* is a graphical representation of the interactions that exist between a set of modules. Examples of metrics that can be derived from a call graph include the number of recursive modules, hierarchical complexity (number of modules on each level of the graph), testability measures, and the average number of calls per module as well as highlighting calls that have not been invoked.

A *control flow graph* is a graphical representation of the algorithmic structure of a module. That is, it depicts the flow of control from the entry point to the exit point of the module. Metrics that can be derived from the control flow graph include the number of nodes, number of edges, the depth of nesting, and adherence to structured programming rules. This also highlights the paths that have not been tested, e.g., unreachable paths, as well as poorly structured or “spaghetti code”.

Possibilities for more detailed forms of analysis also exist; but methods such as formal proof are often hindered by the lack of formal software development in industry and suitable verification systems.

In general, there is a lack of standards in the area, which means that objectives are often unclear and benefits are difficult to quantify. Moreover, some language features (such as the aliasing of variables and the use of pointers for parameter passing in C) cause inherent problems for analysis tools. It is possible to subsume some forms of static analysis into the language and compiler. Certain advances even suggest the incorporation of more detailed forms of analysis, such as proof of code against its specification, but this would require rigorous specification and design information. This must be taken into account early in the life cycle.

Tools for static analysis typically exhibit an architecture that typically involves a front end that collects data from specific languages and expresses it in a general notation and a back end that performs calculations on data and displays results. The output can typically be in the form of metrics, graphs, or text. But generally tools tend to vary in scope and applicability. It has been observed that metrics may be defined differently for different languages and across different tools.

Dynamic Analysis

Dynamic analysis can be viewed as a means of collecting and analyzing data about the software during the testing activity. Dynamic analysis techniques that require extra instructions to be inserted into the code are called *intrusive*. *Nonintrusive* analysis usually involves running the analyzer on a separate system or environment.

Dynamic analysis involves analyzing software and in so doing using the results of execution. The software must be executed according to its intended usage. This can take various possible forms:

- **Program instrumentation**

This involves inserting various kinds of *probes* into the code to provide a profile of execution – for example, coverage analysis – statement counters are used to determine path or statement coverage by test data; dynamic assertions – predicates concerning values of variables that must hold at a certain point are inserted.

Inserting probes may increase the size of the code significantly, leading to problems, especially in real-time systems.

- **Failure analysis**

An important, though obvious, approach is the collection of failure data for the executing software system. This can be used in reliability estimation.

- **Performance characteristics**

Dynamic analysis is also concerned with the performance characteristics of software such as response times and resource usage.

As with static analysis, the facility and efficiency with which dynamic analysis can be carried out is enhanced dramatically through the use of good support tools.

Tools Development

A range of tools can be used for static and dynamic analysis and they vary in their capability. As expected, static analysis tools are concerned with static analysis of source code and certain facilities based on the same; dynamic analysis tools are based on testing and the conclusions that can be drawn from the test. On occasion, both sets of facilities can be found within one tool.

Typically the main objective of static-analysis-based tools is to provide an evaluation of the quality of a developed software system. A front end may accommodate different languages (e.g., Ada, C, Pascal) but the back end will be similar for each. But typically code is the only product analyzed. Metrics of various kinds can be gathered automatically. Possible metrics include the following:

- **Metrics gathering**

A range of metrics can be gathered automatically. These are related to various aspects of the software such as size, structure, complexity, and documentation. The possibilities are many but, for instance, these could include the number of statements, program length, number of noncyclic paths, number of jumps, comments frequency, average statement size, and number of input/output points. Experience will tell that, for good quality programs, these various metrics should normally lie within a particular range, the latter being defined by a low value and a high value.

The visual display of metrics data allows them to be comprehended more easily and areas for attention to be identified. The concept of a Kiviat diagram or graph is often used; this provides two concentric ellipses, the outer one defined by the high values of a range, and the inner one defined by the lower values. Then when actual values are plotted, they all ought to lie between the two ellipses. Any failure to do so is highlighted. This allows the results for different metrics to be compared, as well as describing which values are beyond their limits.

- **Classification analysis**

On the basis of ranges as defined by the low and the high levels, conclusions can be reached about different sections of the code. To be precise, different sections may be classified as acceptable, readable, testable, or self-describing, thus

implying something about the existence of commentary of some kind. And other parts may be flagged as unusual or worthy of further scrutiny.

- **Quality factor indicator**

On the basis of above classification, sections of the code can be categorized and a report produced automatically. This can identify good sections of code and also sections that merit further attention.

Typically, e.g., in the form of a pie chart, a report can be produced showing the distribution of components according to their quality factor. For instance, a report might indicate that, of the existing modules, about 69% were accepted and about 10% required further documentation. A list of modules that make up each percentage can then be obtained.

Such reports are useful as a management tool to control activity. This is often known as *management by metrics*.

Dynamic analysis can take place during the testing phase to supplement this activity, or by rerunning the test cases and thus simulating the activity. The latter method is used when the product is being evaluated, for example, for efficiency or reliability, by an independent body.

Toward Reliable Software

A whole host of factors contribute toward the desirable goal of achieving reliable software. Included in these is a set of consideration such as

- capturing the role that the software has to play in precise detail, i.e., its functionality, and importantly, also the considerations such as performance;
- protecting the software from interference from elsewhere, e.g., from malicious software such as viruses but also from deviant behavior of other software in the vicinity (which may contaminate, for instance);
- adopting certain approaches to design, e.g., those associated with human–computer interface considerations and these in turn may be associated with human conditions such as color blindness, or the requirement that no single error should bring the system to a dangerous state; there are design techniques in software engineering that accommodate such considerations;

- recognizing the role of certain software process issues, e.g., choosing a compiler or software tool of high quality.

Yet despite best endeavors with respect to the design of software and its testing, it is recognized that faults may still remain in software. Software fault tolerance techniques exist to provide another additional kind of approach.

Software Fault Tolerance

There are various approaches to software fault tolerance.

Exception Handling. Typically, as a result of execution of a program some fault occurs; in many modern languages (e.g., Java, Ada) this causes a flag or *exception* to be raised. Examples of exceptions are division by zero, arithmetic overflow, certain kinds of inappropriate input, etc.

There is also the potential for the fault to be handled either at the component level or at the system level as appropriate. The “repair” (to the system state, *not* the program) is dealt with by an *exception handler* – a piece of code to which control is passed, and which is then executed prior to control returning to the main program sequence. Normally, exception handling is done by that designer who is most likely to understand the ramifications of an exception being raised. To illustrate, if an erroneous piece of data is passed to a component, this is likely to be handled at the system level; or, if an arithmetic error leading to overflow occurs in a component, it is likely to be handled at the component level.

N-Version programming. N-version programming (also known as *diverse programming*) involves a variety of N ideally independently and diversely produced routines satisfying the same specification being written in isolation from one another. Though this is the principle, achieving design diversity in practice is not always straightforward; see [11]. When a result is sought, voting takes place and broadly speaking, the result gaining most votes is the answer delivered.

There are different strategies for choosing N and for deciding how the voting should proceed. For instance, a system may be deemed to operate satisfactorily if at least three out of five components agree. Typically $N \geq 3$, although $N = 2$ is sometimes used

for systems such as railway signaling systems; in the event of disagreement in such systems, a fail-safe situation needs to be established, e.g., all the trains stop.

Voting typically has the effect of masking single faults. If $N \geq 5$ there is the potential to mask more than one fault. So voting provides adjudication and has the capacity to detect errors. Indeed, the behavior of individual components can be monitored and a judicious approach to replacing unreliable components by more reliable components can lead to ever more reliable overall systems. This same strategy can also be used to gradually introduce new technologies and new developments in a controlled manner by monitoring their performance.

Possible damage caused by faulty software must be limited and must not extend to other versions. So these should not affect global variables, e.g., run each version on a separate processor if necessary. Recovery is effectively achieved by ignoring a version.

Voting must be simple (e.g., majority) and ideally exact. The situation is often complex. If real numbers are involved, then rounding errors may be present. Then inexact voting results, and this increases complexity. If large arrays are involved, that can be time consuming and can lead to problems of performance. And, of course, programs can fail by giving the wrong result, looping indefinitely, failing to achieve performance targets, behaving in a destructive manner, and so on.

Concurrent implementation of N-version programming is often possible.

Recovery Blocks. In using recovery blocks, again a number of routines are written (in isolation) ideally using different approaches. In addition, an acceptance test is provided and the first routine to satisfy the acceptance test is selected. The structure of a recovery block typically takes the form of a number of alternative actions to be taken. Their format is captured in the following:

ENSURE	Acceptance test
BY	Primary module
ELSE_BY	First alternative
ELSE_BY	Second alternative ...
ELSE_BY	Nth alternative
ELSE_ERROR	Exception raised

The primary module is the preferred option. The sequence of events is as follows: on entry to the

recovery block, a recovery point P is established; on reaching this, record the state S of the system. The primary module runs and produces a result; if this is accepted by the acceptance test, then normal execution can continue, i.e., typically, the statement following the recovery block is executed; if the acceptance test fails, then recreate the recorded state S of the system at the recovery point P and invoke the next alternative. Continue until the result passes the acceptance test or there are no alternatives remaining. In the latter case, an exception is raised.

The complexity of acceptance tests has to be limited; otherwise they may introduce more faults into the system, e.g., violation of performance targets. The design of good acceptance tests is not a straightforward matter; most do not guarantee correct behavior and additional checks are needed if that is a requirement. Acceptance tests need to be simple, especially in real-time systems where there need to be guarantees of performance. On occasion, null alternatives can be used in recovery blocks to provide an effective fault tolerant strategy for real-time systems.

It is possible to have different strategies in relation to the alternatives within the recovery block. For instance, all alternatives are independent and of equal status; alternatively, prioritize the alternatives (e.g., for efficiency reasons), or finally have degraded alternatives.

In the use of recovery blocks, it is desirable that execution of a routine should not adjust the environment, since the effect of that execution may have to be discarded. Thus reestablishing the state S of the system must be simple.

Comparison of Mechanisms. A comparison can be offered between the N-version programming and recovery block concepts. In N-version programming, the voting check is a powerful mechanism for error detection but inexact voting causes complexity in the software and this can introduce new faults. In recovery blocks, acceptance tests are often difficult to devise and often do not provide a guarantee of correct execution. Generally, in both cases, the tendency exists for different version of components to access common data structures and this tends to compromise the independence of designs.

The principles behind software fault tolerance can be applied equally at the component level and at the systems level. When applied to individual components, redundancy is tailored to the specific

needs and structure of those components. At the component level, redundancy can be effective but can only deal with those faults specific to the component in question. At the systems level, redundancy is more expensive because of its generality but can consequently deal with a greater range of faults.

References

- [1] Pressman, R.S. (2005). *Software Engineering: A Practitioner's Approach*, 6th Edition, McGraw-Hill, New York.
- [2] Littlewood, B., Bloomfield, R., Courtois, P.-J., McGettrick, A., Randall, B., Bainbridge, L., Cribbens, A.H., Hall, R.S., Hamilton, V.L.P. & Oddy, G. (1998). *The use of Computers in Safety-critical Applications*, Final report of the study group on the safety of operational computer systems, Expanded report from the ACSNI Committee of the Health and Safety Commission, HMSO.
- [3] McGettrick, A.D. (1982). *Program Verification using Ada*, Cambridge University Press, Cambridge.
- [4] McGettrick, A.D., Traynor, O. & Duffy, D. (1993). *Verification Aspects of the PROSPECTRA Project*, Springer Lecture Notes No. 680 in Computer Science, Springer, pp. 129–144.
- [5] Woof, D.A., Goldstein, M. & Coolen, F.P.A. (2002). Bayesian graphical models for software testing, *IEEE Transactions on Software Engineering* **28**(5), 510–525.
- [6] Zhu, H., Hall, P.A.V. & May, J.H.R. (1997). Software unit test coverage and adequacy, *ACM Computing Reviews* **29**(4), 366–427.
- [7] Jelinski, Z. & Moranda, P.B. (1972). Software reliability research, in *Statistical Computer Performance*, W. Freiberger, ed, Academic Press, New York, pp. 465–484.
- [8] Musa, J. (1975). A theory of software reliability and its applications, *IEEE Transactions on Software Engineering* **SE-1**, 312–327.
- [9] Littlewood, B. (1981). Stochastic reliability growth: a model for fault removal in computer programs and hardware designs, *IEEE Transactions on Reliability* **R-30**, 313–320.
- [10] Littlewood, B. (1991). Software reliability modelling, in *The Software Engineer's Reference Book*, J.A. McDermid, ed, Butterworth-Heinemann, Chapter 31.
- [11] Littlewood, B., Popov, P. & Strigini, L. (2001). Modelling software design diversity – a review, *ACM Computing Surveys* **33**(2), 177–208.

ANDREW D. MCGETTRICK

Soil Contamination Risk

Soil is generally defined as the unconsolidated mineral or organic material on the immediate surface of the earth that serves as a natural medium for the growth of land plants. Additional functions of soil should be fully recognized, such as harboring groundwater supplies, producing and recycling organic material, storing and recycling organic waste [1]. Soil is a “vital” medium hosting most of the biosphere and is essentially a nonrenewable natural resource owing to the very slow processes for its generation. Contamination is one of the major threats to soil conservation, others being erosion, sealing, organic matter decline, salinization, compaction, and landslides [2]. Soil-contaminating substances can be roughly divided into heavy metals, organic contaminants such as *persistent organic pollutant* (POPs) (see **Persistent Organic Pollutants**), and nutrients. Pollutants may enter the soil through direct application, as is the case with pesticides and fertilizers (included sludge amendments), accidents, spills, or improper uses of industrial waste (see **Hazardous Waste Site(s)**). Dry deposition of industrial air emissions or long-range transport of volatile substances can also result in soil contamination. Contamination may have an impact on most soil functions including the support to human living. Since soil is the interface between area, water, and rock compartments and the platform for human activities and landscape, the term *land contamination* is often preferred to that of soil contamination. The term *land* indicates the need for an integrated approach to contaminated soil, surface water nearby, and groundwater beneath. It also clearly refers to value placed on land and soil by humans: land is often a private property and soil is a part of the landscape.

Some key concepts in the management of contaminated soils are the following:

- Contaminated soil should be closely related to the management of air and water compartments (see **Air Pollution Risk; Water Pollution Risk**).
- The definition of sustainable solutions requires a proper consideration of the temporal and spatial dimensions.
- Long-term care objectives should be defined, taking into consideration the mobility and degradability of contaminants.

- A distinction is usually made between the historic contamination (inherited from the past) and the prevention of future contamination.
- With regard to the spatial dimension, the remediation of contaminated land is closely linked to spatial planning and is usually driven by the “fitness for use” concept.
- A distinction is usually made between local and diffuse contamination, the former being a problem to be solved by the landowners, and the latter being a problem related to multiple stakeholders and environmental protection strategies on a larger scale.

In the following sections the fundamentals of both human health and ecological risk assessment for soil contamination will be introduced. The application of risk assessment in the derivation of soil quality standard (SQS) and site-specific assessment will be described thereafter.

Human Health Risk Assessment

The human health risk assessment of contaminated soils implies the analysis of the likelihood and magnitude of adverse effects on human health arising from the direct or indirect exposure to toxic substances in soil.

The potential exposure pathways of human beings to soil contaminants include e.g., direct contact, inhalation, ingestion of particles, and drinking of contaminated groundwater. Which are the most relevant exposure pathways mainly depend on the physico-chemical properties of the contaminants.

The ingestion of contaminated soil is one of the main exposure pathways for immobile contaminants (those strongly adsorbed onto soil particles). Several studies proved that children can ingest significant amount of soil per day: e.g., the Exposure Factor source book published by the United States Environmental Protection Agency (USEPA) [3] reports measured amounts of ingested soil ranging from *ca.* 100 to 1400 mg kg⁻¹ day⁻¹ (lower and upper percentiles figures) with means of 383 (soil ingestion only) and 587 mg day⁻¹ (soil and dust ingestion).

For volatile contaminants, rather than evaporation in open air, the accumulation in enclosed spaces (through cracks in walls and pavements) appears to be the major exposure pathways.

The ingestion of contaminated groundwater appears to be the most relevant exposure pathway for relatively soluble compounds. Relatively soluble contaminants can leach through the unsaturated soil layer to the groundwater body below. If the aquifer is exploited for domestic uses, contaminants are ingested by drinking uses or inhaled during shower.

Diet exposure is by far the most relevant exposure pathway for bioaccumulating substances. Contaminants can be taken in by vegetables. They accumulate in roots, plants, and fruits and, consequently, in meat and dairy products.

In some cases, the contact between contaminated groundwater and surface water can lead to bioaccumulation in fishes.

Potential adverse effects on human health include irritation, sensitization, acute or chronic systemic toxicity, mutagenic and carcinogenic effects, and fertility and developmental effects, depending on the toxicological properties of the contaminants. A further insight on the assessment of human health risk *via* environmental exposure to hazardous substances in soil or waste matrices can be gained in the encyclopedia entry **Hazardous Waste Site(s)**.

Ecological Risk Assessment

Ecological risk assessment (*see* **Ecological Risk Assessment**) – applied to soil contaminants – concerns the following adverse effects [4]:

- effects on soil functions, and particularly on the capacity of soil to act as a substrate for plants and those organisms important for proper soil functioning and nutrient cycle conservation;
- effects on plant biomass production;
- effects on soil, above-ground, and foliar invertebrates, which are the basis of the terrestrial food web and cover essential roles as pollinators, detritivores, saprophages etc.;
- effects on terrestrial vertebrates;
- accumulation of toxic compounds in food items and through the food chain (as it is also considered in relation to human health).

Typical terrestrial receptors include soil microbes, invertebrates (e.g., beetles) including soil-dwelling invertebrates (e.g., insect larvae, worms, nematodes), plants, amphibians, reptiles, birds, and mammals. The exposure to soil contaminants includes direct

contact routes, like passive diffusion or active uptake, for plants and soft-bodied soil-dwelling organisms; incidental ingestion of soil particles for hard-bodied terrestrial invertebrates (e.g., beetles, centipedes) and, above-soil wildlife (e.g., small mammals and birds). Inhalation of highly volatile substances can be a significant exposure pathway for wildlife in the case of accumulation in enclosed spaces (e.g., lairs). In general, above-soil wildlife is mainly exposed to pollutants bioaccumulated through the food web (secondary poisoning) [5].

Ecological risk assessment of soil contaminants deals with many sources of uncertainty. First of all, it is very difficult to extrapolate effects to the ecosystem from the effects evidenced on a single species. Being true for the aquatic ecosystem, this concern is even more relevant for the terrestrial ecosystem owing to its overwhelming complexity. The risk assessment usually relies on results from laboratory tests with species representative of that environmental compartment. The extrapolation from monospecies tests to multispecies ecosystem is usually obtained by application of assessment factors, which is a very rough and conservative approach. How effects on a single species are reflected on populations, communities and the ecosystem is poorly known. Phenomena such as adaptation and pollution-induced community tolerance have been demonstrated in soil microbial communities. In this context, mesocosm studies can provide interesting results, but they are time consuming, costly, and not easily applicable to regulatory risk assessment.

The second source of uncertainty is associated with the poor understanding and modeling of environmental availability and bioavailability of soil contaminants. Exposure assessment to toxic chemicals in soil requires information on the concentration that is available to the organism in soil, which may be only one fraction of the total concentration, i.e., the bioavailable fraction. *Bioavailability* can be defined as the dynamic process that comprises the physicochemical desorption and the physiological driven uptake of the contaminant by the terrestrial organism. Soil pollutants can be sequestered by adsorption onto mineral surfaces and organic matter, diffusion in microporous media, and encapsulation. For organic pollutants, adsorption (and absorption) is usually related to the organic matter content, while for *heavy metals* parameters like pH, clay content, and

cation exchange capacity play a major role. Moreover, the bioavailability of metals depends on the chemical speciation of the dissolved fraction, which in turn depends on complex geochemical equilibriums. Refined methods for the estimation and incorporation of bioavailability in the risk assessment process are needed [6–8]. Related to the bioavailability concept is the need to normalize ecotoxicological data according to the type of soil used in the test. For comparison purposes the results from various tests should be recalculated for standard soil characteristics (e.g., 10% organic matter and 25% of clay).

The third relevant source of uncertainty is the poor availability of ecotoxicological data about terrestrial organisms as compared to aquatic organisms. This is usually overcome by applying equilibrium partitioning methods to aquatic organism ecotoxicological data. This method relies on two pragmatic (but weak) assumptions: that bioavailability, bioaccumulation, and toxicity are closely related to pore water concentrations and that sensitivities of aquatic organisms and terrestrial organisms are comparable. The poor availability of data significantly limits the possibility of applying probabilistic effect assessments, like the derivation of species sensitivity distributions (statistical distribution describing the variation in toxicity of a certain compound or mixture among a set of species) [9].

The fourth source of uncertainty is the oversimplification of the compartmental approach. The distinction between soil and above-ground compartments does not properly address the complexity of the terrestrial ecosystem [10].

Derivation of Soil Quality Standards (SQS)

Environmental quality standards for soil contamination have been issued in several countries and are usually based on quantitative risk assessment methodologies.

For the derivation of SQS, the risk assessment is applied to standard scenarios, including exposure pathways of concern, types of human and ecological receptors, and toxicological endpoints. In practice, the definition of standard scenarios incorporates relevant decisions about what to protect and to which extent. For example, in some cases only human health is considered and potential effects to ecological receptors are not. The contamination of groundwater is

not always included, but addressed separately. As for ecological receptors, whether to include microbial communities (biogeochemical cycles), soil organisms and plants, terrestrial wildlife, or aquatic organisms in (hypothetically affected) surface water bodies is a matter of choice. For human health risk assessment, potentially highly exposed subpopulations are not always considered: e.g., people consuming vegetables grown on contaminated soils or children with pica behavior (voluntary ingestion of significant amount of soil). Besides the definition of conceptual exposure models, conservativeness of exposure input values may also vary according to the concern for that specific exposure pathway. For toxicological and ecotoxicological assessment, the definition of endpoints of concern is usually inspired by the greatest conservativeness. Variability can be found on setting assessment factors and protection levels in the derivation of predicted no-adverse-effect concentrations. For example, when species sensitivity distributions are applied in ecological risk assessment, acceptable concentrations can be set to protect a certain percentage of species, usually ranging from 50 to 95% according to the objective of the SQS application. The rationale behind different choices in SQS derivation methods can be related to scientific arguments, but more often it depends on geographical, sociocultural, and political factors. It should be recognized, therefore, that SQS derivation methods are the product of sound science and conventional assumptions depending on risk societal valuing. Differences among procedures adopted by various national and international authorities may lead to wide variations (up to four orders of magnitude) of SQS values for the same contaminant, while the identification and distinction of reasons behind differences can be extremely difficult [11].

Site-Specific Risk Assessment

Site-specific risk assessment aims at estimating the risk for human health or ecological receptors at a specific site. Differing from SQS derivation methods, site-specific assessment takes into account local conditions that affect exposure or effects.

It should be borne in mind that soil contamination may concern small spots as well as very large areas or regions, according to the type of polluting source and contaminant mobility. The larger the contaminated area, the more important it is to characterize

the spatial distribution of the contaminant. Owing to spatial heterogeneity and the often accidental nature of contamination, concentrations of pollutants may vary remarkably over very short distances. Moreover, it is common and often unavoidable to have relatively small data sets derived from limited sampling in both the horizontal and vertical dimensions. Among different data interpolation methods to estimate the concentration distribution over the site, geostatistical methods like kriging [12] have the merit of being based on the proved spatial correlation of the observed concentrations and providing an indication of the uncertainty of the estimate [13, 14].

References

- [1] Doelman, P. & Eijssackers, H. (eds) (2004). *Vital Soil: Function, Value and Properties. Developments in Soil Science*, Elsevier, Amsterdam, Vol. 29.
- [2] European Commission (2006). *Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions. Thematic Strategy for Soil Protection*, COM(2006)231 final.
- [3] US EPA (1997). *Exposure Factors Handbook*, EPA600-P-95-002Fa, Office of Research and Development, National Center for Environmental Assessment, Washington, DC, available from <http://www.epa.gov/ncea/exposfac.htm> (accessed Oct 2006).
- [4] CSTEE (2000). *CSTEE Opinion on the Available Scientific Approaches to Assess the Potential Effects and Risk of Chemicals on Terrestrial Ecosystems*, DG SANCO, Brussels.
- [5] Fairbrother, A. & Bruce, H. (2005). Terrestrial ecotoxicology, in *Encyclopedia of Toxicology*, 2nd Edition, P. Wexler, ed, Elsevier, Oxford.
- [6] Jensen, J. & Mesman, M. (eds) (2006). Ecological risk assessment of contaminated land decision support for site specific investigations, Report no 711701047, RIVM.
- [7] Lanno, R.P. (ed) (2003). Contaminated soils: from soil-chemical interactions to ecosystem management, in *Proceedings of Workshop on Assessing Contaminated Soils, Pellston, 23-27 September 1998*, Society of Environmental Toxicology and Chemistry (SETAC), SETAC Press, Pensacola, p. 427.
- [8] Peijnenburg, W.J.G.M., Posthuma, L., Eijssackers, H.J.P. & Allen, H.E. (1997). A conceptual framework for implementation of bioavailability of metals for environmental management purposes, *Ecotoxicology and Environmental Safety* **37**, 163–172.
- [9] Posthuma, L., Suter II, G.W. & Traas, T.P. (eds) (2002). *Species Distributions in Ecotoxicology*, Lewis Publishers, Boca Raton.
- [10] Tarazona, J.V. & Vega, M.M. (2002). Hazard and risk assessment of chemicals for terrestrial ecosystems, *Toxicology* **181–182**, 187–191.
- [11] Carlon, C. (ed) (2007). *Derivation Methods of Soil Screening Values in Europe. A Review and Evaluation of National Procedures Towards Harmonisation Opportunities*, European Commission Joint Research Centre, EUR 22805, pp. 306.
- [12] Isaaks, E.H. & Srivasta, R.M. (1989). *An Introduction to Applied Geostatistics*, Oxford University Press, New York.
- [13] Carlon, C., Critto, A., Nathanail, P. & Marcomini, A. (2000). Risk based characterisation of a contaminated industrial site using multivariate and geostatistical tools, *Environmental Pollution* **111**(3), 417–427.
- [14] Gay, R.J. & Korre, A. (2006). A spatially-evaluated methodology for assessing risk to a population from contaminated land, *Environmental Pollution* **142**, 227–234.

Related Articles

Cumulative Risk Assessment for Environmental Hazards

Environmental Health Risk

Environmental Health Risk Assessment

Environmental Risk Assessment of Water Pollution

CLAUDIO CARLON

Solvency

Solvency is the ability of an insurance company to pay future claims as they fall due. This involves that the insurer must have sufficient assets not only to meet the liabilities but also to satisfy statutory financial requirements.

It is important for the supervisor to not only ensure the protection of the policyholders but also ensure the stability of the financial market.

The available solvency margin (ASM) is the difference between assets (A) and liabilities (L), i.e.,

$$ASM = A - L \quad (1a)$$

This definition, in terms of a solvency margin, was first given in [1]. It is necessary to distinguish between this actual margin and one required by the regulator. The theoretical requirement can in some jurisdictions be a minimum amount, and in others just a target or an early warning signal. If the insurer has an ASM above the required level, it can continue to be a going concern. Otherwise it has to be a dialogue between the insurer and the supervisor on what steps to take to get the ASM above this level. Some systems, e.g., the National Association of Insurance Commissioners' (NAIC) risk-based system in the United States, have an intervention ladder between the upper target and the absolute minimum level.

We assume a jurisdiction with two regulatory capital requirements;^a the target will be called *the solvency capital requirement* (SCR) and the absolute minimum will be called *the minimum capital requirement* (MCR). Ideally we have

$$MCR < SCR \leq ASM \quad (2)$$

Two concepts of solvency could be defined: (a) the position when liabilities could be paid at a breakup or at a runoff of the company or when it could be transferred to a willing partner and (b) the position when the company can pay all its debts as they mature (the going concern approach).

The first position may be obtained when $ASM \leq MCR$ and the second when $ASM \geq SCR$. This is also discussed in [2, 3].

One way of describing the strength of solvency is to evaluate the *ruin probability* (see **Ruin Theory**), i.e., the probability that the insurer having an initial $ASM \geq SCR$ will become insolvent during a

chosen time horizon $(0, T)$, see [4]. In the legal meaning insolvent means the break of the MCR, i.e., $ASM \leq MCR$. T is usually chosen according to the accounting period, i.e., at least a 1-year period.

In the literature there are a number of different formulas for the assessment of the capital requirement, especially for the MCR. In the beginning of the 1990s, a tendency was to seek rules that take into account all risks that an insurance company is facing. Systems of this kind are usually called *risk-based* and hence a risk-based capital requirement. A thorough discussion of the insurers' risks is given in [5, 6].

Historical Treatment of Solvency

The pioneering works done by Cornelis Campagne in the Netherlands at the end of the 1940s and by Teivo Pentikäinen in Finland in the beginning of the 1950s are important, as they introduced the solvency research for insurance undertakings, see e.g. [1, 2, 7]. The works done by Campagne became the framework for the solvency directives within the European Union (EU). The first EU directives framed in the 1970s have been amended twice during the 1980s and 1990s. On the basis of the discussion in reference [8], the EU Parliament adopted revised directives in 2002, Solvency I, and at the same time initiated a future risk-based system. This new system will be implemented within Europe around 2012. The basic ideas of Solvency II are given in [9].

As many countries in Europe had got the insurance directives during the 1970s implicating minimum solvency margins, research on solvency assessment was initiated. Work was done not only in, e.g., United Kingdom [10–13] and the Netherlands [3, 14] but also in Finland [15–17] and Norway [18–20]. The best reference and summary of different solvency assessment methods used in the middle of the 1980s are given in [3].

The research and works done were all stepwise toward a risk-based capital (RBC) approach. The NAIC introduced RBC systems in 1992 and 1993. At the same time the Canadian Office of the Superintendent of Financial Institutions (OSFI) introduced a risk-based system for life insurers. Risk-based systems were also introduced in Australia, Singapore, and Japan. While waiting for a new European system, different approaches were discussed or introduced in

United Kingdom, Switzerland, the Netherlands, Denmark, and Sweden, see e.g. [6].

Historically, there have been problems in comparing the available and required solvency margins between companies (and especially between companies in different countries). Assets have either been defined as book values or market values. The main problem, though, has been the technical provisions^b as these have included implicit margins to protect policyholders. These margins have been set by the actuaries and have reflected the prudence of the company. In the EU the incomparability of the provisions has been recognized and discussed.

Risk-Based Systems

The International Actuarial Association (IAA) and the International Association of Insurance Supervisors (IAIS) have issued standards and guidance regarding the assessment of insurer solvency.^c The main aim is to set up standards that make it possible to compare the valuation of the reserves, the capital requirement, etc., in different jurisdictions.

The total balance sheet approach was introduced by IAA (see [5]). It should not only recognize the asset and liability sides of the balance sheet, but also the interdependence between them and the impact of the SCR, MCR, and the eligibility of capital covering the requirements. The technical provisions are the reserves set aside to cover the liabilities the company face according to the insurance contracts.

In the banking supervisory system, *Basel II* (see **From Basel II to Solvency II – Risk Management in the Insurance Sector**), a three-pillar approach was introduced. A similar approach will be adopted within the EU Solvency II system. It consists of

- Pillar I: The quantitative requirements
- Pillar II: The qualitative requirements
- Pillar III: Statutory and market reporting.

The first pillar includes the calculation of the SCR according to a standard model (e.g., factor based), or the introduction of partial or full internal models. It also includes rules on provisioning and eligible capital. The second pillar focuses on the supervisors and their review process, e.g., a company's internal control and risk management, the approval of using partial or full internal models in Pillar I and its validation. The supervisor can also impose a company

to increase its SCR (capital add-ons) if it believes that the capital is not adequate or that the management is not sufficient. The third pillar includes the reporting to both the supervisor and the market promoting the market discipline and greater transparency, including harmonization of accounting rules.

A total balance sheet approach is based on common valuation methodologies. It should make optimal use of information provided by the financial markets [9] in getting market values when they exist ("mark-to-market") or getting market consistent values when they do not exist ("mark-to-model"). This is called *the economic value*. In a traditional actuarial valuation of present value a deterministic interest rate function is normally used, but in a market consistent valuation the deterministic interest rate is changed for a stochastic function, a deflator, reflecting the market price. The valuation should be prospective and all cash flows related to assets and liabilities should be discounted and valued at current estimate (CE). This valuation should use a relevant risk-free yield curve and be based on current, credible information and realistic assumptions. A risk margin covering the uncertainty linked to future cash flows over their whole time horizon can be added to the CE. An alternative in using a deflator function in market consistent valuation is to use a replicating portfolio in valuing the liabilities, i.e., a portfolio of assets that replicates the cash flow of liabilities most closely. One such technique is to use a valuation portfolio (VaPo), see [25–28].

A risk-adjusted CE of the liabilities is called *economical* technical provisions or market value of liabilities (MVL). The corresponding risk-adjusted assets is called *market value of assets* (MVA). The risk margins should be determined in a way that enables the insurance obligations to be transferred to a third party or to be put in runoff, cf. e.g. [9].

In the valuation procedure, hedgeable assets and liabilities should be valued by a mark-to-market approach, as any risk margins are implicit in observed market prices. Nonhedgeable assets and liabilities should be valued by a mark-to-model approach, i.e., a CE requires the calculation of an explicit risk margin. This is illustrated in Figure 1.

In terms of a total balance sheet approach, Equation (1a) is written as

$$ASM = MVA - MVL \quad (1b)$$

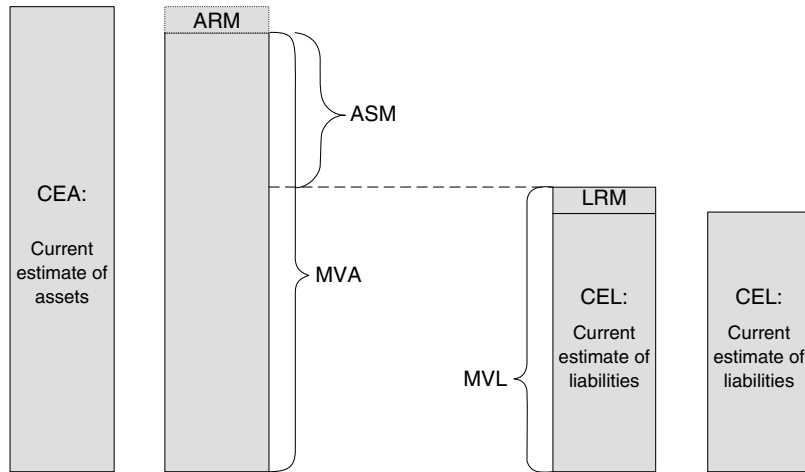


Figure 1 Current estimates of assets and liabilities are calculated, namely, CEA and CEL, respectively. An asset risk margin, ARM, is then deducted from the CEA to get the market values of assets, MVA. A liability risk margin, LRM, is then added to the CEL to get the market value of liabilities, MVL

The asset risk margin (ARM) is a risk margin deducted from the CE of assets owing to uncertainty in models and parameters. Valuations of derivative instruments may include uncertainty in the models used, see e.g. [21], giving rise to an uncertainty in the valuation.

As there is no liquid market for insurance risks, we need to use a mark-to-model approach to determine the MVL, i.e., to model a risk margin on top of the current estimate of liabilities (CEL). This liability risk margin (LRM) should take account of the uncertainty of models and parameters, and be such that the insurance contracts could be sold to a “willing buyer” or put in runoff. In economic terms the risk margin is often called a *market value margin* (MVM).

In Australia, the risk margin for nonlife insurance is calculated as the 75th percentile of the distribution function where the unbiased mean equals CEL. Using an economic approach, a proxy of the LRM (or MVM) can be given by a cost-of-capital (CoC) approach. “The CoC approach bases the risk margin on the theoretical cost to a third party to supply capital to the company in order to protect against risks to which it could be exposed” [22, 23]. The CoC approach was first introduced in the solvency context in the Swiss solvency test, see [6, 24].

The MVL and the capital requirement, in terms of SCR, have somewhat different roles in a solvency regime. The LRM is a safeguard to the CEL.

The SCR provides further safeguard to the policyholders by protecting both the MVL and the MVA and their interaction. The SCR should be calibrated such that it could withstand current year claims’ experience in excess of CE and that assets still exceed the MVL at the end of a defined time horizon, with a certain degree of confidence, say 99.5%, or that the available capital (ASM) could withstand a range of predefined shocks or stress scenarios over the defined time horizon.

Let $X = MVL$ and Y be two stochastic variables, where Y is the assets covering the MVL, see Figure 2. The SCR is now defined as a function of $Z = h(X, Y)$. We can assume an unknown and probably skewed distribution function of Z with $\mu_Z = MVL$. The SCR is now defined as a stochastic variable: $SCR = f(Z) - \mu_Z$, where $f(Z)$ is usually measured by the *value at risk* (VaR) (see **Value at Risk (VaR) and Risk Measures**) or *TailVaR*^d risk metrics. The capital requirement (SCR) is thus the difference between the upper quantile (if using VaR) in the skew distribution and the mean (“assets covering the MVL”), see Figure 2. The function $h(X, Y)$ includes both the risks inherent in the liabilities (X) and the assets covering them (Y) as well as the interaction between the assets and liabilities.

To model the SCR, the IAA [5] has proposed five main *risk categories*, mainly based on the Basel II accord and insurance characteristics: (a) insurance

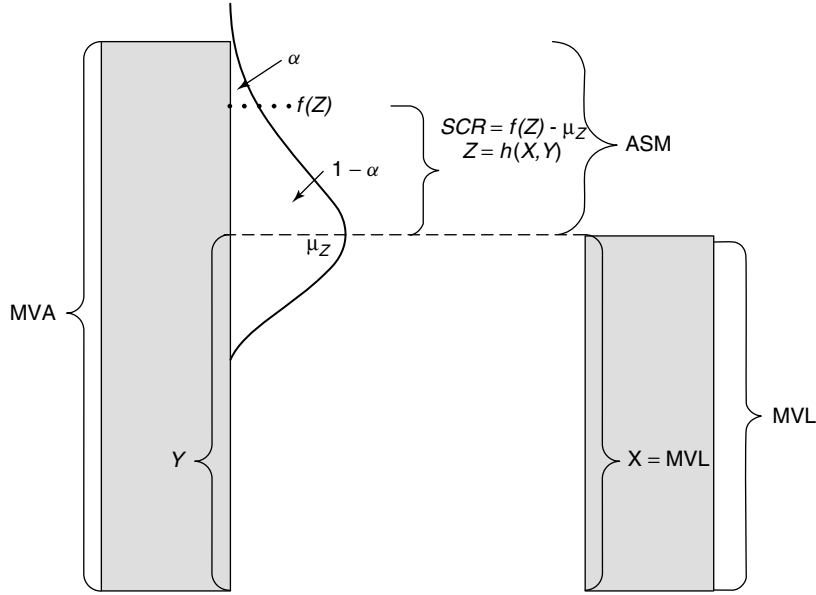


Figure 2 The SCR is calculated as the difference between a function of $h(X, Y)$ and the mean of the distribution. The distribution is a function of the $X = MVL$ and the corresponding assets covering the liabilities (Y)

risk (or underwriting risk), (b) credit risk, (c) market risk, (d) operational risk, and (e) liquidity risk. If a risk category is not possible to model, it should be treated as a Pillar II assessment, i.e., under the supervisory review process. The five risks are used both for solo entities and for insurance groups or financial conglomerates. There is also a sixth main risk category that can be introduced for groups or conglomerates: the group or participating risk. Examples of the latter type are an internal reinsurance program within an insurance group or the possibility that a bank is insuring its credit risks to an insurance company in the conglomerate.

The modeling is made in a top-down fashion. Each of the main risk categories is in the next step of modeling split up into subrisks, which in turn could be split up into sub-sub risks, etc. On the contrary, the calculation of the capital requirement is made in a bottom-up approach, see [29], starting from the lowest level.

To model the SCR we use a pragmatic solution. We use the notation of the standard deviation principle as the base-line approach [5] and apply the VaR and the TailVaR as risk metrics. The VaR approach is illustrated in Figure 2. We have two main problems to deal with: the nonnormality and the nonlinearity.

With four main risk categories the SCR could be written as

$$C = \sqrt{\begin{aligned} &C_1^2 + C_2^2 + C_3^2 + C_4^2 + 2\rho_{12}C_1C_2 \\ &+ 2\rho_{13}C_1C_3 + 2\rho_{14}C_1C_4 + 2\rho_{23}C_2C_3 \\ &+ 2\rho_{24}C_2C_4 + 2\rho_{34}C_3C_4 \end{aligned}} \quad (3)$$

where $C_i = k\sigma_i$, and k is a quantile function that is clearly defined for a standard normal distribution.

Nonnormality

Using a first-order “normal power (NP) approximation” provides us a solution to the nonnormality problem. VaR (quantile) and TailVaR (expected shortfall)^c of a skew standard distribution $F(\cdot)$ can be expressed in terms of the VaR (quantile) and TailVaR of a standard normal distribution $\Phi(\cdot)$. For the VaR case see [6, 30, 31] and for the TailVaR case see [32, 33].

To be more pragmatic, even if we have started with $C_* = k(\gamma_1)\sigma_*$, we may choose to let C_* be the result of scenarios or stress tests (i.e., the capital charge based on stress tests).

Nonlinearity

In assessing diversification effects, IAA [5] has proposed the use of *Copulas* (see **Copulas and Other Measures of Dependency**) as they can recognize

dependencies in the tail of the distributions. Extreme events (*see* **Extreme Value Theory in Finance**) and tail dependencies are important for the insurance industry. In reference [29] it is proposed, as a pragmatic solution, to adjust the correlation matrix with tail correlations instead of using the more complex copulas approach.

Accounting

In order not to burden the insurance undertaking with both a statutory requirement and an accounting requirement, it would be desirable to have one reporting system that could be used by both purposes. The International Accounting Standard Board (IASB) has been working toward a consistent approach to insurance accounting. The work done by IASB has been in coordination with the United States Financial Accounting Standard Board (FASB).^f A presentation of IASB's work is given in reference [34].

End Notes

- ^a. If we only have one level of requirement, then $SCR = MCR$.
- ^b. As an example: Within EU the solvency requirement for life insurance is mainly 4% of the technical provisions. However, as these are calculated with prudent assumptions, e.g., in mortality, the more prudent a company is, the higher its requirement is!
- ^c. See further reading.
- ^d. The TailVaR is a coherent risk metric.
- ^e. We assume that the distributions are continuous.
- ^f. For more information see its website: www.fasb.org.

References

- [1] Pentikäinen, T. (1952). On the net retention and solvency of insurance companies, *Skandinavisk Aktuarietidskrift* **35**, 71–92.
- [2] Campagne, C. (1961). *Standard Minimum De Solvabilité Applicable Aux Entreprises D'assurances*, Report of the OECE, March 11. Reprinted in *Het Verzekerings-Archief* deel XLVIII, 1971–1974.
- [3] Kastelijn, W.M. & Remmerswaal, J.C.M. (1986). *Solvency, Surveys of Actuarial Studies*, 3, Nationale-Nederlanden N.V., Rotterdam.
- [4] Pentikäinen, T. (2004). *Solvency, Encyclopedia of Actuarial Science*, J. Teugels & B. Sundt, eds, John Wiley & Sons, New York.
- [5] IAA (2004). *A Global Framework for Insurer Solvency Assessment*, Ontario.
- [6] Sandström, A. (2005). *Solvency: Models, Assessment and Regulation*, Chapman & Hall, Boca Raton, ISBN: 1-58488-554-8.
- [7] Campagne, C., van der Loo, A.J. & Yntema, L. (1948). Contribution to the Method of Calculating the Stabilization Reserve In Life Assurance Business. *Gedenboek Verzekeringskamer 1923-1948*, Staatsdrukkerij- En Uitgeverijbedrijf, Den Haag, pp. 338–378 (in both Dutch and English).
- [8] Müller report (1997). Report of the Working group “Solvency of Insurance Undertakings” set up by the Conference of the European Union Member States, DT/D/209/97. Available at <http://www.ceiops.org>.
- [9] EU Commission (2006). *Amended Framework for Consultation on Solvency II*, MARKT/2515/06, European Commission, Internal Market and Services DG, Financial Institutions, Insurance and Pensions.
- [10] Daykin, C.D. (1984). The development of concepts of adequacy and solvency in non-life insurance in the EEC, 22nd *International Congress of Actuaries*, Sydney, pp. 299–309.
- [11] Daykin, C.D., Devitt, E.R., Kahn, M.R. & McCaughan, J.P. (1984). The solvency of general insurance companies, *Journal of the Institute of Actuaries* **111**, 279–336.
- [12] Daykin, C.D., Bernstein, G.D., Coutts, S.M., Devitt, E.R.F., Hey, G.B., Reynolds, D.I.W. & Smith, P.D. (1987). Assessing the solvency and financial strength of a general insurance company, *Journal of the Institute of Actuaries* **114**, 227–310.
- [13] Daykin, C.D. & Hey, G.B. (1990). Managing uncertainty in a general insurance company, *Journal of the Institute of Actuaries* **117**, 173–277.
- [14] De Wit, G.W. & Kastelijn, W.M. (1980). The solvency margin in non-life insurance companies, *ASTIN Bulletin* **11**, 136–144.
- [15] Pentikäinen, T. (ed) (1982). *Solvency of Insurers and Equalization Reserves. Volume I General Aspects*, Insurance Publishing Company Limited, Helsinki, ISBN 951-9174-20-6.
- [16] Rantala, J. (ed) (1982). *Solvency of Insurers and Equalization Reserves. Volume II Risk Theoretical Model*, Insurance Publishing Company Limited, Helsinki, ISBN 951-9174-20-6.
- [17] Pentikäinen, T., Bonsdorff, H., Pesonen, M., Rantala, J. & Ruohonen, M. (1989). *Insurance Solvency and Financial Strength*, Finnish Insurance Training and Publishing Company Limited, Helsinki, ISBN 951-9174-75-3.
- [18] Norberg, R. & Sundt, B. (1985). Draft of a system for solvency control in non-life insurance, *ASTIN Bulletin* **15**(2), 149–169.
- [19] Norberg, R. (1986). A contribution to modelling of IBNR claims, *Scandinavian Actuarial Journal* (3–4), 155–203.
- [20] Norberg, R. (1993). A solvency study in life insurance, *III AFIR Colloquium*, Rome, 30 March – 3 April 1993.

- [21] Cont, R. (2006). Model uncertainty and its impact on the pricing of derivative instruments, *Mathematical Finance* **16**(3), 519–547.
- [22] Comité Européen des Assurances (2006). *CEA Document on Cost of Capital*, Brussels, April 21, 2006.
- [23] CEA-CRO Forum (2006). *Solutions to Major Issues for Solvency II*, February 17, 2006.
- [24] SST (2004). *The Risk Margin for the Swiss Solvency Test*, Swiss Federal Office for Private Insurance, September 17, 2004.
- [25] Bühlmann, H. (2002). New math for life actuaries, *ASTIN Bulletin* **32**(2), 209–211.
- [26] Bühlmann, H. (2003). On teaching actuarial science, guest editorial, *British Actuarial Journal* **9**(3), 491–492.
- [27] Bühlmann, H. (2004). Multidimensional valuation, *Finance* **25**, 15–29.
- [28] Wüthrich, M., Bühlmann, H. & Furrer, H. (2007). *Market-Consistent Actuarial Valuation*, Springer verlag, ISBN: 978-3-540-73642-4.
- [29] Groupe Consultatif (2005). *Diversification*, Technical Paper, Version 3.1, 17 October 2005 (written by Henk van Broekhoven).
- [30] Beard, R.E., Pentikäinen, T. & Pesonen, E. (1984). *Risk Theory, The Stochastic Basis of Insurance, Monographs on Statistics and Applied Probability* 20, 3rd Edition, Chapman & Hall, ISBN: 0-412-24260-5.
- [31] Daykin, C.D., Pentikäinen, T. & Pesonen, M. (1994). *Practical Risk Theory for Actuaries, Monographs on Statistics and Applied Probability* 53, Chapman & Hall, ISBN: 0-412-42850-4.
- [32] Christoffersen, P. & Gonçalves, S. (2005). Estimation risk in financial risk management, *Journal of Risk* **7**, 1–28.
- [33] Giamouridis, D. (2006). Estimation risk in financial risk management: a correction, *Journal of Risk* **8**, 121–125.
- [34] Wright, P. (2006). A view of the IASB's work on accounting for insurance contracts and financial

instruments, *Report Presented at the 28th International Congress of Actuaries*, Paris.

Further Reading

As the European Solvency II project and the IAIS solvency project are evolving, we recommend the web sites of these and other organizations for further reading:

EU Solvency II project: http://ec.europa.eu/internal_market/insurance

International Actuarial Association, IAA: www.actuaries.org

International Association of International Supervisors, IAIS: www.iaisweb.org

International Accounting Standards Board, IASB: www.iasb.org

Groupe Consultatif Actuariel Européen, GC: www.gcactuaries.org

Committee of European Insurance and Occupational Pensions Supervisors, CEIOPS: www.ceiops.org

Related Articles

Asset–Liability Management for Life Insurers

Asset–Liability Management for Nonlife Insurers

ARNE SANDSTRÖM

Spatial Risk Assessment

Spatial risk assessment concerns the analysis of a risk outcome within a geographically bounded region (such as a country, county, or census tract). The risk could be defined to be health risk or risk of adverse events, such as accidental occurrences or toxic spillages (see **What are Hazardous Materials?**). In general, *risk* can be broadly defined to mean the chance of a particular outcome (the probability of loss or threat). Loss or threat could be to health of people or ecological systems, either by a health insult or, for example, criminal activity. Alternatively, loss could be financial in terms of fraud or monetary loss. In engineering applications, loss or threat may be related to failure of components. Many of the techniques that will be mentioned here have general applicability across these different risk scenarios. However, for simplicity, the focus is on health risk and situations where an adverse health outcome is to be detected in space. A general reference in the area of risk assessment is [1].

Detection of Elevated Risk Areas

Georeferenced (or *spatial*) data can be analyzed for evidence of adverse outcomes. Maps of disease incidence for example (i.e., the spatial distribution of case addresses of disease) can display peaks of risk and these can be analyzed with respect to “significant” excesses of risk. The assessment of the degree of significance is a statistical issue. A range of methods are available for the assessment of such localized areas of risk. Often these areas are called *clusters*, and *cluster detection* is a major activity within public health, crime, and financial surveillance. Clusters can suggest localized environmental risk elevation or, in the case of crime statistics, some elevated crime risk. Spatial clustering can be detected *via testing* or *via modeling*. There has been large literature in the area of cluster testing, and software is readily available as special packages or functions (e.g., SatScan <http://www.srab.cancer.gov/satscan/>, ClusterSeer <http://www.terraser.com/products/clusterseer.html> or DCluster in R <http://cran.r-project.org/>). Modeling of risk is less well developed but has seen some recent advances using Bayesian

hierarchical modeling (see e.g. Lawson and Denison [2]) (see **Bayesian Statistics in Quantitative Risk Assessment**).

Risk Detection

The basis of the detection of areas of elevated risk is a criterion for labeling areas as “in excess”. Definition of what constitutes an area of excess and also what constitutes significance of the area is important. One definition of excess leads to what is known as “*hot spot*” clustering. This is where any area, regardless of shape or size, is declared to be a cluster if the risk estimate in the area is “significantly” high. Often this is a natural definition in that we are often interested in finding any unusual elevated risk. However, sometimes the risk excess can have a particular form and it is only when this form is found that a significance can be assessed. For example, it might be that the shape of the cluster or spatial integrity is important. A cluster of relatively regular conical or circular shape might suggest a central source for the risk. In the case of putative source risk assessment (see e.g. Elliott *et al.* [3, Chapter 9]), the aim is to detect any evidence for an association between the adverse health outcomes and a fixed known location (believed to be the source of risk). This putative source could be the location of an incinerator or nuclear plant or waste dump site. An early example was John Snow’s analysis of cholera deaths in relation to the Broad Street pump in London [4]. More recent examples include the Armadale (UK) lung cancer epidemic, and the Sellafield nuclear reprocessing plant (UK) and the risk of childhood leukemia (see e.g., Lloyd [5]; Kinlen [6]) (see **Radioactive Chemicals/Radioactive Compounds**). Around fixed locations, it is feasible to set up statistical models for the analysis where the outcome variables are (disease) incident counts (or at a finer level, event locations themselves) and the explanatory variables are exposure surrogates such as distance and direction around the fixed location. In this case, the focus is usually on evidence for a decline with distance from the source or concentration around a directional axis. This often implies that some form of circular clustering is assumed. Such models are simple to fit using standard software such as R or SAS.

When the location of clusters or areas of elevated risk are unknown, the problem is more difficult to

analyze statistically. If one adopts a method that can assess whether the localized risk is elevated, then that method can be used for *hot spot* detection with small areas. If on the other hand the clusters are thought to be larger than just one area, methods that use neighborhoods might need to be employed. Neighborhood methods assume that contiguity is important in clustering.

Definition of Clusters and Clustering

Two extreme forms of clustering can be defined. These two extremes represent the spectrum of modeling from nonparametric to parametric forms and there are appropriate statistical models and estimation procedures associated with these forms. First, as many researchers may not wish to specify *a priori* the exact form/extent of clusters to be studied, a nonparametric definition is often the basis adopted. Without any assumptions about shape or form of the cluster, the most basic definition would be as follows: *any area within the study region of significantly elevated risk* [7, p. 112].

This definition is often referred to as *hot spot clustering*. In essence, any area of elevated risk, regardless of shape or extent, could qualify as a cluster, provided the area meets some statistical criteria. Note that it is not usual to regard areas of significantly low risk to be of interest, although these may have some importance in further studies of the etiology of a particular disease.

Second, at the other extreme, we can define a parametric cluster form: *the study region displays a prespecified cluster structure*.

This definition describes a parameterized cluster form that would be thought to apply across the study region. Usually, this implies some stronger restriction on the cluster form and also on some region-wide parameters that control the shape and size of clusters. For example, infectious diseases may display particular forms of clustering depending on the spread rate and direction.

Nonspecific Clustering

Nonspecific clustering is the analysis of the overall clustering tendency in a study region. This is also known as *general clustering*. As such, the assessment of general clustering is closely akin to the assessment

of spatial autocorrelation. Hence, any model or test relating to general clustering will assess some overall/global aspect of the clustering tendency of the disease of interest. This could be summarized by a model parameter (e.g., an autocorrelation parameter in an appropriate model) or by a test that assesses the aggregation of cases of disease. For example, a model often assumed for disease relative risk estimation (the Besag, York, and Mollié (BYM) model) uses a correlated prior distribution that has a parameter that describes the clustering behavior. It should be noted at this point that the general clustering methods discussed above can be regarded as *nonspecific* in that they do not seek to estimate the spatial locations of clusters but simply to assess whether clustering is apparent in the study region. Any method that seeks to assess the locational structure of clusters is defined to be *specific*.

An alternative nonspecific effect has also been proposed in models for tract-count or case-event data. This effect is conventionally known as *uncorrelated heterogeneity* (or overdispersion/extra-Poisson variation in the Poisson likelihood case).

Specific Clustering

Specific clustering involves the analysis of the locations of clusters. Location could be very important in the assessment of risk localization. This approach seeks to estimate the location, size, and shape of clusters within a study region. For example, it is straightforward to formulate a nonspecific Bayesian model for case events or tract counts that includes heterogeneity. However, specific models or testing procedures are less often reported. Nevertheless, it is possible to formulate specific clustering models for the case-event and tract-count situation.

Focused versus Nonfocused Clustering

Another definition of clustering seeks to classify the methods based on whether the location or locations of clusters are known or not. Focused clustering is *specific* and usually seeks to analyze the clustering tendency of a disease or diseases around a known location. Often this location could be a *putative pollution source* or health hazard (as discussed above). Nonfocused clustering does not assume knowledge of a location of a cluster but seeks to assess either

the locations of clustering within a map or the overall clustering tendency. Hence, nonfocused clustering could be specific or nonspecific.

Hypothesis Tests for Clustering and Cluster Modeling

The literature of spatial risk assessment of health data has developed considerably in the area of hypothesis testing and, more specifically, in the sphere of hypothesis testing for clusters. Very early developments in this area arose from the application of statistical tests to spatiotemporal clustering. Spatiotemporal clusters are short duration bursts of excess risk, which are spatially compact, and are a particularly strong indicator of the importance of a spatial clustering phenomenon. As noted above, distinction should be made between tests for general (nonspecific) clustering, which assess the overall clustering pattern of the disease, and the *specific* clustering tests where cluster locations are estimated.

For case events (where the residential address location of cases is known), a few tests have been developed for nonspecific clustering. Specific nonfocused cluster tests address the issue of the location of putative clusters. These tests produce results in the form of locational probabilities or significances associated with specific groups of tract counts or cases. Openshaw and coworkers [8] first developed a general method that allowed the assessment of the location of clusters of cases within large disease maps. The method was based on repeated testing of counts within circular regions of different sizes. An alternative statistic has been proposed by Kulldorff and Nagarwalla [9] (the scan statistic). The test can be applied to both case events and tract counts.

Focused tests have also been developed and there is now a range of possible testing procedures (see e.g. Lawson [7, Chapter 7]).

Cluster modeling has seen some development but has not developed as fully as testing procedures. Usually the successful models are Bayesian with prior distributions describing the risk aggregation behavior (see e.g. Lawson and Denison [2]).

References

- [1] Covello, V. & Merkhofer, M. (1993). *Risk Assessment Methods: Approaches for Assessing Health and Environmental Risks*, Plenum Publishing, New York.
- [2] Lawson, A.B. & Denison, D. (2002). *Spatial Cluster Modelling*, 1st Edition, Chapman & Hall, London.
- [3] Elliott, P., Wakefield, J., Best, N. & Briggs, D. (2000). *Spatial Epidemiology: Methods and Applications*, Oxford University Press, London.
- [4] Snow, J. (1854). *On the Mode of Communication of Cholera*, 2nd Edition, Churchill Livingstone, London.
- [5] Lloyd, O. (1982). Mortality in a small industrial town, in *Current Approaches to Occupational Health* – 2, A. Gardner, ed, Wright.
- [6] Kinlen, L.J. (1995). Epidemiological evidence for an infective basis in childhood leukaemia, *British Journal of Cancer* **71**, 1–5.
- [7] Lawson, A.B. (2006). *Statistical Methods in Spatial Epidemiology*, 2nd Edition, John Wiley & Sons, New York.
- [8] Openshaw, S., Charlton, M.G., Wymer, C. & Craft, A.W. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets, *International Journal on Geographical Information Systems* **1**, 335–335.
- [9] Kulldorff, M. & Nagarwalla, N. (1995). Spatial disease clusters: detection and inference, *Statistics in Medicine* **14**, 799–810.

Related Articles

Environmental Hazard

Environmental Health Risk

Statistics for Environmental Toxicity

ANDREW B. LAWSON

Spatiotemporal Risk Analysis

Spatiotemporal risk analysis focuses on the study of the geographical pattern of risk and its variation over time. The *spatial analysis* of risk usually consists of a static examination of a surface to assess where unusual risk patterns occur. Inevitably, with almost all risk subjects, whether health- or crime- or finance-related, patterns of risk vary with time and often it is this time variation (or *temporal* variation) that is of interest. Unusual risk patterns could take a variety of forms when expressed as features on a spatial surface. Often clusters of risk are the focus, and these could be circular, oblong, linear, or have a relatively indeterminate pattern. In addition, in *spatiotemporal analysis* of risk, we are interested in temporal effects and so we might want to assess how *change* occurs in the spatial risk surface. The detection of changes in risk and, in this case, risk surfaces is a major focus area for *surveillance*. Disease surveillance, for example, when applied to spatial data, focuses on the ability to detect changes in disease risk over space and time (see Farrington and Andrews [1] and Lawson [2]). There are a range of approaches available to this end, depending on what type of unusual risk pattern or aberration is focused on.

Aberration types

1. Temporal

In temporal surveillance, a time series of events is available (usually either as counts of disease in intervals or as a point process of reporting or notification times). With time series of discrete counts, the aberrations that might be of interest (suggesting “important” disease changes) are highlighted in Figure 1. In this figure, there are three main changes (A, B, C). The first aberration (A) is a sharp rise in risk (jump) or a change point in level. The second aberration (B) that might be of interest is a cluster of risk (an increase and then a decrease in risk). However, this can only be detected retrospectively. The third aberration (C) is an overall process change where not only the level of the process is changed but also the variability is increased or decreased.

When a point process of event times is observed, then action may need to be taken whenever a new event arrives. Aberrations that are found in point processes can take the form of unusual aggregations of points (temporal clusters), sharp changes in rate of occurrence (jumps), and overall process change (as above).

2. Spatiotemporal

If a spatial domain (map of events) is to be monitored, then spatial and spatiotemporal aberrations must be considered. Spatial aberrations could consist of discontinuities in risk between regions (jumps in risk across region boundaries), spatial clusters of risk (localized aggregations of cases of disease), and long-range trend in risk.

When maps are monitored in time, then *changes* in the above features might be of interest. In the infectious disease case, for example, the development of new clusters of infectious disease in time might signal localized outbreaks.

Multivariate Issues

In surveillance systems designed for prospective surveillance, whether for finance or health or other foci, many of the above features would have to be detectable across a wide range of data streams. Not only would whole data streams be monitored, but subsets of data streams may be required to be examined. For example, for early detection of effects, it may be important to monitor “high-risk” subsets (such as old or young age-groups for health data). Also correlations between subsets may be important

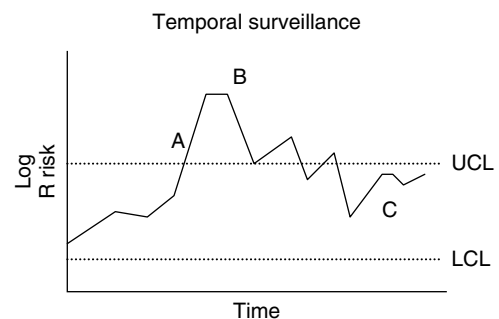


Figure 1 Schematic features of changes/aberrations found in temporal surveillance (Log R risk: log relative risk)

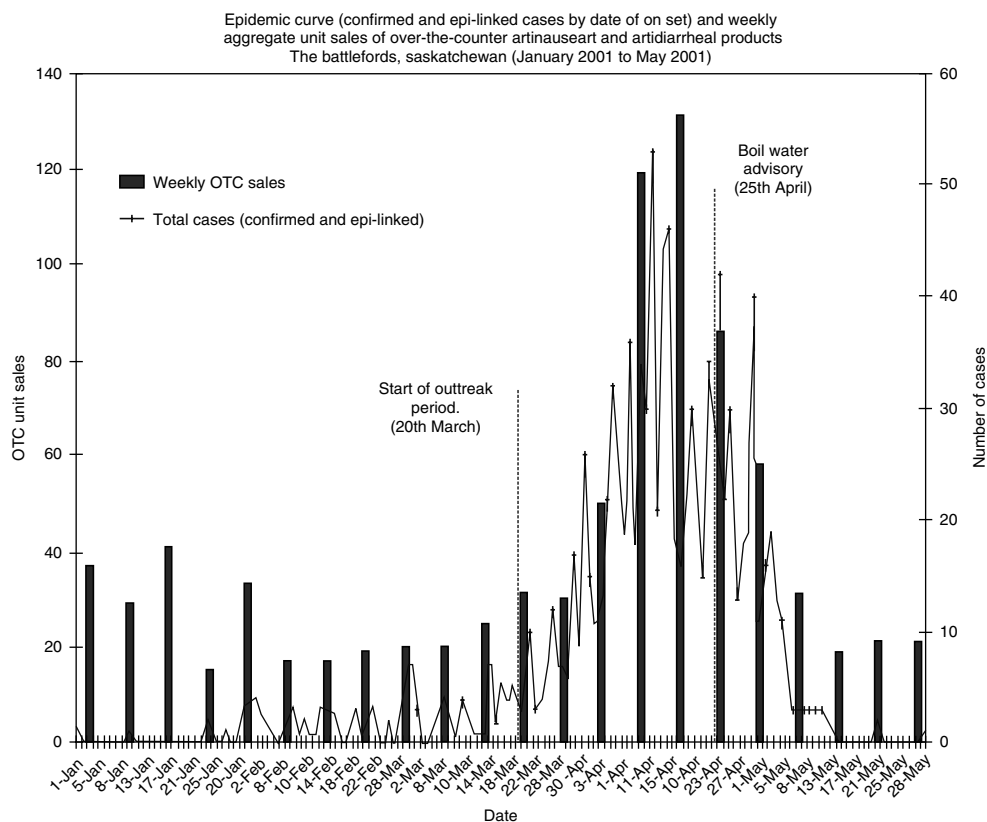


Figure 2 Battlefords, Saskatchewan: epidemic curve and time series of over-the-counter (OTC) pharmaceutical sales from January to May 2001 [Reproduced from Edge *et al.*, *Canadian Journal of Public Health* **95**, 2004. © Canadian Public Health Association.]

for making detection decisions. In addition, if maps are to be monitored over time as well as time series, then the problem intensifies considerably. The *Biosense* system developed by centers for disease control (CDC) (<http://www.cdc.gov/phn/component-initiatives/biosense/index.html>) has a multivariate and mixed stream (spatial and temporal) capacity.

Data Collection

Data for space–time risk analysis are often difficult to gather. For health systems, this would mean obtaining historical records to allow estimation of “normal” background risk evolution and also a spatial resolution fine enough to allow monitoring of risk spatially. For health data, very fine spatial resolution is often not available owing to confidentiality. For

crime or financial analysis, this fine scale may not be useful. In addition, early detection of events (such as bioterrorism infections or credit card fraud events) would be useful. Hence ancillary variables may need to be collected and monitored for this purpose. In public health analysis this is known as *syndromic surveillance* [3]. Figure 2 provides an example of some syndromic data where over-the-counter pharmaceutical sales are used to make early detection of gastrointestinal disease outbreaks.

Models

One of the current concerns about large-scale surveillance systems is the ability to correctly model the variation in data and to properly calibrate sensitivity and specificity with such a

multivariate and multifocus task. The application of Bayesian hierarchical modeling has been advocated for surveillance purposes and this may prove to be useful in its ability to deal with large-scale systems evolving in time in a natural way. For example, recursive Bayesian learning could be important. Clearly, computational issues could be paramount here given the potentially multivariate nature of the problem and the need to optimize computation time. Sequential Monte Carlo methods have been proposed to deal with computational speedups [4–6].

References

- [1] Farrington, P. & Andrews, N. (2004). Outbreak detection: application to infectious disease surveillance, in *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*, R. Brookmeyer & D. Stroup, eds, Oxford University Press, New York.
- [2] Lawson, A.B. (2004). Some issues in the spatio-temporal analysis of public health surveillance data, in *Statistical Principles and Methods for Public Health Surveillance*, R. Brookmeyer & D. Stroup, eds, Oxford University Press, New York.
- [3] Lawson, A.B. & Kleinman, K. (2005). *Spatial and Syndromic Surveillance for Public Health*, John Wiley & Sons, New York.
- [4] Kong, A., Lai, J. & Wong, W. (1994). Sequential imputations and Bayesian missing data problems, *Journal of the American Statistical Association* **89**, 278–288.
- [5] Doucet, A., De Freitas, N. & Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*, Springer, New York.
- [6] Doucet, A., Godsill, S. & Andrieu, C. (2005). On sequential Monte Carlo sampling methods for Bayesian filtering, *Statistics and Computing* **10**, 197–208.

ANDREW B. LAWSON

Stakeholder Participation in Risk Management Decision Making

Stakeholder participation in risk-management decision making has had a checkered history. The nature of this has been perhaps best captured by MacKean and Thurston [1] who likened stakeholder participation to “eating spinach”. While people know that spinach is good for them in theory, they do not necessarily want to eat it themselves or even see it served at their dinner table. As a result of this attitude, there continues to be a gap between the concept and reality of stakeholder participation [2].

Nonetheless, the value of stakeholder participation in risk decision making is becoming increasingly recognized. This paper explores the benefits and challenges of stakeholder involvement, some of the factors contributing to successful stakeholder participation in the risk-management process, and the evaluation of stakeholder participation processes.

For the purposes of this discussion, “stakeholder” is being defined in terms of the broader concept of “interested and affected parties” coined by the US National Research Council [3]. “Affected parties” are “people, groups, or organizations that may experience benefit or harm as a result of a hazard, or of the process leading to risk characterization, or of a decision about risk. They need not be aware of the possible harm to be considered affected” [3, p. 214]. “Interested parties” are “people, groups, or organizations that decide to become informed about and involved in a risk characterization or decision-making process. Interested parties also may or may not be affected parties” [3, p. 215]. “Interested and affected parties” include risk practitioners, risk researchers, decision and policy makers, and the various “public”. They may be representatives of government, academia, industry, nongovernmental organizations, advocacy groups, and the general public, among others. However, the primary focus of many efforts in stakeholder participation has been on the specific needs of the “public”, and how, when, and why they should be involved in risk decision-making processes (*see Scientific Uncertainty in Social Debates Around Risk*).

Benefits and Challenges

Over the past 30 years of formalized risk decision making, the value of stakeholder participation has become increasingly evident. Involving stakeholders produces decisions that are responsive to varying interests and values, including those of the community. Stakeholder participation has also helped to prevent or resolve conflicts, build trust between various parties, and inform all parties about different aspects of the risk [4]. A review of 239 published case studies of stakeholder involvement in environmental decision making conducted by Beierle [5] showed that intensive stakeholder processes were more likely to result in higher quality decisions.

However, many risk professionals and decision makers are not yet “sold” on the value and feasibility of stakeholder participation. Some of the concerns frequently raised by agencies dealing with risk are that stakeholder participation will encourage the mobilization of antagonistic interests, will be too costly (in terms of both time and money), will be taken over by special interests, will undermine their authority, will be lacking in technical knowledge, and will be over dramatized [6]. There is also concern that trying to seek a compromise between a “squeaky wheel” and a more “moderate” stakeholder using traditional public participation may increase the likelihood of more extreme decisions [7].

Still others contend that stakeholder participation (particularly public participation) and scientific rigor are mutually exclusive processes [8]. Yet it must be remembered that different perspectives cannot change scientific facts or the laws of nature. What they can do is change how this information is interpreted and used in decision-making processes, thereby validating decisions about these risks [2]. Yosie and Herbst [9] suggested that stakeholder participation might actually function as a “social peer review” process that complements the scientific peer review process. In studies on risk assessment approaches for genetically modified foods, it was demonstrated that “more inclusive styles of decision making are not only more democratic but also more scientifically robust” [10, p. 129], on the basis that a strictly “scientific” approach will inevitably lead to an overly narrow focus of enquiry and policy judgment. Juanillo [11] also concluded that truly participatory communication on the risks and benefits of agricultural biotechnology leads to better regulatory policies.

However, many risk-management processes continue to be dominated by those conducting technical risk assessments, and provide only limited opportunities for interested parties to participate fairly and competently in the decision-making process [12, 13]. As was noted by the US National Research Council [3, p. 87], behavior based on the assumption that risk analysis is a matter for experts only, “may lead some of the interested and affected parties to feel disenfranchised from the regulatory process and either withdraw from the policy arena or seek unconventional ways to interfere with the process.”

Overall, the benefits associated with a participatory risk decision-making process are irrefutable. Most agencies dealing with the management of risks have acknowledged (through voluntary and/or regulated means) their responsibility to enable stakeholders to actively participate in the process. Likewise, stakeholders are increasingly aware of their right to be part of an informed process.

Successful Stakeholder Participation

Much has been written on “lessons learned” about past stakeholder (particularly public) participation efforts [8, 13–16]. While it is generally agreed that more research is required to evaluate all the elements that may contribute to a successful participatory process, it has been amply demonstrated that a variety of approaches can be employed [17]. It has also been clearly shown that a “one size fits all” approach for participation processes is neither feasible nor desirable. Each situation must be assessed in terms of the goals of participation, the design of the process and the context of the specific risk situation [18].

However, some general guidance may be found in the “Guidelines to Stakeholder Involvement” developed by the US Presidential/Congressional Commission on Risk Assessment and Risk Management [19]. Additional guidance can be found in Jardine *et al.* [20]. These considerations (modified from both sources) are summarized below:

- Stakeholder participation must be meaningful. Agencies should only involve stakeholders in the process if they are willing and able to respond to stakeholder input in making the final risk decisions. Involving stakeholders when the decision has already been made is unethical and only serves to increase frustration and distrust.
- Attempts should be made to engage all interested and affected parties so that a variety of perspectives are considered. If appropriate, stakeholders should be provided with incentives to ensure an “equal playing field” – for example, if agency personnel are being compensated to participate in the process as part of their jobs, it may be only fair for other stakeholders to receive some recognition of their time and expertise.
- Involvement of all parties should occur at the beginning of the process when the problem is defined. Failure to involve people at this stage in the process frequently results in either “solving the wrong problem” or in overlooking some key aspect of concern.
- The purpose of the process, the roles of those involved, and the design of the process should be clearly explained and agreed upon by everyone involved in the process.
- All interested and affected parties should have equal access to relevant information and support a mutual learning environment. Where possible, stakeholders should be empowered to make decisions by providing them with whatever resources are necessary to help them fully gather and understand all the relevant information.
- All parties should be willing to participate and work together. They must be prepared to be flexible, to negotiate, to listen, and to learn from different viewpoints.
- All those involved in the process should be given credit for their roles in the final decision. Too often, the coordinating agency takes full credit for the work of many stakeholders.
- Stakeholder involvement should be made part of the agency’s mandate and mission. Sufficient resources (including time) must be allocated to ensure that the process is initiated early and sustained through to the final risk-management decision.
- The “nature, extent, and complexity of stakeholder involvement should be appropriate to the scope and impact” [19, p. 16] of the risk problem and decision. For decisions that affect many people and/or are highly contentious, a formalized process involving all potentially interested and affected parties is warranted. For decisions with minimal impact and/or potential controversy, a more informal process of those directly involved might be more appropriate.

Table 1 Stakeholder participation goals and criteria for evaluating success in risk decision-making processes^(a)**Goal 1: Incorporating different values into decisions**

How much influence are the various stakeholders having on decisions made?

Whose values are represented?

Are the participants representative of the group they are representing?

Are all the interests at the table?

Are processes in place for soliciting input from interested and affected parties not at the table?

Goal 2: Improving the substantive quality of decisions

Are stakeholders improving decisions through creative problem solving, innovative ideas, or new information?

Are the decisions superior to likely alternatives in terms of cost-effectiveness, joint gains, the opinions of participants, or other measures?

Do participants add information, provide technical analysis, contribute innovative ideas, or contribute a holistic perspective?

Goal 3: Resolving conflict among competing interests

Was conflict that was present at the beginning of the process resolved by the end?

Was conflict avoided by avoiding contentious issues?

Was conflict avoided because certain parties were excluded or chose not to participate?

Goal 4: Building trust

Was mistrust of agencies or of interest groups that was present at the beginning of the process lessened by the end?

If trust was high at the beginning of the process, was this maintained or did it decrease?

Goal 5: Information exchange

Did all stakeholders learn enough about the issue to actively engage in decision making?

Was there educational outreach to the wider public?

Did the agencies involved learn and acknowledge both public concerns and values, and local knowledge of the environment and the community?

^(a) Reproduced from [4] with permission from Resources for the Future, © 2002

Evaluating Stakeholder Participation

As with risk communication, evaluation is a valuable but often overlooked component of stakeholder involvement processes (*see Evaluation of Risk Communication Efforts*). Without evaluation, there is no means of determining if the stakeholder participation process has been successful. Learning from successes and failures improves subsequent efforts. Conducting an evaluation is dependent on setting appropriate goals at the outset of the process. Beierle and Cayford [4] outlined five general goals for stakeholder participation. These are summarized in Table 1, together with criteria for evaluating if the participation process has been successful in meeting each goal.

Summary

Despite the acknowledged challenges in involving stakeholders in risk decision processes, the demonstrated benefits are incontrovertible. Furthermore, people have the right to be able to participate in

decision that affect them and the people they love; particularly when such decisions are based upon uncertain evidence. However, all parties would agree that one of the biggest obstacles in the employment of stakeholder participation and deliberation in risk decision-making processes is how this may be best undertaken. The “right” process of participation will need to be determined on the basis of the nature of the risk and the impact of the decision.

As a word of caution, stakeholder involvement should not be expected to necessarily prevent or end controversy. However, if properly employed, effective participation of interested and affected parties should result in productive discussion of disputes, and will hopefully engender the negotiation and compromise required to develop a decision acceptable to all.

References

- [1] MacKean, G. & Thurston, W. (1999). Chapter five: a Canadian model of public participation in health care planning and decision making, in *Efficiency vs Equality: Health Reform in Canada*, M. Stingl & D. Wilson, eds, Fernwood Books Limited, Halifax, pp. 55–69.

- [2] Jardine, C.G., Predy, G. & MacKenzie, A. (2007). Stakeholder participation in investigating the health impacts from coal-fired power generating stations in Alberta, Canada, *Journal of Risk Research* **10**(5), 693–714.
- [3] U.S. National Research Council (1996). *Understanding Risk: Informing Decisions in a Democratic Society*, National Research Council, National Academy Press, Washington, DC.
- [4] Beierle, T.C. & Cayford, J. (2002). *Democracy in Practice: Public Participation in Environmental Decisions*, Resources for the Future, Washington, DC.
- [5] Beierle T.C. (2002). The quality of stakeholder-based decisions, *Risk Analysis* **22**(4), 739–749.
- [6] Senecah S. (2004). The trinity of voice: the role of practical theory in planning and evaluating the effectiveness of environmental participatory processes, in *Communication and Public Participation in Environmental Decision Making*, S.P. Depoe, J.W. Delicath & M.-F.A. Elsenbeer, eds, State University of New York Press, New York, pp. 13–34.
- [7] Kinsella, W.J. (2004). Public expertise: a foundation for citizen participation in energy and environmental decisions, in *Communication and Public Participation in Environmental Decision Making*, S.P. Depoe, J.W. Delicath & M.-F.A. Elsenbeer, eds, State University of New York Press, New York, pp. 83–89.
- [8] McDaniels, T., Gregory, R.S. & Fields, D. (1999). Democratizing risk management: successful public involvement in local water management decisions, *Risk Analysis* **19**(3), 497–510.
- [9] Yosie, T.F. & Herbst, T.D. (1998). *Using Stakeholder Processes in Environmental Decision Making: An Evaluation of Lessons Learned, Key Issues, and Future Challenges*, Prepared by Ruder Finn Washington and ICT Incorporated.
- [10] Scott, A. (2001). Technological risk, scientific advice and public ‘education’: groping for an adequate language in the case of GM foods, *Environmental Education Research* **7**(2), 129–139.
- [11] Juanillo, N.K. Jr. (2001). The risks and benefits of agricultural biotechnology. *American Behavioral Scientist* **44**(8), 1246–1266.
- [12] Snary, C. (2002). Risk communication and the waste-to-energy incinerator environmental impact assessment process: a UK case study of public involvement, *Journal Environmental Planning and Management* **45**(2), 267–283.
- [13] Petts, J. (2004). Barriers to participation and deliberation in risk decision: evidence from waste management, *Journal of Risk Research* **7**(2), 115–133.
- [14] Renn, O., Webler, T. & Wiedemann, P. (eds) (1995). *Fairness and Competence in Public Participation: Evaluating Models for Environmental Discourse*, Kluwer Academic Press, Dordrecht.
- [15] U.S. Environmental Protection Agency (2001). *Stakeholder Involvement and Public Participation at the U.S. EPA*, United States Environmental Protection Agency, Office of Policy, Economics, and Innovation, EPA-100-R-00-040, January 2001.
- [16] Mock, G.A., Vanasselt, W.G. & Petkova, E.I. (2003). Rights and reality: monitoring the public’s right to participate, *International Journal of Occupational and Environmental Health* **9**, 4–13.
- [17] Kessler, B.L. (2004). *Stakeholder Participation: A Synthesis of Current Literature*, Prepared by the National Marine Protected Areas Center in Cooperation with the National Oceanic and Atmospheric Administration Coastal Services Center.
- [18] Rowe, G. & Frewer, L.J. (2004). Evaluating public participation exercises: a research agenda, *Science, Technology and Human Values* **29**(4), 512–556.
- [19] U.S. Presidential/Congressional Commission on Risk Assessment and Risk Management (1997). *Framework for Environmental Health Risk Management*, Final Report, Washington, DC, Vols 1 and 2.
- [20] Jardine, C.G. (2003). Development of a public participation and communication protocol for establishing fish consumption advisories, *Risk Analysis* **23**(3), 461–471.

Related Articles

Risk and the Media

Role of Risk Communication in a Comprehensive Risk Management Approach

CYNTHIA G. JARDINE

Statistical Arbitrage

Historical Context

In the 1960s, there were two basic approaches to stock-market analysis and investment strategies: fundamental analysis and technical analysis. *Fundamental analysis* (or value investing) was best exemplified by Graham and Dodd [1, 2]. Fundamental analysis focused solely on evaluation of the strength of the company as exhibited by its balance sheet, the strength of the industrial sector within which the company operated, and the economic outlook for the market as a whole and the company's sector in particular. Graham and Meredith [3] tried to educate the reader in the art and science of reading and interpreting financial statements. The result of this analysis was the identification of *value stocks* and *overvalued stocks*. The intelligent investor's portfolio would be long in the value stocks and short in the overvalued stocks (although Graham was not one to advocate short selling). Since the publication of Graham's books, value investing has been a standard approach to investing and is buttressed by an army of stock and industry analysts in all investment banks and brokerage houses; however, the method has periodically gone into and out of favor.

Technical analysis is, perhaps, best exemplified by Edwards *et al.* [4], whose work was first published in 1948 by Edwards and Magee, and is now in its eighth edition. Technical analysis is applied to the time series of stock prices and possibly also trading volume. The methodology assumes that stock prices are driven by supply and demand, and that certain stock prices, or at least trends, can be predicted from the history of that stock's time series of prices and trading volume. Interestingly, Edwards and Magee paid no attention whatsoever to the balance sheet, nor to any fundamental analysis of the company, industrial sector, or the economy. The buy and sell decisions were based solely on the stock-price pattern. More recent books on technical analysis have advocated combining it with fundamental analysis for stock selection.

The world of stock-price prediction entered a new phase in 1965 with the publication of Samuelson [5]. This was followed by the seminal paper of Fama [6], which coined the term *efficient market hypothesis* (EMH). This hypothesis is often defined

only informally. The interested reader might look at the website <http://www.e-m-h.org> for a brief history, various definitions, and relevant references. Fama [6, p. 383] gives the following description of the EMH:

... and investors can choose among the securities that represent ownership of firms' activities under the assumption that security prices at any time "fully reflect" all available information. A market where prices always "fully reflect" available information is called "efficient".

Fama [6, p. 383] goes on to conclude that "with but a few exceptions, the efficient markets model stands up well". This would mean that one can ignore the entire history of the stock-price time series (and that of all other stocks as well), paying attention only to current prices, and one will do as well as an individual who engages in fundamental or technical analysis.

The EMH is certainly counter intuitive and not easily believable; however, extensive empirical studies were conducted, and it was not rejected, at least for a substantial period of time. It was the dominant approach taken by the academic community for many years.

Malkiel's book [7] (now in its 8th edition) is one that was to become very popular in the mainstream press. Malkiel has been a constant, strong voice supporting the EMH (which he calls the random walk hypothesis). More recently, Malkiel [8] asserted:

The way I put it in my book, *A Random Walk Down Wall Street* first published in 1973, a blind-folded chimpanzee throwing darts at the *Wall Street Journal* could select a portfolio that would do as well as the experts. Of course, the advice was not literally to throw darts, but instead to throw a towel over the stock pages—that is, to buy a broad-based index fund that bought and held all the stocks in the market and that charged very low expenses.

The next 20 years might be called the *era of the efficient market Hypothesis*; however, there have been many attacks on it. Furthermore, the EMH did not seem to deter financial institutions from employing stock and industry analysts and selling information. Finally, many notable Wall Street figures rejected the EMH. For example, in an interview on April 3, 1995 issue of *Fortune* magazine, Warren Buffett said, "I'd be a bum in the street with a tin cup if the markets were efficient".

Beginning in 1985 and continuing well beyond 2000, there were a number of papers in the academic literature that called into question the EMH and assessed investment and trading strategies that could result in excess profit beyond market returns. It is, in fact, very easy to find many publications, especially in the popular press, which provide investment and trading strategies that appear to offer phenomenally good results. Many of those publications do not provide sufficient information about the details of the strategy, so that they are impossible to replicate. Others do not use sound statistical methodology for example, not realizing that strategies that are developed using in-sample data must be cross validated while using out-of-sample data. Finally, many papers do not make any assessment of the risks associated with the strategy they propose or investigate.

Arbitrage, Statistical Arbitrage, and Trading Strategies

Is “Riskless Arbitrage” Truly Riskless?

The concept of arbitrage is closely related to the “law of one price” which asserts that if two assets have the same payoff in every state, then they must have the same price. Going further, if two assets have the same payoff, but not the same price, then a riskless profit can be obtained by simultaneously trading both assets. Such a trade would be a *riskless arbitrage*. It is generally accepted that market forces will ensure that prices converge to their equilibrium levels, and the law of one price will be reestablished.

The reader should realize that there are few true arbitrage opportunities that exist. One situation occurs when the same instrument is priced differently in different markets. Shleifer and Vishny [9] give an example of two Bund futures contracts to deliver DM250 000 in face value of German bonds at a future time T , traded on two different markets. They suppose that the two identical contracts have different prices: one sells for DM240 000 and the second for DM245 000. An arbitrageur sells the higher priced contract and buys the lower priced one. Although he is perfectly hedged at time T , he must put up good faith money for each contract. If the prices of the contracts diverge further before time T , the arbitrageur must invest more money in the meantime. Depending on the extent to which such movements continue, he runs the risk of having to liquidate his

position at a loss prior to realizing the “sure profit” at time T . The authors go on to identify further complexities that bring forth other types of risks associated with even a simple “riskless” trade. Thus, the point they make is that even in the most obvious case of a pure arbitrage, opportunity risks may still be present.

Arbitrage-based pricing is exemplified by delta hedging and the Black–Scholes formula which leads to a unique price for the option, and if the price of the option were to differ, that difference presents an arbitrage opportunity. However, readers must be aware that this unique price is calculated under the assumption of a specific model, and if the model is wrong, then the price is wrong, thus the arbitrage opportunity is not free of risk, in this case *model risk* (see **Model Risk**).

Finally, default (see **Default Risk**) and credit risk (see **Mathematical Models of Credit Risk**) have entered the picture. With the failure of long-term capital management and with many other corporate defaults, it became apparent that the standard riskless arbitrage model needed to be extended to consider the possibility of *default risk*.

What Is Statistical Arbitrage?

Statistical arbitrage uses market models or statistical models to identify instruments that, according to the models, are priced inconsistently and whose prices are expected to return to what the models predict.

Hogan *et al.* [10, p. 526] say “We define statistical arbitrage as a long horizon trading opportunity that generates a riskless profit”. More precisely, they let $v(t)$ represent the discounted cumulative profits from a zero-initial-cost, self-financing trading strategy. The strategy is called a *statistical arbitrage* if the following three conditions hold:

$$\lim_{t \rightarrow \infty} E[v(t)] > 0 \quad (1)$$

$$\lim_{t \rightarrow \infty} \Pr[v(t) < 0] = 0 \quad (2)$$

$$\lim_{t \rightarrow \infty} \frac{\text{Var}[v(t)]}{t} = 0 \quad (3)$$

The authors then develop statistical tests that can be used with historical data to test whether specific strategies meet the definition of statistical arbitrage.

In this article, we think of statistical arbitrage in the following way. While arbitrage involves the trading of instruments that are identical or equivalent (e.g., a call and the hedging portfolio that replicates the call), but that have inconsistent prices, statistical arbitrage involves trading instruments that are different, but whose prices are mutually inconsistent according to some market model or historical price analysis. The price inconsistencies are identified by statistical techniques applied to historical data. A statistical arbitrage trading strategy places a set of trades on the inconsistently priced instruments, expecting those prices to return to their historical relationship. The statistical methods focus on the identification of *persistent anomalies* that violate the efficient market hypothesis, which can be used to create a trading strategy to generate profit with high probability.

The reader should be aware that there is no single widely accepted definition of statistical arbitrage. Indeed, the term is often used to connote a specific trading strategy, usually *pairs trading* (see section titled “Pairs Trading”).

Specific Trading Strategies

The focus of this section will be on some of the major statistical arbitrage strategies. These will include pairs trading, momentum, value and contrarian trading strategies, volatility trading, and technical analysis. In addition, we will introduce the concept of cointegration, originated by Engle and Granger [11], to broaden the concept of correlation, and we will see how it applies to pairs trading.

Pairs Trading

Suppose that an asset price process is governed by a mean-reverting process given by

$$dS_t = \alpha(\bar{S} - S_t) dt + \sigma(S_t) dW_t \quad (4)$$

where $\sigma(S_t)$ is some volatility function. Thus in the long run, the asset has an average price of $\bar{S}$. If the price at time t , S_t , is above $\bar{S}$, then a trader could short this asset in the expectation that, if the price process continues to be governed by this mean-reverting process with level $\bar{S} < S_t$, he or she will be able to sell and close the position at $\bar{S}$, thus

making a profit. Conversely, if the current asset price, S_t satisfied $S_t < \bar{S}$, then a trader could go long in the asset, expecting to close the position at $\bar{S}$, again making a profit.

We don't expect stocks to follow such a mean-reverting process (or, indeed, to be stationary); however, a pairs trader looks for a pair of stocks whose price difference or a weighted and shifted linear combination of which (to account for location and scale differences) follows a mean-reverting process. If one can find such a pair, then when the price of one of the two (say stock A) is high relative to the price of the other (say stock B), then a pairs trader would short a suitable amount of stock A and go long in a corresponding amount of stock B. When the prices of the two stocks go back into balance, then the trader closes the position. A pairs trading strategy can be an example of a *market neutral strategy*, one whose profitability does not depend upon the direction of the market. Rather it depends upon prices reverting to their long run equilibrium values relative to each other, not with respect to the market. The success of pairs trading as a statistical arbitrage strategy is evaluated in [12].

Pairs trading sounds like a “sure thing,” but there are several questions that need to be addressed:

- *how to find stocks A and B with this long run mean-reverting property?*
- *when to put on the position?*
- *how much of each stock to buy?*
- *when to close the position?*

One naïve approach to pairs trading is to calculate the correlation between all pairs of the price series. One would then trade those pairs with the highest correlations, whenever the price series diverged sufficiently. However, there is no justification for assuming that the difference between the two price series is stationary, hence new methodology is needed to handle this more general situation. This brings us to cointegration, which we discuss next.

Cointegration

Cointegration was developed by Engle and Granger [11] as a means of modeling dynamic codependencies in multivariate time series. While several individual series may themselves be nonstationary, we say

that they are *cointegrated* if we can find a linear combination that is stationary. This topic generalizes the notion of correlation, and an elementary discussion can be found in [13, Chapter 12]. Other good sources of information about cointegration and statistical tests for it include [14, 15].

A pairs trader would seek pairs (or triples, etc.) of stocks that are cointegrated. When the stationary linear combination of prices deviates far from its mean, a position could be opened with the amounts of each stock determined by the coefficients of the linear combination. When the stationary linear combination returned close to its mean, the position could be closed. One could also close a position at a loss if further deviation of the stationary linear combination from its mean appears to suggest that the stationary relationship between the stocks has changed.

Value and Contrarian Strategies

Value and contrarian strategies are based on fundamental analysis. One considers the intrinsic value of a stock and compares it with its market value using, for example, its book-to-market or P/E ratio as an indicator. A stock with a high book-to-market value is considered to be undervalued (and is a candidate for purchase in a value trading strategy), whereas a stock with a low book-to-market value is considered to be overvalued (and is a candidate for short selling in a value trading strategy). Such a strategy is also called *contrarian* as it favors purchase of stocks with share prices that are relatively low or which are “out of favor,” while it shorts stocks that are in favor.

DeBondt and Thaler [16] were among the first to show systematic anomalies in that “loser portfolios” (those with stocks that are out of favor) consistently outperformed the market. Conversely, “winner portfolios” (those with stocks that had recently outperformed the market) subsequently did much worse than the market. The authors related this phenomenon to “investor overreaction,” and cited principles of behavioral decision theory to justify their theory. Subsequent studies for the next decade would cite this paper and relate their work to it.

This work was significantly strengthened by Lakonishok *et al.* [17] who continued the ideas and considered trading strategies that went long in “value stocks” and short in “glamor stocks”. Membership in these two categories was defined in terms of some financial ratios including book to market, cash flow

to price, and P/E ratio. The authors clearly showed that this strategy led to returns in excess of market returns. They also dealt with the question whether the strong performance of the strategies was the result of their being inherently more risky.

Momentum Strategies

Everyone has heard the advice “buy low, sell high”. Momentum strategies follow the maxim “buy high, sell higher”. Suppose there is some sort of positive news (e.g., unexpectedly high earnings) that cause a company’s stock price suddenly to rise. Momentum strategies are based on the idea that the stock-price process will exhibit “momentum” and will rise even higher. Of course, there is a question as to whether there is any momentum effect at all, but if there is, the trader needs a stopping strategy to lock in the hoped for profits from additional price increases. Interestingly, a momentum trading strategy would seem to be essentially the direct opposite of the contrarian/value strategies discussed above.

The following papers utilize momentum strategies; however, the ways in which the portfolios are selected are different: [18–21]. These papers show overall that momentum-based strategies can lead to excess returns. The strategies that these authors use begin with a criterion defining momentum such as returns over a specific past period or closeness to the high price over a past period. They then study portfolios that are long in the stocks (or groups of stocks) that have high values of the momentum criterion and short in the stocks that have low values. There is also a position that argues against the efficacy of momentum strategies. Two such papers are [22, 23]. Among other things, they raise the issue of transaction costs. Their thesis is that the momentum strategies defined by the other authors show excess profits only because the transactions costs are ignored or understated.

Volatility Arbitrage Trading

The price of an option on an underlying tends to increase with increasing volatility of the underlying. One can take the formula for an option price together with the current option price and infer the value of the volatility (called the *implied volatility*). It is, however, possible that one may also have independent estimates of volatility, and those estimates may differ from the implied volatility. Popular ways to

estimate volatility range from naïve sample standard deviations of log returns to ARCH, GARCH, and EGARCH models, and the reader is referred to [13] for descriptions of these methods.

If, for example, the implied volatility is higher than the estimated volatility, then one would say that the option is overpriced with respect to the estimated volatility. This would lead a trader to short an option (or write an option) at the current price. One could also create a delta neutral portfolio by buying delta shares of the underlying and shorting the option. Of course, if the estimated volatility were to be larger than the implied volatility, then one would infer that the option is relatively underpriced and want to go long in the option and short in the stock. Strategies of this type, are discussed by Connolly [24]. Another type of volatility trade is a *straddle* in which one buys (or sells) both a call and a put option on the same underlying at the same strike with the same maturity. Straddles provide partial hedging against market movements. Noh *et al.* [25] do a study of methods for estimating volatility and compare them on the basis of how they would be used when choosing which straddles to invest in.

Market Factors

One of the classic studies that identified persistent anomalies is published in [26, 27] by Fama and French.

The abstract from [26] summarizes the results of this study:

This paper identifies five common risk factors (*see Integration of Risk Types*) in the returns on stocks and bonds. There are three stock-market factors: an overall market factor and factors related to firm size and book-to-market equity. There are two bond-market factors, related to maturity and default risks. Stock returns have shared variation due to the stock-market factors, and they are linked to bond returns through shared variation in the bond-market factors. Except for low-grade corporates, the bond-market factors capture the common variation in bond returns (*see Mathematical Models of Credit Risk*). Most important, the five factors seem to explain average returns on stocks and bonds.

Although multiple factors are cited, the work is most closely associated with company size in that stocks of small firms generally outperform stocks of large firms. A generalization of the Fama and

French factors was introduced by Carhart [28]. This paper adds a fourth factor, momentum, to the three stock-market factors identified by Fama and French. Furthermore, this paper is also interesting, because it looks at the performance of mutual funds.

Analysis of Trading Strategies

An important consideration, in addition to identifying potentially profitable trading strategies, is providing broader assessments of those strategies and statistical arbitrage in general. Some papers that have attempted such evaluations include [29, 30].

While considering trading strategies, one should also be cognizant of a variety of practical issues that can have a significant impact on their performance; however, some of these are not addressed in the various papers cited above. These might be grouped into four broad categories:

- *Transaction costs*

This category would include factors such as the charge to execute the trade, the price at which the trade is executed (i.e., bid-ask spread), the cost of odd-lot transactions, and fractional shares. An additional issue is the liquidity of the stock, and whether a purchase will influence the execution price. Lack of liquidity was a crucial ingredient in the failure of long-term capital management as described by Lowenstein [31].

- *Position holding costs and tax consequences*

Many trading strategies create portfolios that have both long and short positions. There can be additional costs to hold short positions. There are also tax consequences associated with trading strategies that can be difficult to evaluate and are largely ignored.

- *Survivorship bias*

Many of the papers mentioned above implement their strategies over a relatively long time period. Often the strategy is applied to a large set of stocks that have been listed continuously over the time period in question. This creates a survivorship bias, since stocks that go out of existence tend to be from companies that perform relatively poorly and go bankrupt or are acquired (*see Default Risk*). When one restricts an evaluation to stocks that survived over the time period in question, one is generally biasing the performance upward.

- *Riskiness of the strategy*

It is possible that a trading strategy might deliver excess profits, but do so because the strategy assumes higher than the market risk. One must control for the riskiness of the strategy.

Technical Analysis

In the academic literature, technical analysis has been considered to be nonsense. For example [7, p. 159] stated that “chart reading must share a pedestal with alchemy”. The assessment of technical analysis given in [7, Chapters 5 and 6] is scathing. However, in more recent times, computer-based techniques drawn from data mining or machine learning have become more popular, and these might be considered to be the twenty-first century version of technical analysis.

We cannot begin to describe here the many technical indicators that are available. They range from simple patterns such as a succession of three up and down cycles in which the middle cycle rises higher than the two outer cycles (head-and-shoulders) to complicated patterns in which the difference of two exponentially weighted moving averages is compared with a moving average of the difference (moving average convergence divergence). Technical indicators generally signal trades for one security at a time, rather than for pairs or large market portfolios.

A serious academic study of technical analysis was done by Lo *et al.* [32]. The authors automate many of the subjective choices for which technical analysis is criticized, and apply their automated procedures to US stocks from 1962 to 1996 and five pairs of technical indicators. They find that returns following the occurrence of one of the indicators do tend to be different from the returns of purchases that ignore the indicators. A critical discussion of this paper appeared in the same journal: [33]. Bauer and Dahlquist [34] describe 120 technical indicators, and do a systematic comparison of the signals that they generate to buy-and-hold strategies.

References

- [1] Graham, B. & Dodd, D. (1934). *Security Analysis*, McGraw-Hill, New York.
- [2] Graham, B. (1949). *The Intelligent Investor*, Harper & Row, New York.
- [3] Graham, B. & Meredith, S. (1937). *The Interpretation of Financial Statements*, Harper & Brothers, New York.
- [4] Edwards, R.D., Magee, J. & Bassetti, C. (2001). *Technical Analysis of Stock Trends*, 8th Edition, St. Lucie Press, Boca Raton.
- [5] Samuelson, P. (1965). Properly anticipated stock prices fluctuate randomly, *Industrial Management Review* **6**, 41–49.
- [6] Fama, E. (1970). Efficient capital markets: a review of theory and empirical work, *Journal of Finance* **25**, 383–417.
- [7] Malkiel, B. (1973). *A Random Walk Down Wall Street*, Norton, New York.
- [8] Malkiel, B. (2003). The efficient market hypothesis and its critics, *Journal of Economic Perspectives* **17**, 59–82.
- [9] Shleifer, A. & Vishny, R. (1997). The limits of arbitrage, *Journal of Finance* **52**, 35–55.
- [10] Hogan, S., Jarrow, R., Teo, M. & Warachka, M. (2004). Testing market efficiency using statistical arbitrage with applications to momentum and value strategies, *Journal of Financial Economics* **73**, 525–565.
- [11] Engle, R. & Granger, C. (1987). Co-integration and error correction: representation, estimation, and testing, *Econometrica* **55**, 251–276.
- [12] Gatev, E., Goetzmann, W. & Rouwenhorst, K. (2006). Pairs trading: performance of a relative-value arbitrage rule, *Review of Financial Studies* **19**, 797–827.
- [13] Alexander, C. (2001). *Market Models*, John Wiley & Sons, Chichester.
- [14] Johansen, S. (1988). Statistical analysis of cointegration vectors, *Journal of Economic Dynamics and Control* **12**, 231–254.
- [15] Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models, *Econometrica* **59**, 1551–1580.
- [16] DeBondt, W. & Thaler, R. (1985). Does the stock market overreact? *Journal of Finance* **40**, 793–808.
- [17] Lakonishok, J., Schleifer, A. & Vishny, R. (1994). Contrarian investment, extrapolation, and risk, *Journal of Finance* **49**, 1541–1578.
- [18] Jegadeesh, N. & Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency, *Journal of Finance* **48**, 65–91.
- [19] Jegadeesh, N. & Titman, S. (2001). Profitability of momentum strategies: an evaluation of alternative explanations, *Journal of Finance* **56**, 699–718.
- [20] George, T.J. & Hwang, C.-Y. (2004). The 52-week high and momentum investing, *Journal of Finance* **59**, 2145–2176.
- [21] Moskowitz, T. & Grinblatt, M. (1999). Do industries explain momentum? *Journal of Finance* **54**, 1249–1290.
- [22] Keim, D.B. (2003). The cost of trend chasing and the illusion of momentum profits, Technical report 11-03, Wharton School.
- [23] Lesmond, D.A., Schill, M.J. & Zhou, C. (2004). The illusory nature of momentum profits, *Journal of Financial Economics* **71**, 349–380.

- [24] Connolly, K.B. (1997). *Buying and Selling Volatility*, John Wiley & Sons, Chichester.
- [25] Noh, J., Engle, R. & Kane, A. (1993). *A Test of Efficiency for the S & P 500 Index Option Market Using Variance Forecasts*, Working paper 4520, NBER.
- [26] Fama, E.F. & French, K. (1993). Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* **33**, 3–56.
- [27] Fama, E.F. & French, K. (1996). Multifactor explanations of asset pricing anomalies, *Journal of Finance* **51**, 55–84.
- [28] Carhart, M.M. (1997). On persistence in mutual fund performance, *Journal of Finance* **52**, 57–82.
- [29] Ahn, D.-H., Conrad, J. & Dittmar, R. (2003). Risk adjustment and trading strategies, *Review of Financial Studies* **16**, 459–485.
- [30] Conrad, J. & Kaul, G. (1998). An anatomy of trading strategies, *Review of Financial Studies* **11**, 489–519.
- [31] Lowenstein, R. (2000). *When Genius Failed: The Rise and Fall of Long-Term Capital Management*, Random House, New York.
- [32] Lo, A.W., Mamaysky, H. & Wang, J. (2000). Foundation of technical analysis: computational algorithms, statistical inference, and empirical implementation, *Journal of Finance* **55**, 1705–1765.
- [33] Jegadeesh, N. (2000). Foundation of technical analysis: computational algorithms, statistical inference, and empirical implementation: discussion, *Journal of Finance* **55**, 1765–1770.
- [34] Bauer, R.J. & Dahlquist, J.R. (1999). *Technical Market Indicators: Analysis and Performance*, John Wiley & Sons, New York.

Related Articles

Numerical Schemes for Stochastic Differential Equation Models

Risk-Neutral Pricing: Importance and Relevance

JOHN P. LEHOCZKY AND MARK J. SCHERVISH

Statistics for Environmental Justice

Definition and Background

The United States Environmental Protection Agency (USEPA) defines the term *environmental justice* as "... the fair treatment and meaningful involvement of all people regardless of race, ethnicity, income, national origin or educational level with respect to the development, implementation, and enforcement of environmental laws, regulations, and policies. Fair treatment means that no population, due to policy or economic empowerment, is forced to bear a disproportionate burden of the negative human health or environmental impacts of pollution or other environmental consequences resulting from industrial, municipal, and commercial operations or the execution of federal, state, local, and tribal programs and policies" [1].

This definition formalized a growing grassroots movement addressing the distribution of environmental hazards across sociodemographic groups, initially questioning whether hazardous waste sites or toxic facilities were disproportionately sited in areas with high minority populations. The movement began with local activists in the late 1970s and early 1980s and received national attention with the release of a 1987 report by the Commission for Racial Justice of the United Church of Christ (UCC) [2] following up on efforts beginning in 1982 in Warren County, North Carolina regarding the siting of a proposed polychlorinated biphenyl (PCB) disposal landfill in an area with a large proportion of African-American residents. In response to the UCC report, an earlier government study [3], and the efforts of the "Michigan Coalition" [4], the US EPA formed an Environmental Equity Working Group in 1990, and the EPA Office of Environmental Equity in 1992, which is now known as the *EPA Office of Civil Rights*.

In order to consider quantitative assessments of environmental justice, it is necessary to understand its regulatory setting. In 1994, President Clinton issued Executive Order 12898 requiring each federal agency to "analyze the environmental effects, including human health, economic, and social effects" of federally sponsored programs. For a more detailed historical development, see [5] and [6], and for the

complete text of Executive Order 12898, see [7]. In effect, Executive Order 12898 links two earlier pieces of legislation, namely, the National Environmental Policy Act (NEPA) of 1969 [8] and Title VI of the Civil Rights Act of 1964 [9]. NEPA established the regulatory requirement of an environmental impact statement (EIS) for all federally sponsored activities "significantly affecting the quality of human environment" and the inclusion of any socioeconomic impacts related to significant changes in the physical environment. Title VI of the Civil Rights Act prohibits any recipient of federal funds from taking action that results in a discriminatory impact on the basis of race, sex, color, national origin, religion, age, or disability. If a recipient of federal assistance is found to have discriminated and voluntary compliance cannot be achieved, the US Department of Justice is allowed to initiate fund termination or other legal action. Liu [6] points out that NEPA is associated with proposed future environmental impacts while Title VI pertains to past, present, and future actions and decisions by any entity receiving federal assistance.

Executive Order 12898 places environmental justice within the same legal setting as other nondiscrimination issues and, as a result, links environmental justice assessments to ongoing judicial and legislative interpretations and opinions regarding the standards of evidence required to assess Title VI complaints. We note that, while the general conceptual issues of environmental justice apply to any population subgroups, most assessments of environmental justice to date default with respect to legal and regulatory issues involving subgroups covered under Title VI and Executive Order 12898, i.e., age, race, income, but particularly race.

The legal and regulatory definitions of environmental justice provide the background for guidelines and proposals for investigative procedures to address complaints. Of particular interest to quantitative risk assessment is the role of statistical summaries and measurements. In other words, what does one need to quantify in order to demonstrate environmental injustice?

The US Department of Justice outlines several key legal concepts and the role of statistical evidence in investigating Title VI complaints [10]. A key distinction is that between "intentional discrimination/disparate treatment" where similarly situated persons (individuals or classes) are treated differently

intentionally, and “disparate impact/effects” where persons are treated differently with or without intent by another party. Proving “intentional discrimination” requires the complainant to prove both disparate impacts as well as the intent behind them. Proving “disparate impact” does not require intent and, in fact, may be the result of a “neutral policy or practice that has the effect of disproportionately excluding or adversely affecting members of a protected group” [10]. Past rulings of the US Supreme Court have upheld a “disparate impact” standard for proving Title VI violations in general [11, 12] with similar supporting decisions from lower courts.

Proving disparate impact often relies on statistical evidence and the US Department of Justice specifies the role of such evidence by stating “If: evidence reveals a statistically significant difference in the treatment of similarly situated classes of different races (or sexes) in the delivery of services (or receipt of benefits, in employment levels, etc.), and there is no legitimate nondiscriminatory reason for the observed pattern of disparities, then it is reasonable to infer that race was a factor in creating such a pattern” [10]. The US Department of Justice Title VI Investigative Manual next stresses that statistical techniques are required to demonstrate that “substantial disparate impact exists and that the observed impact is statistically significant” [10].

Most legal and research investigations of environmental justice focus on measuring disparate impact in environmental exposures rather than disparate impact in subsequent outcomes related to the exposure. A focus on outcomes complicates analysis since it would require one to show differences between subgroups with respect to both exposures and outcomes, and to show that the observed outcome difference is caused by the exposure difference.

Data Sources

Typical assessments of environmental justice in the United States utilize data from at least two sources: demographic profiles for small enumeration areas (e.g., census tracts or block groups) and exposure estimates across a study area of interest. Both values vary across space, leading to a central role for geographic information systems (GISs) in environmental justice assessments [6] (*see Geographic Disease Risk; Hotspot Geoinformatics*). Population demographics

often reveal higher local concentrations of minority and/or low-income populations, and investigation hinges on whether these local concentrations are associated with local concentrations of high exposure to environmental hazards.

Reliable demographic data are obtained from census projections, but sources of environmental exposure data vary from study to study. In some cases, only the location of the source (proposed or actual) is known and exposure estimates are based on approximated or modeled values based on actual or estimated releases from the source. Few assessments use monitored or measured exposure values since many assessments (consistent with the link to EIS preparation) are in response to planned rather than existing sources of environmental hazard (*see Environmental Monitoring*).

Statistical Methods

When summarizing statistical methods for environmental justice assessments, two broad categories of comparisons emerge based on the availability and/or extent of exposure data. In the absence of measured or approximated exposure values, the first set of comparative methods revolves around proximity-based exposure classifications. Such classifications may be based on simple distance radii with individuals residing within a given radius classified as exposed and those residing outside of the radius classified as unexposed. More sophisticated options are also available within GISs such as classifications based on travel times or distances along road networks, but the general principle is the same. In this setting, one assesses environmental justice by comparing population demographics across exposure classes. For instance, one could compare the proportions classified as African-American for individuals residing in the exposed and nonexposed areas. Such applications are very common in the environmental justice literature, and allow assessments using only locations of hazards rather than measured exposure values. However, many reports illustrate that such approaches can lead to very different conclusions for different exposure-defining radii, partially due to geographic heterogeneities in the distribution of demographic subpopulations (e.g., the population within 1 mile of a proposed site location might contain a majority of one racial group while the population

within 3 miles contains a majority of another group) and also potentially due to confounding variables impacting both demographics and exposure (e.g., land values, neighborhood average incomes) [6, 13–15]. The confounding issue may be addressed, at least in part, through the use of regression-based methods measuring associations between population proportions and exposure classifications while adjusting for other covariates. Greenberg [13] and Liu [6] discuss the advantages and disadvantages of regression-based methods to adjust associations between subpopulation proportions and exposure classification for potentially confounding factors. Such analyses will adjust the measured associations but are still subject to the selection of the exposure-defining classification, and to the effects of potential collinearities between covariates such as race and income.

The second set of comparative methods involves quantitative exposure values or surrogates. As noted above, many environmental justice studies involve planned or potential exposures or emissions rather than measured ambient or personal exposures. As a result, very few published environmental justice assessments make use of monitored exposure data, and most utilize some sort of proximity-based exposure surrogate. Once one has obtained exposure measures (e.g., through ambient or personal monitoring) or exposure surrogates (e.g., based on proximity) for each individual or census subregion, one defines comparison groups by the subpopulations of interest (e.g., racial or ethnic groups), and then compares the distributions of exposure measurements or surrogates between these groups. Such comparisons are appealing in that comparisons of exposures between demographic subgroups seem to coincide better with the question of interest than do comparisons of demographics between exposure groups. Comparisons of exposures (or exposure surrogates) could be straightforward two-sample tests comparing means or medians, or could involve comparisons of the exposure distributions for each subpopulation using tests such as Kolmogorov–Smirnov tests, probability plots, relative distribution functions [16], or integrated cumulative exposure distributions [15].

The key factors in deciding between comparisons of population structure across exposure surrogate groups and comparisons of experienced exposure surrogates across population groups appear to be data availability and justification of a single exposure or set of exposures of interest. In the case

of monitored ambient exposures, measurements may be unavailable for the time periods of interest for retrospective studies of existing hazards, or simply not in existence for proposed future sources of hazard. If, as is usually the case, exposure data are not available, proximity-based measures often act as surrogates either for exposure classification (the radius-based classification) or as exposure surrogates themselves (e.g., inverse-distance approximations [15]). In some cases, some exposure measurements may be available, but the driving justification for a particular assessment of environmental justice may be based on exposures to a collection of hazards (from a single or multiple sources) rather than to a particular measurable substance or substances. In such cases, various collective and/or cumulative exposure indices may be proposed or used, such as the population emissions index (PEI) [17], which is based on a weighted average of emissions within each population subgroup with weights based on population proportions within small enumeration districts.

In summary, environmental justice represents a maturing issue comprised of complex interrelated legal, regulatory, risk, demographic, exposure, and statistical issues. An appreciation of each of these areas provides a context for improved data collection, more accurate and reliable assessment methodologies, and appropriate responses.

References

- [1] United States Environmental Protection Agency (1998). *Final Guidance for Incorporating Environmental Justice Concerns in EPA's NEPA Compliance Analyses*, U.S. Government Printing Office, Washington, DC, p. 2.
- [2] United Church of Christ (1987). *Toxic Wastes and Race in the United States: A National Report on the Racial and Socio-Economic Characteristics of Communities with Hazardous Waste Sites*, New York.
- [3] United States General Accounting Office (1983). *Siting of Hazardous Waste Landfills and Their Correlation with Racial and Socio-Economic States of Surrounding Communities*; United States Government Printing Office, Washington, DC.
- [4] Bryant, B. & Mohai, P. (1992). The Michigan conference: a turning point, *EPA Journal* **18**, 9–10.
- [5] Sexton, K., Olden, K. & Johnson, B.L. (1993). Environmental justice: the central role of research in establishing a credible scientific foundation for informed decision making, *Toxicology and Industrial Health* **9**, 685–727.
- [6] Liu, F. (2001). *Environmental Justice Analysis: Theory, Methods, and Practice*, CRC Press, Boca Raton.

- [7] Institute of Medicine (1999). *Toward Environmental Justice: Research, Education, and Health Policy Needs*, National Academy Press, Washington, DC.
- [8] 42 United States Code § 4321.
- [9] 42 United States Code §§ 2000d.
- [10] United States Department of Justice (1998). *Investigating Procedures Manual for the Investigation and Resolution of Complaints Alleging Violations of Title VI and other Nondiscriminatory Statutes*, Department of Justice, Civil Rights Division, Coordination and Review Section, Washington DC, pp. 71, 85–86.
- [11] Guardians Association v. Civil Service Commission of New York City, (1983). 463 U.S. 582.
- [12] Alexander v. Choate, (1985). 469 U.S. **287** 293–294.
- [13] Greenberg, M.R. (1993). Proving environmental inequity in the siting of locally unwanted land uses, *Risk: Issues in Health and Safety* **4**, 235–252.
- [14] Anderton, D.L., Anderson, A.B., Oakes, J.M. & Fraser, M.R. (1994). Environmental equity: the demographics of dumping, *Demography* **31**, 229–248.
- [15] Waller, L.A., Louis, T.A. & Carlin, B.P. (1997). Bayes methods for combining disease and exposure data in assessing environmental justice, *Environmental and Ecological Statistics* **4**, 267–281.
- [16] Handcock, M.S. & Morris, M. (1999). *Relative Distribution Methods in the Social Sciences*, Springer, New York.
- [17] Perlin, S.A., Setzer, R.W., Creason, J. & Sexton, K. (1995). Distribution of industrial air emissions by income and race in the United States: an approach using the toxic release inventory, *Environmental Science and Technology* **29**, 69–80.

Related Articles

Cumulative Risk Assessment for Environmental Hazards

Disease Mapping

Environmental Health Risk

History and Examples of Environmental Justice

Stakeholder Participation in Risk Management Decision Making

LANCE A. WALLER

Statistics for Environmental Mutagenesis

Laboratory experiments to assess the toxic potential of hazardous substances (*see What are Hazardous Materials?*) often generate data whose statistical analysis depends on the nature of the endpoint and the aspects of the particular assay under study [1]. Among major endpoints of risk analytic value, environmental damage to genetic components such as DNA is of increasing interest [2, 3]. Mutation induction, or *mutagenesis*, and other forms of genotoxicity due to environmental exposures represent fruitful areas of study, particularly as continuing biotechnological advances allow for more precise genetic study [4].

Mutagenesis is perhaps the most varied of any toxicological mechanism, since the potential components of genetic damage are so numerous. Indeed, there exist well over 100 different genotoxicological test systems, employing all sorts of animal and microbial systems [5]. Statistical themes associated with environmental mutagenesis experiments include proper identification of the statistical nature of the outcome data [6–10], construction of appropriate biomathematical models to characterize dose–response and other features of genetic damage [11–16], sample size determinations [17, 18], and the selection of statistical tests based on these models [19, 20]. An introduction to these various issues is given in the compilation by Kirkland [21] and the review by Edler [22]. Herein, attention will be directed at some of the basic issues of dose–response testing and analysis useful for risk assessments of environmental mutagens. For an informative article that emphasizes the risk regulatory questions associated with mutagenesis and genetic toxicity data, see Dearfield and Moore [23].

Microbial Systems in Environmental Mutagenesis: the *Salmonella* Assay

Perhaps the most well known of all modern mutagenicity assays is the Ames/*Salmonella* microsome assay [24], employing the bacteria *Salmonella*

typhimurium to identify damage to DNA after exposure to toxic environmental agents (*see Environmental Hazard*). The assay is based on strains of the bacterium unable to synthesize histidine, an amino acid required for growth. This production deficiency can be reversed into a production capability *via* point mutations at selected sites on the bacterial genome. DNA damage is indicated by mutation of the bacteria from histidine dependency to a self-sustaining state. Thus, the mutated cells can grow in a microenvironment (such as a petri dish) containing only minimal amounts of histidine; greater mutant yield at higher exposures to the environmental agent suggests that mutagenesis increases with increasing dose. Observational accuracy of the assay is enhanced by use of a selective medium for the growth environment, so that only mutant colonies grow after exposure to the environmental toxin.

The observations from this microbial assay are the mutated colony counts, Y_{ij} , for the j th replicate plate at the i th dose ($i = 1, \dots, T$; $j = 1, \dots, R_i$). Typically, the number of replicate plates is taken to be constant, e.g., $R_i = 3 \forall i$. Table 1 presents an example of data from a *Salmonella* assay, using the chemical agent 1,3-butadiene. A natural candidate for the sampling distribution of Y_{ij} is the Poisson distribution for count variables. Indeed, many processes that drive the Poisson process are active with the colony count data seen here. These include (a) the microbes' responses are independent, (b) each locally organized group of microbes experiences the same environment (ignoring controlled variables such as dose), (c) the plated number of microbes is large, say 10^8 , and (d) the probability that a plated microbe gives rise to a visible mutant colony, is small here, between 10^{-5} and 10^{-9} [14]. A fifth assumption crucial to adoption of the Poisson distribution is, however, that both the environment and the number of microbes plated should remain relatively stable across replicate plates (within a dose group). If this is not the case, extra-Poisson variability results, and one is forced to model or otherwise account for the consequent overdispersion in Y_{ij} .

Testing for Extra-Poisson Variability

Testing for extra-Poisson variability is critical when analyzing count data from environmental mutagenesis experiments. Standard trend tests for

Poisson-distributed counts become too sensitive in the presence of overdispersion, increasing type I error rates above the nominally specified α level [25]. Thus, the data analyst must be aware of the nature and level of any overdispersion in the data before selecting a test for dose–response.

Extra-Poisson variability in a set of ostensibly homogeneous observations is assessed *via* a variance-to-mean comparison. The test employs Fisher’s dispersion statistic [26, 27]:

$$X_i^2 = \sum_{j=1}^{R_i} \frac{(Y_{ij} - \bar{Y}_i)^2}{\bar{Y}_i} \quad (1)$$

where $\bar{Y}_i$ is the i th dose sample mean ($i = 1, \dots, T$). Asymptotically, X_i^2 is distributed as $\chi^2(R_i - 1)$, hence $X_i^2 > \chi_{1-\alpha}^2(R_i - 1)$ implies significant extra-Poisson variability. This test is $C(\alpha)$ -optimal [28] against the negative binomial distribution, a standard alternative to the Poisson model. To extend equation (1) to multiple test groups, simply aggregate the individual X_i^2 statistics into a sum: $X^2 = \sum_{i=1}^T X_i^2$. Asymptotically, X^2 is distributed as $\chi^2(\sum_{i=1}^T R_i - T)$. (A slightly more powerful test of extra-Poisson variability may be constructed using the similar statistic $C^2 = \sum_{i=1}^T \sum_{j=1}^{R_i} (Y_{ij} - \bar{Y}_i)^2 / \bar{Y}$, where $\bar{Y} = \sum_{i=1}^T R_i \bar{Y}_i / \sum_{i=1}^T R_i$ [29].) Dean and Lawless [30] discuss additional extensions to more complex regression settings; also see Kim and Park [31] and Xiang and Lee [32].

A common application of these tests is to large, controlled databases where similar control trials and replicate data are available. Margolin *et al.* [14] reported on such data, demonstrating that the Poisson model often inadequately describes statistical variability when analyzing *Salmonella* mutagenicity experiments. Further research has suggested that *Salmonella* data often exhibit extra-Poisson variability, and that the parent distribution of the data may be better described by a negative binomial distribution [19]. Other publications indicate, however, that the Poisson model can be acceptable in select instances [33], particularly for test data obtained by imposing strict protocol adherence and by harvesting all the data on the same day [34, 35]. Given these considerations, testing for extra-Poisson variability is of clear necessity when analyzing *Salmonella* dose–response data (see **Dose–Response Analysis**).

Trend Tests for Dose–Response

For mutagenicity experiments where no extra-Poisson variability is evidenced, the well-known Cochran–Armitage trend test for dose–response may be employed. Under Poisson sampling, the test is known to be optimal against any monotone dose–response function [36]. Given the observed counts Y_{ij} ($j = 1, \dots, R_i; i = 1, \dots, T$), and denoting x_i as some increasing score (dose, log-dose, etc.) associated with treatment level i , the test statistic is

$$Z_{CA} = \frac{\sum_{i=1}^T X_i \left[\left(\sum_{j=1}^{R_i} Y_{ij} \right) - R_i \bar{Y} \right]}{(\bar{Y} S_x^2)^{1/2}} \quad (2)$$

in which $S_x^2 = \sum_{i=1}^T R_i (x_i - \bar{x})^2$ and $\bar{x} = \sum_{i=1}^T R_i x_i / \sum_{i=1}^T R_i$. Asymptotically, Z_{CA} is distributed as standard normal, and significant dose–response is suggested when Z_{CA} is larger than an appropriate upper- α critical point from a standard normal distribution, z_α .

If the response is overdispersed and the Poisson model is invalid, a form of Cochran–Armitage trend test may still be constructed. Under a negative binomial parent model, one simply replaces $(\bar{Y} S_x^2)^{1/2}$ in equation (2) with an updated estimator based on estimating the negative binomial variance [19]. The result is

$$Z_{NB} = \frac{\sum_{i=1}^T x_i \left[\left(\sum_{j=1}^{R_i} Y_{ij} \right) - R_i \bar{Y} \right]}{[\bar{Y} (1 + \hat{\delta} \bar{Y}) S_x^2]^{1/2}} \quad (3)$$

The dispersion estimator, $\hat{\delta}$, is calculated *via* the method of moments, maximum likelihood, maximum quasi-likelihood, or more advanced techniques [37–40]. As above, significant dose–response is suggested when Z_{NB} is larger than a standard normal critical point, z_α .

When the parent distribution of the data is not known or indeterminate, it is possible to construct a generalized score statistic for testing trend [41].

Table 1 *Salmonella* mutagenicity results for 1,3-butadiene in strain TA1535

Dose (ppm)	Mutants per plate, Y_{ij}			Average plate count, $\bar{Y}_i$
0.000	20	31	27	26.0
0.002	100	92	89	93.7
0.007	147	123	178	149.3
0.014	216	170	181	189.0
0.020	176	154	183	171.0
0.030	154	153	149	152.0

This is

$$Z_{GS} = \frac{\sum_{i=1}^T R_i(x_i - \bar{x})\bar{Y}_i}{\sqrt{\sum_{i=1}^T R_i(x_i - \bar{x})^2 \sum_{j=1}^{R_i} (Y_{ij} - \bar{Y})^2}} \quad (4)$$

similar to the basic statistics in equation (2) or equation (3). Z_{GS} is distributed asymptotically as standard normal; significant dose–response is suggested when $Z_{GS} > z_{\alpha}$.

Trend Test under Nonmonotone Dose–Response

An important, additional feature illustrated by the data in Table 1 is a curious downturn in the response at higher doses. This form of nonmonotone dose–response is common in the *Salmonella* assay [42], and is often observed with other mutagenesis assays as well [43].

In *Salmonella*, the downturn phenomenon is driven by a number of possible mechanisms, the most common of which is a consequence of the experimental scheme employed to identify the mutational events. Since mutagenesis is identified by growth on a selective medium, any other mechanism that hinders growth will compete with the desired outcome. Thus, e.g., high exposures of the toxic agent can lead to cell death or perhaps chemical/threshold induced increases in DNA repair. These factors conspire to *reduce* the yield of mutated cells by killing or neutralizing the microbes before mutations can be observed. The result is a downturn at the higher doses, called an *umbrella response* [44]. As might be expected, simple tests for monotone trend such as Z_{CA} , Z_{NB} , or Z_{GS} can exhibit large losses in power to detect any increase in dose–response in the presence

of a downturn [41, 45], and suitable alternatives must be employed.

Nonlinear biomathematical and empirical models for this downturn phenomenon are available, which, when coupled with the negative binomial sampling model on the Y_{ij} 's, involve fairly advanced statistical methodology [46, 47]. A simple alternative is to identify statistically the point of maximal departure from background mutation levels, and then test for increasing trend up to that point. Various procedures have been suggested for such an approach [48, 49], the majority of which analyze the ranks of the observations *via* “nonparametric” techniques based on the ranks of the observations.

For example, a nonparametric approach developed by Simpson and Margolin [49] estimates recursively the point of maximal response (the *umbrella point*), and then tests for a significant increase in dose–response up to that point. One begins by calculating a series of two-sample Mann–Whitney statistics [50], W_i , comparing the i th dose group with – collectively – all preceding dose groups, starting at $i = T$ and working backwards. That is, W_T tests whether the response information at $i = T$ departs significantly above the pooled response information at $i = 1, \dots, T - 1$. If so, an estimate, h , of the umbrella index is $h = T$. If not, one discards the information at $i = T$, and repeats the process at $i = T - 1$. In this way, the test employs the values of W_i ($i = T, T - 1, \dots, 2$) to estimate the umbrella index *via* recursive pretesting. In effect, it selects this estimate as the largest dose index, h , such that W_h is larger than a specified critical value c_h ; i.e., $h = \max\{i \in \{2, \dots, T\} : W_i > c_i\}$, where $c_2 = 0$ and $c_i = c_i(q)$ ($i = 3, \dots, T$) is the q quantile of the distribution of W_i under the null hypothesis of no dose–response. The tuning parameter $q \in (0, 1)$ must be prespecified, and is usually set at or near $q = 0.5$ [45].

Once the umbrella index, h , is estimated, the recursive procedure calculates the conditional Jonckheere–Terpstra trend statistic $U_h = W_1 + \dots + W_h$ [51], and an increasing dose–response is suggested when U_h is larger than the $(1 - \pi)$ quantile of the distribution of U_i under the null hypothesis of no differences among the doses ($i = 2, \dots, T$), for $0 < \pi < 1$. For fixed q and prespecified significance level α , a conservative choice for π is $\pi = \alpha(1 - q)/(1 - q^{T-1})$. A computing algorithm for these calculations is described in Simpson and Dallal [52], and extensions and other research issues involved with rank-based testing for *Salmonella* data are discussed in Schumacher and Schmoor [42] and Cologne and Breslow [53]. Also see the review by Kim and Margolin [20].

Example: 1,3-Butadiene

Consider the example of the airborne toxin 1,3-butadiene, the data for which appear in Table 1. These values represent *Salmonella* mutagenic response after gaseous exposure to the chemical, where an increase in the number of observed colonies suggests a mutagenic effect.

Begin by testing whether the data exhibit any overdispersion, relative to the Poisson sampling model. Employing the aggregated dispersion statistic using equation (1) to the data in Table 1 gives $X^2 = 22.134$ on 12 df ($P = 0.004$). Strong departure from Poisson variability is evidenced. If the negative binomial sampling model were considered as an adequate replacement to the Poisson for these data, then a trend test for increasing dose–response based on equation (3) gives $Z_{NB} = 1.285$, with one sided $P = 0.099$. Marginal suggestion of an increasing trend is given.

For these data, however, a downturn in the dose–response is evident. This could affect the trend test’s ability to identify properly a positive dose–response. Application of the recursive test for umbrella response is therefore appropriate: setting the tuning parameter to $q = 0.5$, the umbrella index is estimated as $h = 6$, with corresponding rank-based trend statistic given by $U_6 = 105.5$. Rejection occurs at significance levels as low as $\alpha = 0.005$, providing strong evidence of a positive dose–response for these mutagenicity data.

References

- [1] Piegorsch, W.W. (1994). Environmental biometry: assessing impacts of environmental stimuli via animal and microbial laboratory studies, in *Handbook of Statistics Volume 12: Environmental Statistics*, G.P. Patil & C.R. Rao, eds, North-Holland/Elsevier, Amsterdam, pp. 535–559.
- [2] Cuenca, P. & Ramirez, V. (2004). Mutagénesis ambiental y el uso de biomarcadores para predecir el riesgo de cáncer (in Spanish), *Revista De Biología Tropical* **52**, 585–590.
- [3] Neri, M., Bonassi, S., Knudsen, L.E., Sram, R.J., Holland, N., Ugolini, D. & Merlo, D.F. (2006). Children’s exposure to environmental pollutants and biomarkers of genetic damage I. Overview and critical issues, *Mutation Research* **612**, 1–13.
- [4] Schmidt, C.W. (2003). Data explosion. Bringing order to chaos with bioinformatics, *Environmental Health Perspectives* **111**, A340–A345.
- [5] Waters, M.D., Stack, H.F., Brady, A.L., Lohman, P.H.M., Haroun, L. & Vainio, H. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans (1987). Activity profiles for genetic and related tests, Appendix I, in *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Genetic and Related Effects: An Updating of Selected IARC Monographs from Volumes 1 to 42*, International Agency for Research on Cancer, Lyon, pp. 687–696.
- [6] Lockhart, A.-M., Piegorsch, W.W. & Bishop, J.B. (1992). Assessing overdispersion and dose–response in the male dominant lethal assay, *Mutation Research* **272**, 35–58.
- [7] Margolin, B.H., Kim, B.S. & Risko, K.J. (1989). The Ames Salmonella/microsome mutagenicity assay: issues of inference and validation, *Journal of the American Statistical Association* **84**, 651–661.
- [8] Radack, K.L., Pinney, S.M. & Livingston, G.K. (1995). Sources of variability in the human lymphocyte micronucleus assay: a population-based study, *Environmental and Molecular Mutagenesis* **26**, 26–36.
- [9] Boerrigter, M.E.T.I. & Vijg, J. (1997). Sources of variability in mutant frequency determinations in different organs of *lacZ* plasmid-based transgenic mice: experimental features and statistical analysis, *Environmental and Molecular Mutagenesis* **29**, 221–229.
- [10] Piegorsch, W.W., Lockhart, A.C., Carr, G.J., Margolin, B.H., Brooks, T., Douglas, G.R., Liegibel, U.M., Suzuki, T., Thybaud, V., van Delft, J.H.M. & Gorelick, N.J. (1997). Sources of variability in data from a positive selection *lacZ* transgenic mouse mutation assay: an interlaboratory study, *Mutation Research* **388**, 249–289.
- [11] Ager, D.D. & Haynes, R.H. (1987). Mathematical description of the interactions between cellular inactivating agents, *Radiation Research* **110**, 129–141.
- [12] Alvord, W.G., Driver, J.H., Claxton, J. & Creason, J.P. (1990). Methods for comparing Salmonella mutagenicity

- data sets using nonlinear models, *Mutation Research* **240**, 177–194.
- [13] Krewski, D., Leroux, B.G., Bleuer, S.R. & Broekhoven, L.H. (1993). Modeling the Ames Salmonella/microsome assay, *Biometrics* **49**, 499–510.
- [14] Margolin, B.H., Kaplan, N. & Zeiger, E. (1981). Statistical analysis of the Ames Salmonella/microsome test, *Proceedings of the National Academy of Sciences of the United States of America* **76**, 3779–3783.
- [15] Roldán-Arjona, T., Pueyo, C. & Haynes, R.H. (1995). Mathematical parameters for quantification of mutational responses in bacteria, *Mutation Research* **346**, 77–84.
- [16] Edler, L. & Kopp-Schneider, A. (1998). Statistical models for low dose exposure, *Mutation Research* **405**, 227–236.
- [17] Margolin, B.H., Collings, B.J. & Mason, J.J. (1983). Statistical analysis and sample-size determinations for mutagenicity experiments with binomial response, *Environmental Mutagenesis* **5**, 705–716.
- [18] Piegorsch, W.W., Margolin, B.H., Shelby, M.D., Johnson, A., French, J.E., Tennant, R.W. & Tindall, K.R. (1995). Study design and sample sizes for a *lacI* transgenic mouse mutation assay, *Environmental and Molecular Mutagenesis* **25**, 231–245.
- [19] Margolin, B.H. (1985). Statistical studies in genetic toxicology: a perspective from the U.S. National Toxicology Program, *Environmental Health Perspectives* **63**, 187–194.
- [20] Kim, B.S. & Margolin, B.H. (1999). Statistical methods for the Ames Salmonella assay: a review, *Mutation Research* **436**, 113–122.
- [21] Kirkland, D.J. (1989). *Statistical Evaluation of Mutagenicity Test Data*, Cambridge University Press, Cambridge.
- [22] Edler, L. (1992). Statistical methods for short-term tests in genetic toxicology: the first fifteen years, *Mutation Research* **277**, 11–33.
- [23] Dearfield, K.L. & Moore, M.M. (2005). Use of genetic toxicology information for risk assessment, *Environmental and Molecular Mutagenesis* **46**, 236–245.
- [24] Ames, B.N., McCann, J. & Yamasaki, E. (1975). Methods for detecting carcinogens and mutagens with the Salmonella/mammalian microsome mutagenicity test, *Mutation Research* **31**, 347–364.
- [25] Piegorsch, W.W. (1993). Biometrical methods for testing dose effects of environmental stimuli in laboratory studies, *Environmetrics* **4**, 483–505.
- [26] Fisher, R.A. (1950). The significance of deviations from expectation in a Poisson series, *Biometrics* **6**, 17–24.
- [27] Fisher, R.A., Thornton, H.G. & MacKenzie, W.A. (1922). The accuracy of the plating method of estimating the density of bacterial populations, *Annals of Applied Biology* **9**, 325–359.
- [28] Neyman, J. & Scott, E.L. (1966). On the use of $C(\alpha)$ optimal tests of composite hypotheses, *Bulletin de l'Institut International de Statistique (Calcutta)* **41**, 477–497.
- [29] Collings, B.J. & Margolin, B.H. (1985). Testing goodness of fit for the Poisson assumption when observations are not identically distributed, *Journal of the American Statistical Association* **80**, 411–418.
- [30] Dean, C. & Lawless, J.F. (1989). Tests for detecting overdispersion in Poisson regression models, *Journal of the American Statistical Association* **84**, 467–472.
- [31] Kim, B.S. & Park, C. (1992). Some remarks on testing goodness of fit for the Poisson assumption, *Communications in Statistics: Theory and Methods* **21**, 979–995.
- [32] Xiang, L. & Lee, A.H. (2005). Sensitivity of test for overdispersion in Poisson regression, *Biometrical Journal* **47**, 167–176.
- [33] Bogen, K.T. (1994). Applicability of alternative models of revertant variance to Ames-test data for 121 mutagenic carcinogens, *Mutation Research* **322**, 265–273.
- [34] Hamada, C., Wada, T. & Sakamoto, Y. (1994). Statistical characterization of negative control data in the Ames Salmonella/microsome test, *Environmental Health Perspectives* **102**(Suppl. 1), 115–119.
- [35] Mahon, G.A.T., Middleton, B., Robinson, W.D., Green, M.H.L., Mitchell, I. & Tweats, D.J. (1989). Analysis of data from microbial count assays, in *Statistical Evaluation of Mutagenicity Test Data*, D.J. Kirkland, ed, Cambridge University Press, Cambridge, pp. 26–65.
- [36] Tarone, R.E. (1982). The use of historical control information in testing for a trend in Poisson means, *Biometrics* **38**, 457–462.
- [37] Clark, S.J. & Perry, J.N. (1989). Estimation of the negative binomial parameter κ by maximum quasi-likelihood, *Biometrics* **45**, 309–316.
- [38] Piegorsch, W.W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter, *Biometrics* **46**, 863–867.
- [39] van de Ven, R. (1993). Estimating the shape parameter for the negative binomial distribution, *Journal of Statistical Computation and Simulation* **46**, 111–123.
- [40] Saha, K. & Paul, S. (2005). Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter, *Biometrics* **61**, 179–185.
- [41] Astuti, E.T. & Yanagawa, T. (2002). Trend test for count data with extra-Poisson variability, *Biometrics* **58**, 398–402.
- [42] Schumacher, M. & Schmoor, C. (1991). Statistical analysis of the Ames assay, in *Statistics in Toxicology*, L. Hothorn, ed, Springer-Verlag, New York, Heidelberg, pp. 5–19.
- [43] Hothorn, L.A. (2004). A robust statistical procedure for evaluating genotoxicity data, *Environmetrics* **15**, 635–641.
- [44] Mack, G.A. & Wolfe, D.A. (1981). K -sample rank tests for umbrella alternatives, *Journal of the American Statistical Association* **76**, 175–181.
- [45] Simpson, D.G. & Margolin, B.H. (1990). Nonparametric testing for dose-response curves subject to downturns: asymptotic power considerations, *Annals of Statistics* **18**, 373–390.

- [46] Breslow, N.E. (1984). Extra-Poisson variability in log-linear models, *Applied Statistics* **33**, 38–44.
- [47] Leroux, B.G. & Krewski, D. (1993). AMESFIT: A microcomputer program for fitting linear-exponential dose-response models in the Ames Salmonella assay, *Environmental and Molecular Mutagenesis* **22**, 78–84.
- [48] Bernstein, L., Kaldor, J., McCann, J. & Pike, M.C. (1982). An empirical approach to the statistical analysis of mutagenesis data from the Salmonella test, *Mutation Research* **97**, 267–281.
- [49] Simpson, D.G. & Margolin, B.H. (1986). Recursive nonparametric testing for dose-response relationships subject to downturns at high doses, *Biometrika* **73**, 589–596.
- [50] Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**, 50–60.
- [51] Jonckheere, A.R. (1954). A distribution-free κ -sample test against ordered alternatives, *Biometrika* **41**, 133–145.
- [52] Simpson, D.G. & Dallal, G.E. (1989). BUMP: a FORTRAN program for identifying dose-response curves subject to downturns, *Computers and Biomedical Research* **22**, 36–43.
- [53] Cologne, J.B. & Breslow, N.E. (1994). Nonparametric regression analysis of data from the Ames mutagenicity assay. *Environmental Health Perspectives* **102**(Suppl. 1), 61–70.

Related Articles

Environmental Health Risk

Environmental Risks

WALTER W. PIEGORSCH

Statistics for Environmental Teratogenesis

Background

Teratogenesis is a process through which embryo/fetal development is altered and an abnormality is induced. The abnormalities can arise from genetic or environmental factors. Any chemical or physical agents that can cause an abnormality in the embryo or fetus are called *teratogens*. Teratology is the study of adverse effects of environment on the developing system. The term *teratogenesis* specifically refers to the development of abnormal cells during embryo/fetal growth. In descriptions of the harmful effects of chemical or physical agents on the developing system, in a broader sense, developmental toxicology is the study of adverse effects on the developing organism that may result from exposure prior to conception (either parent), during prenatal development, or postnatal development to the time of sexual maturation. Reproductive toxicology is the study of adverse effects on the reproductive system of males and females, as well as on postnatal maturation and reproductive capacity of offspring, and the cumulative effects on generations. Adverse developmental and reproductive effects include effects on male and female fecundity, spontaneous abortion, infant and child death, congenital malformations, growth retardation, and mental retardation. Because of ethical and practical considerations, the major source of information for studying potential developmental and reproductive effects of chemicals on humans is generally the laboratory animal experiments (*see Cancer Risk Evaluation from Animal Studies*). Mice, rats, and rabbits are the commonly used species in developmental and reproductive studies.

Animal developmental and reproductive studies have been divided into three segments. These are referred to as *Segment I* (fertility and general reproductive studies), *Segment II* (teratology or developmental toxicity studies), and *Segment III* (perinatal and postnatal studies) [1]. The three segments of investigations are generally required by regulatory agencies in the preclinical testing of new drugs or environmental agents such as pesticides and

industrial chemicals. Testing protocols for pharmaceutical chemicals and nonpharmaceutical chemicals are similar; however, the data for nonpharmaceutical chemicals would be used to establish an acceptable level of exposure using quantitative risk assessment. Each animal is monitored throughout the experiment to detect toxic and pharmacologic effects of the test compounds. In a Segment II study, the pregnant dams are normally sacrificed just prior to term, and fetuses are then examined to assess developmental effects.

A developmental experiment typically consists of an untreated control and three dosage groups. The highest dose is chosen to produce minimal maternal toxicity, ranging from marginal body weight (BW) reduction to not more than 10% mortality. The low dose should be a no-observed-adverse-effect level (NOAEL) for both maternal and developmental toxicity. The mid-level dose is usually halfway between the low and high doses. A fourth dosage group may be added to avoid excessive dosage intervals and to assess dose-response relationships for quantitative risk assessment. The dose should ideally be the amount of active substance at the affected tissue site. The estimate of the internal tissue-site dose requires data from toxicokinetic and/or toxicodynamic experiments. In the absence of this information, the amount of the internal dose is assumed to be proportional to the administered or exposure dose. Typically, dosage is measured in milligrams per kilogram body weight per day. Animals can be assigned to dosage groups totally randomly or by stratified BW (but randomly within BW classes).

The developmental endpoints can generally be divided into two categories: parental and embryonic/fetal endpoints. Since the test chemical is administered to the adult animal, the adverse effect occurs in the female that receives the chemical or that is mated to a male that receives the chemical and the treatment affects the embryos/fetuses indirectly *via* the dam. Thus, the development of an individual embryo/fetus in a dam is not an independent process. The embryo/fetal responses from the same dam are expected to be more alike than responses from different dams. This phenomenon is referred to as the *litter effect*. The experimental unit in the analysis should be the litter with the embryo/fetal responses representing multiple observations from a single experimental unit. Failure to account for the intralitter correlations by using each fetus as the experimental unit can

underestimate the variance of an effect and reduce the validity of the analysis.

Statistical analyses of various developmental endpoints have been concerned with two aspects: qualitative testing for detecting adverse effects and quantitative estimation for risk assessment. The qualitative testing is to determine if there is a statistically significant difference between the control and dosed groups. The quantitative estimation is to determine a safe level of human exposure from an assessment of the dose–response relationship. The quantitative risk estimation is typically used for environmental or workplace exposures and does not apply to pharmaceutical drugs.

Dose–Response Experiments

Dose–response models are usually formulated to establish a mathematical relationship between the observed response and the dosage level (*see Dose–Response Analysis*). Risk assessment aims at deriving an acceptable level corresponding to little or no risk or estimating a risk at a given exposure level from a fitted dose–response function. The preliminary step is to generate dose–response data, typically from experiments in laboratory animals.

Consider an experiment of g groups, a control ($d_1 = 0$) and $g - 1$ dose groups ($d_2, \dots, d_g$). Assume that the i th group contains n_i female animals. Let y_{ijk} be the response from the k th fetus out of n_{ij} examined or tested for a particular developmental outcome, $1 \leq i \leq g$, $1 \leq j \leq n_i$, and $1 \leq k \leq n_{ij}$. Note that the y_{ijk} may be an indicator variable representing the presence or absence of a particular effect or a continuous variable representing a magnitude of the outcome. In dose–response studies, the mean parameters are of interest in terms of a marginal model. The mean parameters in a marginal model have a population-averaged interpretation in the sense that the effect of the dose (and other covariates) is averaged across the animals in the same dose group. It is commonly assumed that the mean responses within a litter are exchangeable [2] having a common mean $E(y_{ijk}) = \mu_{ij}$ and a common correlation $\text{Corr}(y_{ijk}, y_{ijk'}) = \phi_i$. The parameter ϕ_i is the intralitter correlation, and it may vary with dose. The mean parameter is related to the dosage level, dose–response function, and a set of covariates $\mathbf{z} = (1, d, z_3, \dots, z_p)$ (e.g., litter size or sex).

The first step in quantitative risk assessment is to select a dose–response function to fit the animal dose–response data. The dose–response function is governed by the nature of the measurement that represents the developmental endpoint of interest. Measurements for developmental toxicity endpoints can be divided into four types: (a) continuous (e.g., fetal body or organ weight), (b) dichotomous (e.g., presence or absence of a malformation), (c) ordered categorical (e.g., a histology score that ranges from 1 – normal to 5 – extremely abnormal), and (d) count (e.g., number of implantations). Quantitative risk assessment has focused on the continuous and dichotomous response endpoints.

Dose–Response Models for Continuous Data

For continuous response data such as a fetal BW, the fetus response y_{ijk} is expressed as a mixed-effects model with two sources of variations: the between litters γ_{ij} and within litters e_{ijk} ,

$$y_{ijk} = \mu_{ij} + \gamma_{ij} + e_{ijk} \quad (1)$$

The mean parameter μ_{ij} represents the fixed effects component. The random components γ_{ij} and e_{ijk} are independently normally distributed with $E(\gamma_{ij}) = 0$, $\text{Var}(\gamma_{ij}) = \sigma_a^2$, and $E(e_{ijk}) = 0$, $\text{Var}(e_{ijk}) = \sigma^2$. Thus, the mean and variance of y_{ijk} are $E(y_{ijk}) = \mu_{ij}$ and $\text{Var}(y_{ijk}) = \sigma_a^2 + \sigma^2$. The intralitter correlation between y_{ijk} and $y_{ijk'}$, for $k \neq k'$ is $\phi = \sigma_a^2 / (\sigma^2 + \sigma_a^2)$.

The mean parameter μ_{ij} can be modeled as a linear function of the dose and covariates, for example,

$$\mu_{ij}(d_i | n_{ij}) = \beta_0 + \beta_1 d_i + \beta_2 d_i^2 \quad (2)$$

The μ_{ij} may also be modeled by nonlinear function, such as the Hill model:

$$\mu_{ij}(d_i) = \beta_0 + \beta_1 \frac{d_i^n}{k^n + d_i^n}, \quad n \geq 1 \quad (3)$$

The maximum-likelihood estimation (MLE) of the parameters of the linear mixed-effects model (2) and the nonlinear mixed-effects model (3) are estimated using reweighted least-squares methods iteratively [3–5].

In quantitative risk assessment, dose–response models for continuous endpoints are often based on

the per-litter response. The fetal toxicity endpoints are analyzed in terms of the average within each litter $y_{ij} = \sum_k y_{ijk}/n_{ij}$. Since litter is the experimental unit, the y_{ij} represents a summarized measure of the fetal responses of the experimental unit. Note that y_{ij} can also be the measurement of the maternal endpoints (e.g., maternal body or organ weight). The y_{ij} has a normal distribution with the mean μ_{ij} . The MLEs of the dose–response coefficients are estimated by the least-squares or nonlinear least-squares method.

Dose–Response Models for Dichotomous Data

Let y_{ijk} indicate the presence or absence of a response. Assume that the mean and variance of y_{ijk} are $E(y_{ijk}) = \mu_{ij}$, and $\text{Var}(y_{ijk}) = \mu_{ij}(1 - \mu_{ij})$ and the correlation is $\text{Corr}(y_{ijk}, y_{ijk'}) = \phi_i$, where $k, k' = 1, \dots, n_{ij}$, and $k \neq k'$. The parameter μ_{ij} is the probability of a developmental effect for the j th litter in the i th group, and ϕ_i is the intralitter correlation coefficient. Parametric multivariate binary models have not been widely used in practice because their MLEs are difficult to compute.

A traditional modeling approach, in terms of the total number of responses, is based on the assumption of exchangeability within a litter. The total number of responses in a litter $y_{ij} = \sum_k y_{ijk}$ has the mean $E(y_{ij}) = n_{ij}\mu_{ij}$ and variance $\text{Var}(y_{ij}) = n_{ij}\mu_{ij}(1 - \mu_{ij})[1 + \phi_i(n_{ij} - 1)]$. The distribution of y_{ij} is a “generalized” binomial. The y_{ij} has an extra-binomial distribution if the intralitter correlation ϕ_i among y_{ijk} ’s is positive, and y_{ij} has a subbinomial distribution if ϕ_i is negative. If $\phi_i = 0$, then all y_{ijk} ’s are independent binary random variables and y_{ij} is a binomial.

The intralitter correlation coefficient is positive ($\phi_i > 0$) for most developmental toxicity endpoints. The beta-binomial model is the distribution commonly used for modeling data with extra-binomial variation. Under the reparameterization $\mu_i = a_i/(a_i + b_i)$ and $\phi_i = (a_i + b_i + 1)^{-1}$, the beta-binomial distribution is

$$P(y_{ij}) = \binom{n_{ij}}{y_{ij}} \frac{B(a_i + y_{ij}, b_i + n_{ij} - y_{ij})}{B(a_i, b_i)} \quad (4)$$

where $B(a_i, b_i) = \Gamma(a_i)\Gamma(b_i)/\Gamma(a_i + b_i)$ and $\Gamma(\cdot)$ is the gamma function, $a_i > 0$, and $b_i > 0$.

For dichotomous responses, the dose–response function $\mu_{ij} = P(d_i)$, relates the probability of

response to the dose. The mean function is often modeled by a logit link function, $\log[\mu_{ij}/(1 - \mu_{ij})]$, in which case the simple linear logistic Dose–response model is

$$P(d_i) = \frac{\exp(\beta_0 + \beta_1 d_i)}{1 + \exp(\beta_0 + \beta_1 d_i)} \quad (5)$$

The logit model (see **Logistic Regression**) has the advantage that the coefficients have the interpretation of being log odds ratios. Chen and Kodell [6] proposed a Weibull dose–response model:

$$P(d_i) = 1 - \exp(-\beta_0 - \beta_1 d_i^w) \quad (6)$$

The power parameter on dose, $w > 0$, increases the flexibility in the shape of the dose–response function. The dose–response model $P(d)$ can include any additional explanatory variables (e.g., fetal weight, litter size). For example, Kodell *et al.* [7] generalized the Weibull model (6) by incorporating the litter size:

$$P(d_i) = 1 - \exp[-\beta_0 - b_1 n_{ij} - (\beta_1 + b_2 n_{ij})d_i^w] \quad (7)$$

For the dichotomous data, the intralitter correlations typically are either modeled as constant across groups or modeled separately for each group. When intralitter correlations are dose dependent, the use of a common intracluster correlation in the beta-binomial model fitting might bias the estimation of dose–response coefficients [8]. On the other hand, the estimates obtained from the fitting of separate correlation parameters might not be reliable [8]. Bowman *et al.* [9] proposed using a logistic dose–response model to fit the intralitter correlations.

The quasi-likelihood approach [10] is an alternative approach to the beta-binomial model. In the quasilikelihood approach, only assumptions regarding the mean and variance are required: $E(y_{ij}|n_{ij}) = n_{ij}\mu_i$ and $\text{Var}(y_{ij}|n_{ij}) = n_{ij}\mu_i(1 - \mu_i)[\phi(n_{ij} - 1) + 1]$. The coefficient β ’s of dose–response functions (5), (6), or (7) can be obtained by solving the quasilikelihood estimating (score) equations. The intralitter coefficient ϕ is calculated by equating with the mean of Pearson chi-square statistics. Estimation details are given by Williams [11]. The estimates of the dose–response coefficient by the quasilikelihood method are often more stable than the MLE estimates obtained by beta-binomial model.

Quantitative Risk Assessment

A fundamental goal of risk assessment is to determine safe levels of human exposure to a chemical. Such estimated levels are commonly referred to as *reference doses* (RfDs). The RfDs are presumed to cause no or negligible health risks in humans. Recently, United States Environmental Protection Agency (USEPA) [12] proposed using the benchmark dose (BMD) approach (*see Benchmark Dose Estimation*) to derive RfD for health risk assessment. The BMD refers to the dose corresponding to a prespecified benchmark response (BMR), an additional adverse response over the background response. The BMD is estimated by fitting a dose–response model to the dose–response data. Because relatively small numbers of animals are exposed to only a few dose levels and from the viewpoint of public health protection, the dose of interest is the lower confidence limit (e.g., 95%) on the BMD, denoted as BMDL. With the BMDL as a starting point-of-departure, uncertainty factors (UFs) are then employed to establish an RfD. The RfD is calculated by dividing BMDL by several UFs: $\text{RfD} = \text{BMDL}/(\text{UF}_1 \times \text{UF}_2 \times \text{UF}_3 \dots)$. The UFs can include interspecies, intraspecies, subchronic-to-chronic, lowest-observed-adverse-effect-level to no-observed-adverse-effect-level (LOAEL-to-NOAEL), etc. For example, using an UF of 1000 ($10 \times 10 \times 10$), the corresponding RfD is $\text{BMDL}/1000$.

Calculations of the MLE of BMD and BMDL are available from the USEPA Benchmark Software, which can be downloaded free of charge from the EPA home page (<http://www.epa.gov>). The MLEs of the coefficient of the dose–response models (2), (3), (5), (6), and (7) and the BMD and BMDL can be obtained using the EPA Benchmark Software. The models (2) and (3) are computed using the litter-based approach, whereas the models (5), (6), and (7) are calculated using the beta-binomial model. The model (7) is named the National Center for Toxicological Research *NCTR model* in the EPA software. Quantitative risk assessment for developmental effects *via* the BMD approach are illustrated below. The EPA Benchmark Software was used in all calculations.

Example 1 For dichotomous endpoints, the dose–response function $P(d)$ is the probability of occurrence of the adverse effect associated with the dose level d . The quantity of interest from the risk

assessment viewpoint is $\pi(d) = P(d) - P(0)$, which measures the excess risk (above the background risk) due to the added dose. For developmental effects, the BMDL corresponding to BMR value between 1 and 10% is used as a starting point-of-departure. The BMR of 10% is the default value recommended by the USEPA.

Example 1 is a study of teratogenic effects from exposure to diethylhexyl phthalate (DEHP). The experiment consisted of five dose groups of CD-1 mice with levels of 0, 44, 91, 191, and 292 mg/(kg BW) administered to pregnant females. This study showed that the DEHP exposure increased the incidence of malformation in fetuses. The complete data set was listed in Kodell *et al.* [7]. Assume the Weibull dose–response model (6) for the malformation effect. The BMD corresponding to 10% excess risk is the dose that satisfies the equation

$$0.10 = \exp(-\beta_0) - \exp(-\beta_0 - \beta_1 d_i^w) \quad (8)$$

The BMD estimate is 98 mg/(kg BW) and the BMDL estimate is 83 mg/(kg BW). If the UF of 1000 is applied, $\text{UF}_1 = 10$ for interspecies, $\text{UF}_2 = 10$ for intraspecies, and $\text{UF}_3 = 10$ for the BMR of 10% (note that the 10% excess risk may be regarded as LOAEL), then the RfD is 0.083 mg/(kg BW).

Example 2 For continuous responses, the BMR represents a change in the means response relative to the background. In the continuous response case, selection of the BMR requires judgment of what level of response should be regarded as adverse. A change in the mean of the control group equal to 1.1 standard deviation is equivalent to defining BMR of 10% for 1% background [13].

Table 1 Body weight gain in female rat from a toxicological experiment

Dose (mg/(kg BW))	<i>n</i>	Mean	Standard deviation
0	10	10.40	4.25
100	10	10.33	2.67
200	10	8.48	1.88
500	10	4.43	3.29
750	10	−4.50	2.93

Example 2 is BW changes in female rats (Table 1). These data represent typical per-litter response. The Hill equation was fit to the data. The BMR is selected

based on the 1.1 SD from the control mean (by selecting the option: BMR Type = Std. Dev and BMR = 1 in the EPA Benchmark Software). The BMD estimate is 377 mg/(kg BW) and the BMDL estimate is 275 mg/(kg BW). If the UF of 100 for interspecies and intraspecies uncertainties is applied, then the corresponding RfD is $275/100 = 2.75$ mg/kg.

Summary Comments

Developmental endpoints consist of both continuous and discrete responses data. The analysis should account for the correlations induced by the clustering effects within litters. For continuous data, the analysis of fetal responses is customarily in terms of the litter mean. However, this approach does account for differences in litter sizes. The mixed-effects model is a general parametric approach to modeling continuous outcomes. For dichotomous response, the fetal responses is summarized in terms of sum of correlated binary variables resulting in a generalized binomial distribution. The distribution for the common developmental endpoints, such as the number of malformations per litter, can be modeled by a beta binomial.

Nonclinical toxicological experiments are generally exploratory in nature with the goal of determining potential adverse effects of a chemical. The probability of detecting a statistically significant effect increases with the increased magnitude of effect and increased sample size. US regulatory guidelines generally recommend about 20 pregnant rodents and 15 nonrodent animals per dosage group. The International Conference Harmonisation guideline recommends use of 16–20 pregnant animals. However, it should be emphasized that sample size determination is preferable to arbitrary requirements for testing 16 or 20 animals as indicated in various guidelines. It is customary to specify a range of the magnitudes of the effect of interest; then an optimal sample size can be calculated to provide a reasonable power of detecting the effect. Sample-size calculations and references for various tests are provided in Desu and Raghavarao [14]. The statistical software PASS can be used for power calculation analysis and sample-size calculation (<http://www.ncss.com>).

The standard approach for assessment of developmental risks of a compound has been based on the analysis of each developmental endpoint separately.

It has been suggested that the developmental toxicity outcomes (i.e., death/resorption, malformation, growth retardation, etc.) may represent different degrees of responses to a toxic insult and occur in a dose-related manner [15]. These developmental outcomes are likely to be correlated. Therefore, a joint analysis of multiple developmental outcomes can have some advantages: it can increase the power of detecting effects if the multiple outcomes are manifestations of some common biological effects and it allows investigations of associations among the multiple outcomes if they are the results of different biological mechanisms. Various multivariate models have been developed for simultaneous analysis of multiple endpoints [16, 17].

References

- [1] Tyl, R.W. & Marr, M.C. (1996). Developmental toxicity testing—methodology, in *Handbook of Developmental Toxicology*, R.D. Hood, ed, CRC Press, New York, pp. 175–225.
- [2] Bowman, D. & George, E.O. (1995). A saturated model for analysis exchangeable binary data: applications to clinical and developmental toxicity studies, *Journal of the American Statistical Association* **90**, 871–879.
- [3] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association* **72**, 320–340.
- [4] Lindstrom, M.J. & Bates, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data, *Journal of the American Statistical Association* **83**, 1014–1022.
- [5] Lindstrom, M.J. & Bates, D.M. (1990). Nonlinear mixed-effects models for repeated measures data, *Biometrics* **46**, 673–687.
- [6] Chen, J.J. & Kodell, R.L. (1989). Quantitative risk assessment for teratological effects, *Journal of the American Statistical Association* **84**, 966–971.
- [7] Kodell, R.L., Howe, R.B., Chen, J.J. & Gaylor, D.W. (1991). Mathematical modeling of reproductive and developmental toxic effects for quantitative risk assessment, *Risk Analysis* **11**, 583–590.
- [8] Williams, D.A. (1988). Estimation bias using beta-binomial distribution in teratology, *Biometrics* **44**, 305–308.
- [9] Bowman, D., Chen, J.J. & George, E.O. (1995). Estimating variance functions in developmental toxicity studies, *Biometrics* **51**, 1174–1176.
- [10] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Model*, 2nd Edition, Chapman Hall, London.
- [11] Williams, D.A. (1982). Extra-binomial variation in logistic linear models, *Applied Statistics* **31**, 144–148.

- [12] U.S. Environmental Protection Agency (2001). *Help Manual for Benchmark Dose Software*, EPA 600/R-00/014F. National Center for Environmental Assessment, Cincinnati Office. Cincinnati, Available: <http://www.epa.gov/ncea/bmds>.
- [13] Crump, K.S. (1995). Calculation of benchmark doses for continuous data, *Risk Analysis* **15**, 79–89.
- [14] Desu, M.M. & Raghavarao, D. (1990). *Sample Size Methodology*, Academic Press, New York.
- [15] Chen, J.J. & Gaylor, D.W. (1992). Correlations of developmental end points observed after 2,4,5-trichlorophenoxyacetic acid exposure in mice, *Teratology* **45**, 241–246.
- [16] Chen, J.J. & Li, L.-A. (1994). Dose-response modeling of trinomial responses from developmental experiments, *Statistica Sinica* **4**, 265–274.
- [17] Catalano, P.J. & Ryan, L.M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes, *Journal of the American Statistical Association* **87**, 651–658.

Further Reading

Kupper, L.L., Portier, C., Hogan, M.D. & Yamamoto, E. (1986). The impact of litter effects on dose-response modeling in teratology, *Biometrics* **42**, 85–89.

JAMES J. CHEN

Statistics for Environmental Toxicity

For the purpose of this article, we define “environmental toxicity” as the study of toxic hazards expressed in any environment and at any level of biological organization ranging from molecular to ecosystem level. We will use “environmental toxicology” synonymously with “environmental toxicity” here. In addition, when we mention statistical methods arising in environmental toxicity assessments, we will provide a citation to a recent environmental toxicology study where these techniques are employed. Frequently, the source of the hazard is of human origin. We include the study of adverse impacts on humans in this definition, considering humans as part of natural food chains and ecosystems. Environmental toxicity or toxicology assessments are conducted (often by statutory mandate) when evaluating the safety of materials or stressors dispersed into the environment from the agrochemical, consumer product, defense, energy, and transportation industries; and from domestic or municipal effluents and runoffs.

Where is the Environmental Impact Observed?

Environmental impacts and the statistics used to model and evaluate these impacts can be considered over multiple levels of biological organization. From the lowest level to the highest level of organization, these include molecular/genomic, tissue/organ, whole organism, population, community, ecosystem, and landscape-scale responses. Environmental toxicity research originally focused on the organismal level, but has simultaneously extended outwards in both directions to lower level responses within organisms (e.g., genomics) and higher level responses across communities of organisms and into landscape-scale responses. We consider, in turn, each level of environmental toxicological responses along with some consideration of the statistical issues that surface with each of these response levels.

Molecular/Genomic Responses

Desire to understand the mode of toxicity has led toxicologists to investigate tissue-specific, cellular, and molecular responses to hazards. In the context

of risk assessment, the investigation of molecular/genomic responses can be used to help with the identification of a compound as a potential hazard (*see* **Environmental Hazard**). Toxic effects at the molecular level include DNA modification and enzyme dysfunction. Cellular and tissue toxicity includes necrosis, cell proliferation, mutation, and cancer. The interest in greater mechanistic understanding has led to greater use of “-omics” techniques (genomics, metabonomics, proteomics, and toxicogenomics). These tools can provide data on the expression of hundreds to thousands of genes for each sample that is processed.

The statistical methods encountered here and in later sections mirror the data that are produced. For DNA modification and genotoxicity responses, toxicologists are often interested in testing for increasing trends in these endpoints. One confounder is that toxin levels that are genotoxic might also be near toxin levels that would be lethal. Thus, one problem encountered in this area is that of testing for increasing dose–response patterns with possible downturns at higher doses. The “-omics” data have been described as “large p, small n” in reflection of the situation that these methods can yield thousands of measurements on tens to hundreds of observations. Thus, toxicologists might be interested in identifying which, if any, gene(s) appears to differentially express given exposure to a toxin. This leads to statistical considerations of *multiplicity* because of testing a large number of candidate genes and descriptions of *false discovery rates* to characterize the risk that genes are incorrectly identified as differentially expressed.

Organ/Tissue Responses

Organisms are often exposed to toxins through the environment (e.g., fish or sea mammals *via* contaminated waters). The measurement of environmental toxin concentrations involves analytic chemistry to quantify the levels and statistical survey sampling to determine locations to select the samples that are evaluated. Analytic instruments have a limit of detection that may result in a number of observations occurring below this detection limit (*see* **Detection Limits**) [1]. One question of interest to environmental toxicologists is how much of this toxin is delivered to tissues that are most sensitive to it. The toxin dose (or some metabolite of the toxin) at the tissue of interest can be thought of as more closely related to

any adverse response that might be observed. Toxicokinetic models in which compartments correspond to particular organs/tissues and transitions between tissues are governed by blood flows and partition coefficients are often employed [2, 3]. The transitions between these models often involve the solution of a system of differential equations. Related kinetic models look at energy transfer in ecosystems [4].

While dichotomous responses (e.g., presence or absence of data) are commonly measured in toxicity studies, other response scales are also used. For example, the severity of tissue damage that might be associated with toxin exposure could be evaluated using a scale with levels “none”, “some”, or “severe”. This ordered severity scale might be analyzed by first dichotomizing the response scale (e.g., comparing “some” or “severe” to “none”) and applying logistic regression techniques or directly analyzing the ordinal response scale using *ordinal regression methods* [5].

Individual Organism Responses

While molecular and tissue level responses can be subtle to observe, toxic responses at the level of an individual organism are often directly observed. One common classification is to describe individual level effects as sublethal (e.g., decrements in growth, development, and reproduction) or lethal effects. The most common statistical methods encountered at this response level include anova methods and regression methods. Testing in the context of analysis of variance models might be used to identify the highest concentration group with a mean response that does not differ statistically from the mean response under control conditions, so-called no observed effect concentrations (NOECs) or the lowest concentration group with a mean response that differs statistically from the mean response observed under control conditions – lowest observed effect concentrations (LOECs). Usually, some multiple comparison procedures such as Dunnett’s procedure are used to identify these concentrations. In the case of dichotomous responses (e.g., survival) or count data (e.g., number of young produced), some transformation of the response might be employed before analyzing the data with normal-theory methods (e.g., arcsine square root for proportions and square root for counts). NOEC/LOEC calculations are still frequently used in environmental toxicology and may serve as the building blocks of other toxicity summaries such as

hazard quotients. A number of statistical critiques of these endpoints have been raised, including the restriction that these endpoints can only be one of the tested concentrations and these values will get smaller in experiments with less variability. Effective concentration estimation and benchmark concentrations (*see Benchmark Dose Estimation*) are suggested as alternatives to NOEC/LOEC calculations [6]. These alternatives involve fitting some regression model to predict the individual organism response as a function of toxin concentration. While normalizing transformations of the responses followed by multiple regression are possible, modeling responses on their observed measurement scale is more common. Generalized linear models are frequently used as a general framework for analyzing individual organism responses [7, 8]. For example, logistic regression (*see Logistic Regression*) or probit regression is usually employed for modeling dichotomous responses [9]. After fitting these models, the concentration that is associated with some specified decrement in response is estimated. A common value that is estimated by inverting these regression models is the concentration associated with 50% mortality, the median lethal concentration or LC₅₀. When analyzing nonlethal responses, the impact level is often described as the effective concentration associated with p% impact or EC_p. The LC_p and EC_p can be considered potency estimates that are derived from inverting a regression model to estimate the concentration associated with some specified impact level. In environmental risk assessment, the potential risk of adverse response at a site might be investigated by constructing a hazard quotient = (environmental exposure concentration)/(toxicity reference value) where the toxicity reference value is a LC_p, EC_p, NOEC, or LOEC [10]. Hazard quotients are deterministic values that have been used to portray potential risk. These values are not actual estimates of risk and do not account for uncertainties in the endpoint estimates. Probabilistic risk assessments use probability distributions of potency estimates and exposure scenarios, in combination with Monte Carlo simulation to derive quantitative estimates of risk along with estimates of variability and uncertainty in these values [11, 12].

Population Responses

Population level effects can be summarized in a variety of terms including mean changes in various

population level traits. A population can be conceptualized as a collection of individuals at various life stages and/or ages. A population can be altered by a number of factors including immigration of new individuals, reduction in reproduction rates, and mortality of early age/stage individuals. The impact of an environmental toxin can be observed in a variety of population level changes. For example, a toxin might reduce fecundity or it might alter the transition of organisms from one life stage to another (e.g., juvenile to adult) or, most dramatically, it might increase the probability of species extinction. Population level studies might investigate life history [13] as well as population growth and recovery [14]. Statistical models for analyzing population impact range from logistic methods for survival and breeding condition to investigating systems of age-structure/life stage stages using Leslie matrices.

Community and Ecosystem Responses

Better appreciation of ecological issues has pushed environmental toxicology in the direction of deeper understanding of species interactions at the community and ecosystem level. The effects of these levels can be evaluated along a variety of different dimensions. What species are most sensitive to a toxin? Does a contaminant biomagnify across trophic levels, i.e., does a toxin increase in predators higher on the food chain because of consuming a persistent toxin in prey? Can a species be determined that would be a sentinel for evaluating whether a hazard has been used? Does exposure to a toxin alter the diversity, richness, or structure of an ecological community? How can distributions of species sensitivity to a toxin be used to set an exposure level that would be protective of most species in an ecosystem? How do environmental pollutants affect biogeochemical cycles and what impact does this have on populations or communities?

Community metrics provide a means of evaluating complex environmental systems. For example, the index of biotic integrity (IBI) is a scale that uses the presence or absence of certain key fish species to “grade” the quality of a stream [15, 16]. The IBI values are then grouped into categories of impact using stream systems with known impact to help set quality categories. This concept of community indexes of biotic integrity has been extended to many other assemblages, including invertebrates [17], plankton [18, 19], birds [20], and ecosystems

[21, 22]. Statistical methods to formalize this classification might include some cluster analytic methods [17]. Another important community response idea that is often used in environmental toxicology, especially as it relates to environmental risk assessment, is the species-sensitivity distribution, which is the cumulative distribution function (CDF) for endpoints derived from concentration–response studies (e.g., EC_p and NOEC). The fifth percentile of this distribution, the HC₅, is frequently estimated and used to represent a concentration that would materially impact 5% of the species while not seriously impacting the other 95% [23, 24]. The estimation of the confidence intervals associated with HC₅ has been a statistical problem addressed by a number of techniques including bootstrapping [25].

Landscape-Scale Responses

The display and analysis of the impact of environmental toxins across a spatial region has led to an increase in the use of spatial statistical methods [26] for modeling and prediction and geographic information systems (GIS) [27, 28] for display. Interestingly, these data share some of the features of the “-omics” data. In other words, a large number of variables are measured on a small number of observations. Finally, Bayesian statistical methods (*see Bayesian Statistics in Quantitative Risk Assessment*) are becoming more commonly encountered in the field of environmental toxicity including assessments of uncertainty and variability in a hierarchical model for a bioaccumulation model [29] and examining latent variables and spatio-temporal trends [30].

Other Issues

There are a number of statistical considerations that are encountered at a number of the ecological measurement scales. A form of censoring (*see Censoring*) occurs when responses are not observed by the scheduled end of observation in a study. This situation leads to the use of survival analytic methods [31]. The analysis of environmental exposures to mixtures of toxins is an area of active investigation [32]. Questions of dose-additivity or response-additivity can be used to determine toxicity units or toxic-equivalence factors that might be part of an environmental risk assessment.

References

- [1] Sinha, P., Lambert, M.B. & Trumbull, V.L. (2006). Evaluation of statistical methods for left-censored environmental data with nonuniform detection limits, *Environmental Toxicology and Chemistry* **25**, 2533–2540.
- [2] Andersen, M.E., Krewski, D. & Withey, J.R. (1993). Physiological pharmacokinetics and cancer risk assessment, *Cancer Letters* **69**, 1–14.
- [3] Bailer, A.J. & Dankovic, D. (1997). An introduction to the use of physiologically based pharmacokinetic models in risk assessment, *Statistical Methods in Medical Research* **6**, 341–358.
- [4] Pery, A.R.R., Flammarion, P., Vollat, B., Bedaux, J.J.M., Kooijman, S.A.L.M. & Garric, J. (2002). Using a biology-based model (DEBTOX) to analyze bioassays in ecotoxicology: opportunities and recommendations, *Environmental Toxicology and Chemistry* **21**, 459–465.
- [5] Chuang-Stein, C. & Agresti, A. (1997). Tutorial in biostatistics a review of tests for detecting a monotone dose-response relationship with ordinal response data, *Statistics in Medicine* **16**, 2599–2618.
- [6] Crump, K. (1984). A new method for determining allowable daily intake, *Toxicological Sciences* **4**, 854–871.
- [7] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition, Chapman & Hall, London.
- [8] Dobson, A. (2002). *Introduction to Generalized Linear Models*, 2nd Edition, Chapman & Hall/CRC, Boca Raton.
- [9] Collett, D. (2003). *Modelling Binary Data*, 2nd Edition, Chapman & Hall/CRC, Boca Raton.
- [10] Han, G.H., Hur, H.G. & Kim, S.D. (2006). Ecotoxicological risk of pharmaceuticals from wastewater treatment plants in Korea: occurrence and toxicity to *Daphnia magna*, *Environmental Toxicology and Chemistry* **25**, 265–271.
- [11] Kentel, E. & Aral, M. (2005). 2D Monte Carlo versus 2D fuzzy Monte Carlo health risk assessment, *Stochastic Environmental Research and Risk Assessment* **19**, 86–96.
- [12] Sander, P., Bergback, B. & Oberg, T. (2006). Uncertain numbers and uncertainty in the selection of input distributions – consequences for a probabilistic risk assessment of contaminated land, *Risk Analysis* **26**, 1363–1375.
- [13] Spromberg, J.A. & Birge, W.J. (2005). Population survivorship index for fish and amphibians: application to criterion development and risk assessment, *Environmental Toxicology and Chemistry* **24**, 1541–1547.
- [14] Barnhouse, L.W. (2004). Quantifying population recovery rates for ecological risk assessment, *Environmental Toxicology and Chemistry* **23**, 500–508.
- [15] Karr, J.R. (1981). Assessment of biotic integrity using fish communities, *Fisheries* **6**, 21–27.
- [16] Weigel, B.M., Lyons, J. & Rasmussen, P. (2006). Fish assemblages and biotic integrity of a highly modified floodplain river, the upper Mississippi, and a large, relatively unimpacted tributary, the lower Wisconsin, *River Research and Applications* **22**, 923–936.
- [17] Bressler, D.W., Stribling, J., Paul, M. & Hicks, M. (2006). Stressor tolerance values for benthic macroinvertebrates in Mississippi, *Hydrobiologia* **573**, 155–172.
- [18] Lacouture, R.V., Johnson, J., Buchanan, C. & Marshall, H. (2006). Phytoplankton index of biotic integrity for Chesapeake Bay and its tidal tributaries, *Estuaries and Coasts* **29**, 598–616.
- [19] Carpenter, K.E., Johnson, J. & Buchanan, C. (2006). An index of biotic integrity based on the summer polyhaline zooplankton community of the Chesapeake Bay, *Marine Environmental Research* **62**, 165–180.
- [20] Bryce, S.A. (2006). Development of a bird integrity index: measuring avian response to disturbance in the Blue Mountains of Oregon, USA, *Environmental Management* **38**, 470–486.
- [21] Mack, J.J. (2006). Landscape as a predictor of wetland condition: an evaluation of the Landscape Development Index (LDI) with a large reference wetland dataset from Ohio, *Environmental Modeling and Assessment* **120**, 221–241.
- [22] Balthis, W., Hyland, J. & Bearden, D. (2006). Ecosystem responses to extreme natural events: impacts of three sequential hurricanes in fall 1999 on sediment quality and condition of benthic fauna in the Neuse River Estuary, North Carolina, *Environmental Monitoring and Assessment* **119**, 367–389.
- [23] Duboudin, C., Ciffroy, P. & Magaud, H. (2004). Effects of data manipulation and statistical methods on species sensitivity distributions, *Environmental Toxicology and Chemistry* **23**, 489–499.
- [24] Frampton, G., Jansch, S., Scott-Fordsmand, J., Rombke, J. & Van den Brink, P. (2006). Effects of pesticides on soil invertebrates in laboratory studies: a review and analysis using species sensitivity distributions, *Environmental Toxicology and Chemistry* **25**, 2480–2489.
- [25] Grist, E., Leung, K., Wheeler, J. & Crane, M. (2002). Better bootstrap estimation of hazardous concentration thresholds for aquatic assemblages, *Environmental Toxicology and Chemistry* **21**, 1515–1524.
- [26] Micheletti, C., Critto, A., Carlon, C. & Marcomini, A. (2004). Ecological risk assessment of persistent toxic substances for the clam *Tapes Philipinarum* in the lagoon of Venice, Italy, *Environmental Toxicology and Chemistry* **23**, 1575–1582.
- [27] Pfleeger, T.G., Olszyk, D., Burdick, C.A., King, G., Kern, J. & Fletcher, J. (2006). Using a geographic information system to identify areas with potential for off-target pesticide exposure, *Environmental Toxicology and Chemistry* **25**, 2250–2259.
- [28] Kapo, K.E. & Burton, G.A. (2006). A geographic information systems-based, weights-of-evidence approach for diagnosing aquatic ecosystem impairment, *Environmental Toxicology and Chemistry* **25**, 2237–2249.
- [29] Lin, H., Berzins, D.W., Myers, L., George, W.J., Abdelghani, A. & Watanabe, K.H. (2004). A Bayesian

approach to parameter estimation for a crayfish (*Procambarus* spp.) bioaccumulation model, *Environmental Toxicology and Chemistry* **23**, 2259–2266.

- [30] Clark, J.S. (2005). Why environmental scientists are becoming Bayesians, *Ecology Letters* **8**, 2–14.
- [31] Collett, D. (2003). *Modelling Survival Data in Medical Research*, 2nd Edition, Chapman & Hall/CRC, London.
- [32] Jonker, M.J., Svendsen, C., Bedaux, J.J.M., Bongers, M. & Kammenga, J.E. (2005). Significance testing of synergistic/antagonistic, dose level-dependent, or dose ratio-dependent effects in mixture dose-response analysis, *Environmental Toxicology and Chemistry* **23**, 2701–2713.

Related Articles

Environmental Carcinogenesis Risk

Statistics for Environmental Teratogenesis

A. JOHN BAILER AND JAMES T. ORIS

Stochastic Control for Insurance Companies

Running a company means to take decisions all the time. One tries to take the decision in an optimal way. Thus, for taking a decision, the management has to solve an optimization problem. Of course, often the optimization problem (*see Asset–Liability Management for Life Insurers; Longevity Risk and Life Annuities; Multiattribute Modeling*) is not explicitly given. The decisions are therefore taken intuitively. However, to get a feeling about which decision should be taken, or to understand how the economy is working, economists, engineers, and mathematicians have developed tools to find or to approximate optimal decisions.

In an insurance company, economic quantities are not the only important factors. An actuary has to take the policy holders' interests into consideration. Thus an actuary has to optimize two contrary positions at the same time: from the point of view of a policy holder, or alternatively the supervisor (like the problems in the sections titled “Optimal Reinsurance” and “Optimal New Business”), from the point of view of a share holder (like in section “Optimal Dividends”), or from some mix of the interests of policy holders and managers (like in section titled “Control of a Pension Fund”).

In this article, we give some examples to illustrate how dynamic optimization problems (*see Role of Alternative Assets in Portfolio Construction; Reliability of Large Systems*) can be solved by stochastic methods. Because we want to keep it simple, we use simple (generic) models and consider simple value functions. For these optimization problems, it will not be difficult to determine the solutions, even though, typically, the solutions have to be obtained numerically. We skip many details. More information and further references can be found in Schmidli [1].

The Basic Model

We start by considering two possibilities to model the surplus of some insurance portfolio. The *classical Cramér–Lundberg model* (*see Large Insurance*

Losses Distributions; Ruin Probabilities: Computational Aspects) in nonlife insurance is

$$X_t = x + ct - \sum_{k=1}^{N_t} Y_k \quad (1)$$

x is the initial capital, i.e., the capital at risk for the considered portfolio of insurance contracts. The premium income is linear with rate c . This is an idealization to make the model tractable. The claim number process is a Poisson process with claim arrival intensity λ (the mean number of claims in a unit interval). The claim sizes $\{Y_k\}$ are independent and identically distributed (i.i.d.) with distribution function $G(y)$ and independent of the claim arrival process $\{N_t\}$. We assume that $G(0) = 0$, i.e., all claims are positive.

Optimal Reinsurance

A main tool for an insurer to decrease his risk is to buy reinsurance (*see Credit Risk Models; Dynamic Financial Analysis; Securitization/Life*). This reduces the capital required by the supervisor. The insurer would then want to know how much reinsurance should be purchased. A popular form of reinsurance is proportional reinsurance. For this form, a retention level $b \in [0, 1]$ is chosen. For a claim Y , the insurer then covers bY ; the rest is covered by the reinsurer. We now allow the insurer to change the retention level continuously. This yields the surplus process (*see Insurance Pricing/Nonlife*) $X_t^b = x + \int_0^t c(b_s) ds - \sum_{k=1}^{N_t} b_{t_k} Y_k$. Here, $b_t \in [0, 1]$ is the retention level chosen at time t , $c(b)$ is the premium rate net the premium rate for reinsurance, and b_{t-} is the retention level (*see Individual Risk Models; Solvency*) chosen immediately prior to t . We suppose that $c(b)$ is strictly increasing in b (more reinsurance costs more), $c(1) = c$ (no reinsurance does not cost anything), and $c(0) < 0$ (full reinsurance is more expensive than first insurance). We define the time of ruin by $\tau^b = \inf\{t : X_t^b < 0\}$ ($\inf \emptyset = \infty$) and the survival probability (*see Default Correlation; Lifetime Models and Risk Assessment*) $\delta^b(x) = \mathbb{P}[\tau^b = \infty]$. The survival probability is a classical measure for the risk and it is also used in the Solvency II agreement (*see Credit Risk Models; Nonlife Insurance Markets*). The goal is to maximize the survival probability. We therefore consider the value function is $\delta(x) = \sup_b \delta^b(x)$, where the maximum

is taken over all strategies $\{b_t\}$, using only the information available until time t .

The problem cannot be solved directly. The following (technical) considerations illustrate how mathematicians can find an equation that the value function should fulfill (see **Multiattribute Value Functions**). Suppose, for the moment, that the strategy $b_t = b$ is constant in a small time interval $(0, h)$ ($h > 0$). After time $T_1 \wedge h = \min\{T_1, h\}$, we follow the optimal strategy, which, we suppose exists for the derivation of equation (5). With probability $e^{-\lambda h}$, there is no claim in $(0, h)$. In this case, the capital at time h is $x + c(b)h$. A claim in $(0, h)$ happens with density $\lambda e^{-\lambda t}$. If a claim happens at time t , then the capital at time t becomes $x + c(b)t - bY_1$. Ruin occurs at time T_1 , if $Y_1 > (x + c(b)t)/b$. Thus we obtain the equation

$$\begin{aligned} \delta(x) &\geq \delta^b(x) = e^{-\lambda h} \delta(x + c(b)h) + \int_0^h \lambda e^{-\lambda t} \\ &\quad \times \int_0^{(x+c(b)t)/b} \delta(x + c(b)t - by) dG(y) dt \end{aligned} \quad (2)$$

Rearranging the terms and dividing by h we obtain

$$\begin{aligned} &\frac{\delta(x + c(b)h) - \delta(x)}{h} - \frac{1 - e^{-\lambda h}}{h} \\ &\quad \times \delta(x + c(b)h) + \frac{1}{h} \int_0^h \lambda e^{-\lambda t} \int_0^{(x+c(b)t)/b} \\ &\quad \times \delta(x + c(b)t - by) dG(y) dt \leq 0 \end{aligned} \quad (3)$$

Supposing that $\delta(x)$ is differentiable, we obtain the limit letting $h \downarrow 0$

$$c(b)\delta'(x) + \lambda \left[\int_0^{x/b} \delta(x - by) dG(y) - \delta(x) \right] \leq 0 \quad (4)$$

This has to hold for all $b \in [0, 1]$. Thus the equation also has to hold if we take the maximum on the left-hand side of equation (4). This motivates the *Hamilton–Jacobi–Bellman equation* (see **Role of Alternative Assets in Portfolio Construction; Reliability Optimization**)

$$\begin{aligned} &\sup_{b \in [0, 1]} c(b)\delta'(x) + \lambda \left[\int_0^{x/b} \delta(x - by) dG(y) - \delta(x) \right] \\ &= 0 \end{aligned} \quad (5)$$

A discussion on why equality should hold instead of the inequality can be found in [1].

Let $b(x)$ be an argument maximizing the left-hand side of equation (5). Because one can argue that $\delta(x)$ is a strictly increasing function, we find that $c(b(x)) > 0$. Let $B = \{b \in [0, 1] : c(b) > 0\}$. For $b \in B$ we have $\delta'(x) \leq \lambda \{c(b)\}^{-1} [\delta(x) - \int_0^{x/b} \delta(x - by) dG(y)]$. Because for $b = b(x)$ the equality has to hold, we get the alternative form of the Hamilton–Jacobi–Bellman equation

$$\delta'(x) = \inf_{b \in B} \lambda \frac{\delta(x) - \int_0^{x/b} \delta(x - by) dG(y)}{c(b)} \quad (6)$$

We can interpret the right-hand side by choosing the $b(x)$ to make the derivative of $\delta(x)$ minimal. This makes sense because we have the boundary condition $\delta(\infty) = 1$, hence the minimal derivative makes $\delta(x)$ maximal.

The derivation of equations (5) and (6) was heuristic. Therefore, one has first to prove a so called *verification theorem*. If we assume that there is a solution $f(x)$ to equation (5) or equation (6) with $f(0) > 0$, then $f(x)$ is strictly increasing and bounded. Moreover, $f(x) = f(\infty)\delta(x)$, and the strategy $b_t = b(X_t)$ is optimal, where $b(x)$ is an argument where the sup (inf) is taken in equations (5), (6), respectively. Thus if a solution to the Hamilton–Jacobi–Bellman equation exists, then we have obtained the correct value function.

Because for any solution $f(x)$ to equation (5), $\alpha f(x)$ is also a solution for any constant $\alpha > 0$, we can start with $f(0) = 1$ for a numerical solution. It follows then that $\delta(x) = f(x)/f(\infty)$. It remains to show that in fact $\delta(x)$ solves equation (5). Because of the verification theorem, it is enough to show that there is a solution to equation (5). For this purpose, we start with $f_0(x) = \delta^1(x)/\delta^1(0)$, where $\delta^1(x)$ is the survival probability if no reinsurance is taken. Then we let recursively $f'_{n+1}(x) = \inf_{b \in B} \lambda \{c(b)\}^{-1} [f_n(x) - \int_0^{x/b} f_n(x - by) dG(y)]$, and $f_{n+1}(x) = 1 + \int_0^x f'_{n+1}(z) dz$. It turns out that $f'_n(x)$ is decreasing in n , and therefore $f_n(x)$ converges to some function $f(x)$ solving equation (6). Thus we have shown that there is a (unique) solution $f(x)$ to equation (6) with $f(0) = 1$.

Example 1 Suppose now that the claim sizes are mixed, exponentially distributed $G(y) = 1 -$

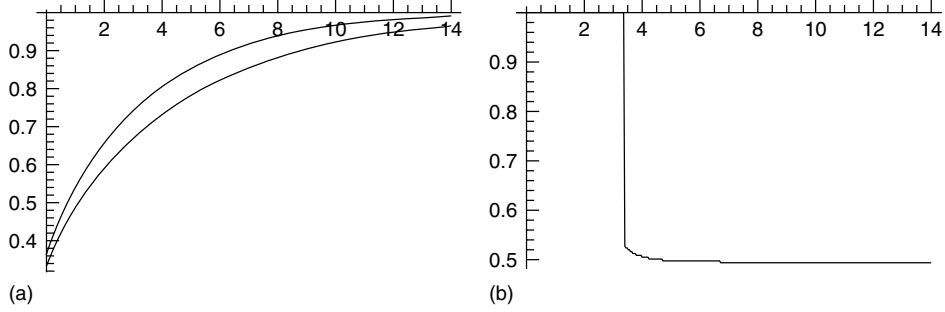


Figure 1 $\delta(x)$ and $\delta^1(x)$ (a) and optimal strategy (b) for mixed exponentially distributed claim sizes

$\frac{1}{3}e^{-0.5y} - \frac{2}{3}e^{-2x}$. One could interpret this distribution function as two sorts of claims: one with mean value 2 appearing with probability $1/3$, the other one with mean value $1/2$, appearing with probability $2/3$. The mean value of the claim sizes is one. We choose $\lambda = 1$ (one claim per time unit) and $c(b) = 1.7b - 0.2$. That is, the premium is $c(1) = c = 1.5$ i.e., the insurer asks for a safety loading of 50% of the net premium. The reinsurer uses a safety loading of 70%. If the process is not controlled, the survival probability is $\delta^1(x) = 1 - 0.606423e^{-0.204667x} - 0.0602437e^{-1.62866x}$. This can be calculated by standard methods. Solving equation (5) gives the optimal survival function $\delta(x)$ and the optimal strategy $b(x)$. Figure 1 shows the uncontrolled survival probability $\delta^1(x)$ and the optimal survival probability $\delta(x)$, as well as the optimal strategy. For small initial capital, one chooses no reinsurance. To buy reinsurance does not strongly reduce the ruin probability. Therefore one wants to get away from zero as quickly as possible. Around $x = 3.3$, one starts to buy reinsurance. The strategy then has a jump.

It can be shown that $\psi(x) = 1 - \delta(x) \sim \xi e^{-Rx}$ as $x \rightarrow \infty$ for some $\xi > 0$. The exponent R is the maximal adjustment coefficient. For our model, we obtain $R = 0.291223$, which is about 50% larger than the adjustment coefficient in the noncontrolled case where $R^1 = 0.204667$. The strategy $b(x)$ converges as $x \rightarrow \infty$ to $b^R = 0.493525$, the retention level for which the maximal adjustment coefficient R is attained. The reduction of the ruin probability is also present for the *value at risk* (see **Credit Migration Matrices; Default Risk**). This is the capital required by the Solvency II rules for insurance companies. $\text{VaR}_{0.05}$ becomes approximately 12.2 for the noncontrolled surplus and approximately 8.5 for

the optimal control. The capital needed for not controlling is therefore approximately 50% larger than for the optimal control.

Optimal New Business

Economists often try to lower the risk by diversification. That is, capital invested is split to independent assets. For an insurer, there is the possibility to participate in some risk exchange. That is, it is possible to buy part of a portfolio from another insurer. We call this risk “new business”. The new business should then cover risks in some area other than the insurer’s original portfolio for the risk to become independent. We simplify the situation in such a way that the insurer is not allowed to sell his original portfolio and can buy or sell part of the new business at any time.

This problem was first treated by Hipp and Takar [2]. The following results can be found in their paper. Suppose the insurer, with the surplus process $\{X_t\}$ modeled by equation (1), has the possibility to buy or sell parts of another surplus process $\{Z_t\}$ with Poisson parameter ℓ , claim size distribution $F(y)$, and premium rate ζ . The net profit condition $\zeta > \ell \int_0^\infty (1 - F(y)) dy$ is fulfilled. At time t , the insurer has the fraction $b_t \in [0, 1]$ of the process $\{Z_t\}$ in his portfolio. Then the surplus of the insurer is $X_t^b := X_t + \int_0^t b_s \zeta - dZ_s$. The goal here is also to minimize the ruin probability. The corresponding Hamilton–Jacobi–Bellman equation is

$$\sup_{b \in [0, 1]} (c + b\zeta)\delta'(x) + \lambda \left[\int_0^x \delta(x - y) dG(y) - \delta(x) \right] + \ell \left[\int_0^{x/b} \delta(x - by) dF(y) - \delta(x) \right] = 0 \quad (7)$$

The alternative form of equation (7) is

$$\delta'(x) = \inf_{b \in [0, 1]} \frac{\lambda \left[\delta(x) - \int_0^x \delta(x-y) dG(y) \right] + \ell \left[\delta(x) - \int_0^{x/b} \delta(x-by) dF(y) \right]}{c + b\zeta} \quad (8)$$

Similarly, for optimal reinsurance, one can prove that there is a unique solution $\delta(x)$ to equations (7) and (8), and this solution is the value function. The strategy $b_t = b(X_t)$ is optimal, where $b(x)$ is an argument at which the supremum on the left-hand side of equation (7) (or infimum in equation (8)) is taken. In the paper by Hipp and Taksar [2], the problem also considered is one in which the process b_t is only allowed to be increasing, i.e., the new risk can only be purchased, but not be sold.

Optimal Dividends

An economist would usually measure the value of a company by the value of the future dividend payments. Thus de Finetti [3] proposed to allow the insurer to pay out a dividend. The postdividend process is $X_t^D = X_t - D_t$, where $\{D_t\}$ is an increasing process with $D_{0-} = 0$. The value of a dividend strategy is $V^D(x) = \mathbb{E}[D_0 + \int_0^{\tau^D} e^{-\delta t} dD_t]$, where $\delta > 0$ is a discount factor, meaning that dividends today are preferred to dividends tomorrow. The time $\tau^D = \inf\{t > 0 : X_t^D < 0\}$ is the time of ruin. At this time, new capital has to be raised to proceed with the business. A share holder is then interested in getting the maximal value of the dividends.

The value function is $V(x) = \sup_D V^D(x)$, where the maximum is taken over all dividend strategies on the basis of the information available, such that $\Delta D_\tau = D_\tau - D_{\tau-} = 0$ (no dividends at ruin).

We start with the classical model (1). There are areas where dividends are paid out and areas where no dividend is paid out. Thus we can expect areas where $\Delta D_t > 0$ and areas where $dD_t = 0$. If $X_t = z$ is such that there is $\varepsilon > 0$ such that no dividends are paid out on $(z - \varepsilon, z)$ and dividends are paid out in $(z, z + \varepsilon)$, then one has to pay out the incoming premiums, i.e., $dD_t = c dt$. On an area where dividends are paid out we obviously have $V'(x) = 1$. On the area where no dividends are paid out one has $V'(x) \geq 1$ because otherwise it would be preferable to pay out the dividends immediately. On

an interval on which no dividends are paid out, the function $V(x)$ has to solve the equation $cV'(x) + \lambda \int_0^x V(x-y) dG(y) - (\lambda + \delta)V(x) = 0$. This can be verified in way similar to inequality equation (4). This leads to the Hamilton–Jacobi–Bellman equation

$$\max \left\{ cV'(x) + \lambda \int_0^x V(x-y) dG(y) - (\lambda + \delta)V(x), 1 - V'(x) \right\} = 0 \quad (9)$$

The proof that $V(x)$ really solves equation (9) is very technical. For details see [1].

The optimal strategy is constructed in the following way. If $V'(x) > 1$, then no dividend is paid out. On interior points where $V'(x) = 1$, the minimal amount is paid out such that a point is reached where the incoming premiums are paid out. On the lower bound of points where dividends are paid out, the incoming premium is paid out as a dividend. That the optimal dividend strategy is of the form described above was first proved by Gerber [4].

Control of a Pension Fund

Consider a pension fund (*see Actuary; Comonotonicity*) and model the outgo for benefits by a diffusion approximation $bt + \sigma_B W_t^B$. One has the possibility to invest into a risky asset and a riskless asset, modeled as a Black–Scholes model (*see Default Risk; Statistical Arbitrage*) given by the (stochastic) differential equations $dZ_t = mZ_t dt + \sigma_1 dW_t^1$ and $dZ_t^0 = \delta Z_t^0 dt$, respectively. The Brownian motions $\{W_t^1\}$ and $\{W_t^B\}$ are supposed to be independent. The pension fund can decide on the proportion θ_t invested into the risky asset at time t and the contribution rate c_t . Then the fund size is described by the stochastic differential equation

$$dX_t^{\theta, c} = [(1 - \theta_t)\delta + m\theta_t]X_t^{\theta, c} dt + \sigma_1 \theta_t X_t^{\theta, c} dW_t^1 + (c_t - b) dt - \sigma_B dW_t^B \quad (10)$$

The fund should keep its size and the contribution rate close to some predefined values $\hat{x}$ and $\hat{c}$, respectively. One therefore chooses a loss function $L(c, x) = (c - \hat{c})^2 + 2\rho(c - \hat{c})(x - \hat{x}) + (k + \rho^2)(x - \hat{x})^2$. This means one punishes deviations from the target values. The parameter ρ allows to skew the punishment, that is, one defines preferences about which directions the deviations should go.

The loss of a strategy $\{\theta_t, c_t\}$ is $V^{\theta,c}(x) = \mathbb{E}[\int_0^\infty e^{-\beta t} L(c_t, X_t^{\theta,c}) dt \mid X_0^{\theta,c} = x]$. The parameter $\beta > 0$ gives some preferences. If β is close to zero, then deviations in the moderate future are punished as well. If β is large, then one wants to be close to the target values now and does not mind much about the future.

The value function is now the minimal expected loss $V(x) = \inf_{\theta,c} V^{\theta,c}(x)$. The corresponding Hamilton–Jacobi–Bellman equation becomes

$$\inf_{\theta,c} L(c, x) + \frac{1}{2}(\sigma_1^2 \theta^2 x^2 + \sigma_B^2) V''(x) + \{(1 - \theta)\delta + m\theta\}x + c - b\} V'(x) - \beta V(x) = 0 \quad (11)$$

The left-hand side is quadratic in c and θ . Thus we have to solve

$$\begin{aligned} & -\frac{1}{4} V'^2(x) + k(x - \hat{x})^2 - \frac{(m - \delta)^2 V'^2(x)}{2\sigma_1^2 V''(x)} \\ & + \frac{1}{2} \sigma_B^2 V''(x) + (\delta x + \hat{c} - \rho(x - \hat{x}) - b) V'(x) \\ & - \beta V(x) = 0 \end{aligned} \quad (12)$$

Trying a solution of the form $V(x) = Ax^2 + Bx + C$ yields the solution

$$\begin{aligned} A &= -\frac{(m - \delta)^2}{2\sigma_1^2} + (\delta - \rho) - \frac{1}{2}\beta \\ &+ \sqrt{\left(-\frac{(m - \delta)^2}{2\sigma_1^2} + (\delta - \rho) - \frac{1}{2}\beta\right)^2 + k} \\ B &= \frac{2[(\hat{c} + \rho\hat{x} - b)A - k\hat{x}]}{A + (m - \delta)^2\sigma_1^{-2} + \beta + \rho - \delta} \\ C &= \beta^{-1} \left[-\frac{B^2}{4} + k\hat{x}^2 - \frac{(m - \delta)^2 B^2}{4A\sigma_1^2} \right. \\ & \left. + \sigma_B^2 A + (\hat{c} + \rho\hat{x} - b)B \right] \end{aligned} \quad (13)$$

Note that the other solution for A is strictly negative and therefore is not the desired solution because we need a convex solution.

It is proved in Cairns [5], see also [1], that the value function really is $V(x) = Ax^2 + Bx + C$ with A, B, C given above. The optimal strategy is of feedback form $\{(\theta_t, c_t)\} = \{(\theta(X_t), c(X_t))\}$ with

$$\begin{aligned} \theta(x) &= -\frac{(m - \delta)(Ax + B/2)}{\sigma_1^2 Ax}, \\ c(x) &= \hat{c} - \rho(x - \hat{x}) - Ax - B/2 \end{aligned} \quad (14)$$

Note that we take a short position (see **Credit Risk Models**) in the risky asset if the reserve lies above $-B/(2A)$. In some sense, we “destroy” wealth if our reserve is too large. This is because the loss function punishes a position above the target value in the same way as a position below. One therefore should choose a more reasonable loss function than quadratic loss.

References

- [1] Schmidli, H. (2008). *Stochastic Control in Insurance*, Springer-Verlag, London.
- [2] Hipp, C. & Taksar, M. (2000). Stochastic control for optimal new business, *Insurance: Mathematics and Economics* **26**, 185–192.
- [3] de Finetti, B. (1957). Su un' impostazione alternativa della teoria collettiva del rischio, *Transactions of the XVth International Congress of Actuaries* **2**, 433–443.
- [4] Gerber, H.U. (1969). Entscheidungskriterien für den zusammengesetzten Poisson-Prozess. Schweizerische Vereinigung der Versicherungsmathematiker, *Mitteilungen* **69**, 185–228.
- [5] Cairns, A.J.G. (2000). Some notes on the dynamics and optimal control of stochastic pension fund models in continuous time, *ASTIN Bulletin* **30**, 19–55.

HANSPETER SCHMIDLI

Stratospheric Ozone Depletion

Stratospheric ozone (O_3) shields the surface of the Earth from solar short-wave ultraviolet B (UVB) radiation. A model of the photochemical process responsible for the formation and destruction of ozone was first proposed by Chapman [1]. Chapman's model involved rapid recycling between oxygen and ozone. This model has since been extended and it is now understood that stratospheric ozone is destroyed not only by oxygen but also by other chemical species. In the 1970s, it was recognized that chlorine could catalyze ozone destruction and, in 1974, Mario Molina and Sherwood Rowland identified man-made chlorofluorocarbons (CFCs) as the main source of ozone-destroying chlorine in the stratosphere [2]. For this work, Molina and Rowland shared the 1995 Nobel prize in Chemistry with Paul Crutzen. Excellent surveys of the processes governing the concentration and distribution of ozone in the stratosphere can be found in Solomon [3] and Staehelin *et al.* [4].

The use of CFCs in refrigerators, aerosol sprays, and cleaning solvents began in the 1930s and, with increased use, their atmospheric concentrations grew rapidly. The discovery of their role in the destruction of stratospheric ozone caused considerable concern about these increasing concentrations. This concern revolved around the potential risks of increased levels of UVB radiation to human health. These risks include elevated rates of skin cancer, eye disease, and immune system suppression [5]. Increased UVB radiation also poses a potential ecological risk owing to its effects on animal health and plant growth.

Concern over ozone depletion led to calls in the 1980s for an international agreement to control the production of CFCs and other ozone-depleting compounds. A central question in the debate over such an agreement concerned the magnitude of the risks posed by ozone depletion. Although the basic chemistry of ozone depletion had been well understood since the work of Molina, Rowland, and others in the 1970s, to assess the risks of ozone depletion to human health and ecosystems, it was necessary to know how the actual concentration of stratospheric ozone had varied – and would likely vary in the future – over particular regions of the Earth's surface. The spatial

and temporal distribution of ozone in the stratosphere is affected by a number of factors, including solar variability, volcanic activity and atmospheric chemistry, temperature, and circulation. A striking example of this is the formation of so-called ozone holes in the spring over the North and South Poles. These dramatic, but temporary, reductions in ozone concentration are associated with atmospheric anomalies called *polar vortices*. The formation of polar vortices has substantial local effects on ozone dynamics. In particular, when a polar vortex forms, extremely cold polar air is essentially isolated from atmospheric circulation. The cold temperatures promote the process of ozone destruction and the isolation prevents relatively ozone-rich air from midlatitudes from flooding in. Although several studies have shown that the depth and spatial extent of the polar ozone holes have increased over time, these phenomena are restricted to polar regions.

The understanding of the role of these and other (possibly unknown) factors in determining the details of ozone concentration was felt by many to be insufficiently developed to serve as the sole basis of a quantitative risk assessment. As a result, statistical analyses of ozone concentrations came to play a central role in the scientific discussion.

Surface measurements of total column ozone were first made in the 1920s using an instrument called the *Dobson spectrophotometer*. Briefly, the Dobson ozone measurement is based on the ratio of intensities of incoming solar radiation at two UV wavelengths, one of which is strongly absorbed by ozone while the other is not. The measurement is calibrated from knowledge of these intensities at the top of the atmosphere. The longest record of Dobson measurements, which began in 1926, is for Arosa, Switzerland. However, it was not until the International Geophysical Year of 1957 that a broad network of Dobson stations was established. Solow [6] reviewed two analyses of the Dobson data by Reinsel and Tiao [7] and Bloomfield *et al.* [8] aimed at estimating a global trend. Although differing in a number of ways, both analyses were based on statistical models incorporating a regional component, a seasonal component, and a secular temporal trend. In addition, both models made allowance for temporal correlation in unmodeled ozone variations. The original papers should be consulted for model details. Reinsel and Tiao represented the trend as constant up to 1970, followed by a linear trend. They found no significant depletion

over the period 1970–1983. This result remained unchanged when a measure of solar variability was included in the analysis. Bloomfield *et al.* assumed that the trend was proportional to an ozone depletion curve based on a photochemical model and also allowed for a solar effect and the effect of atmospheric nuclear tests. The best-fitting depletion curve was around 10% of that predicted by the photochemical model. In overall terms, these two analyses found little evidence of global ozone depletion over the observation period. Other analyses of global total column ozone found similar results.

Dobson measurements of total column ozone can be converted into estimates of the vertical profile of ozone concentration using the Umkehr method. The Umkehr method is based on a division of the atmosphere into separate layers. The distribution of ozone across these layers is estimated from the variation in the Dobson measurement as the sun subtends an angle of 60–90° near dawn and dusk. Because photochemical models suggest that ozone depletion should be greatest in the upper atmosphere, Reinsel and Tiao repeated their analysis for the upper five Umkehr layers. In doing so, they had to deal with a large number of missing observations. The extended analysis found significant ozone depletion in the upper three Umkehr layers.

In addition to the ground-based Dobson data, satellite measurements of ozone have been available since the late 1970s. The longest record, which began in 1978, is for the total ozone mapping spectrometer (TOMS) aboard the NIMBUS 7 satellite. This instrument operates on the same general spectrometric principles as the Dobson instrument, but uses reflected, rather than incident, solar radiation. An advantage of the TOMS data over the ground-based data is its global coverage. Niu and Tiao [9] used the TOMS data to investigate trends at different latitudes in total column ozone over the period 1978–1990. A central part in their analysis was the incorporation of a spatial–temporal autocorrelation (STAR) model into the error process (*see Spatial Risk Assessment*). Again, the original paper should be consulted for model details. Assuming linear trends, Niu and Tiao found rates of depletion increasing asymmetrically with latitude from essentially 0° at the equator to 4% per decade at 65°N and 10% per decade at 65°S.

By the mid-1980s, the risks of ozone depletion were felt to be substantial enough that an international agreement called the *Montreal Protocol* was

signed in 1987 to limit the production of CFCs and other ozone-depleting compounds. The Montreal Protocol, which was amended in 1990 and 1992, has successfully reduced CFC emissions. As the lifetime of CFCs in the atmosphere is on the order of 50–100 years, the full effect of the Montreal Protocol will not be felt for decades. Efforts have been made, however, to identify an effect on ozone concentration. Newchurch *et al.* [10] analyzed data from two satellite experiments – the Stratospheric Aerosol and Gas Experiment (SAGE) and the Halogen Occultation Experiment (HALOE) – over the period 1979–2000 to address this issue. Briefly, the analysis began by fitting a statistical model similar to those described above to data over the period 1979–1996. An attempt was made in this model to account for the effect of solar variability and the quasi-biennial oscillation (QBO), an atmospheric oscillation of period ~2.5 years. Also, ozone observations contaminated by volcanic eruptions and other aerosol effects were omitted from the analysis. In the second part of the analysis, the cumulative sums (CUSUMs) of deviations from the fitted model for the out-of-sample period 1997–2000 were used to detect a decrease in the rate of ozone depletion. The results suggest that such a slowdown had indeed occurred. Future statistical work in this area would benefit from going beyond relatively simple model specifications – involving, for example, linear trends – to incorporating the output of physical models as in Bloomfield *et al.* [8]. A more recent example of this kind of work is described in Guillas *et al.* [11], in which a physical model of chemical transport is used to assess ozone recovery following the implementation of the Montreal Protocol.

References

- [1] Chapman, S. (1930). On ozone and atomic oxygen in the upper atmosphere, *Philosophical Magazine* **10**, 369–383.
- [2] Molina, M.J. & Rowland, F.S. (1974). Stratospheric sink for chlorofluoromethanes: chlorine atom catalyzed destruction of ozone, *Nature* **249**, 820–821.
- [3] Solomon, S. (1999). Stratospheric ozone depletion: a review of concepts and history, *Reviews of Geophysics* **37**, 275–316.
- [4] Staehelin, J., Harris, N.R.P., Appenzeller, C. & Eberhard, J. (2001). Ozone trends: a review, *Reviews of Geophysics* **39**, 231–290.
- [5] Longstreth, J.D., De Gruijl, D., Kripke, F.R., Takizawa, Y. & van der Leun, J.C. (1995). Effects of

- increased solar ultraviolet radiation on human health, *Ambio* **24**, 153–165.
- [6] Solow, A.R. (1994). Statistical methods in atmospheric science, *Handbook of Statistics* **12**, 717–734.
- [7] Reinsel, G.C. & Tiao, G.C. (1987). Impact of chlorofluoromethanes on stratospheric ozone, *Journal of the American Statistical Association* **82**, 20–30.
- [8] Bloomfield, P., Oehlert, G., Thompson, M.L. & Zeger, S.L. (1983). A frequency domain analysis of trends in Dobson total ozone records, *Journal of Geophysical Research* **88**, 8512–8522.
- [9] Niu, X. & Tiao, G.C. (1995). Modeling satellite ozone data, *Journal of the American Statistical Association* **90**, 969–983.
- [10] Newchurch, M.J., Yang, E., Cunnold, D.M., Reinsel, G.C., Zawodny, J.M. & Russell, J.M. (2003). Evidence for slowdown in stratospheric ozone loss: first stage of ozone recovery, *Journal of Geophysical Research* **108**, D4507.
- [11] Guillas, S., Stein, M.L., Wuebbles, D.J. & Niu, X. (2004). Using chemistry transport modeling in statistical analysis of stratospheric ozone trends from observations, *Journal of Geophysical Research* **109**, D22303.

Related Articles

Global Warming

Risk from Ionizing Radiation

ANDREW R. SOLOW

Stress Screening

Stress screening (*see Credit Migration Matrices; Extreme Values in Reliability; Reliability Integrated Engineering Using Physics of Failure*) is a combined stress and functional test applied to products prior to dispatch to prevent the delivery of parts with latent defects that may cause failure at some time during use. Manufacturers use it as a final check process to limit the risk of failure within the guaranteed life and avoid the resulting consequences. Stress screening is very rarely used for mass produced items and is only applied to medium- and low-volume products in particular circumstances. It is sometimes employed when product manufacture involves materials or processes that are difficult to fully control and are likely to include or create latent defects. It is more often used for high-quality, low-volume, complex products, where the consequences of service failure are severe. The consequences may be financial owing to contractual penalties or warranty or may have individual or more general public safety issues. A very simple example of stress screening, although never so-called, is the proof (over pressure) test of a pressure vessel. The stress applied during the screening is usually environmental, e.g., temperature or vibration, and hence its alternative name of environmental stress screening (ESS) (*see Accelerated Life Testing*). The test condition usually comprises more than one stress and can also include product-related loads, e.g., voltage, in an electrical product. Traditionally, the stress applied is restricted to specification limits but, more recently, it has become accepted to use greater stresses in order to increase screen effectiveness. The screening regime is related to a stress testing reliability enhancement process at design stage that increases applied stress to the point of product failure. This determines the margin over the specification limit built into the design and the results can be used to set safe screening levels outside the specification.

One should never use stress screening as a replacement for manufacturing process control or to enable the fielding of poorly designed products. Ensuring that a product is fit for purpose by development testing at the design stage and controlling the materials and processes used in manufacture are much more cost-effective methods of minimizing service failures.

Stress screening is testing designed to overstress latent manufacturing defects with a potential to cause field failure so that they can be detected and not delivered. Stress screening is useful in identifying batch-related and systematic manufacturing problems so that corrective action can be undertaken avoiding product recalls and is a process that enables a continuous improvement environment.

Typical Applications

Applications are typified by low-volume products with manual intervention in final assembly and where the consequences of service failure are severe. In the aviation industry, particularly for electromechanical and electronic systems, stress screening is used extensively and often mandated by the customer. For example, despite the redundancy embedded in the design of a full authority digital engine control (FADEC) controlling a large civil airplane engine, stress screening is used to reduce the risk of service failure and engine run down. The difficulty we experience is in specifying the screening strategy and stress levels so that maximum benefit follows, i.e., 100% detection of defective products with no weakening of good products.

Developing a Stress Screening Regime

There has been a recent tendency, driven by proactive equipment suppliers, to try and impose a standard basic screening regime with variation in stress level alone. This is ill-conceived as every product and application should be considered independently. The key factors to consider in determining the need to screen and selecting the applied stresses are product construction and application environment.

The structure of the product is important as some sublevel screening may be desirable, particularly if appropriate stress cannot be easily applied to individual parts when the product is fully assembled. For example, a process flow involving the screening of individual devices and printed circuit board assemblies as well as the completed product is often used for high-reliability electronic products.

Most product failures, even in electronics, are thermomechanical in nature and an appreciation of the field environmental stresses likely to promote latent defects to failure is essential. Transport and storage conditions should also be considered, as for

benign applications they may be the worst conditions experienced. However, the best guide is provided by an analysis of the field failures of a similar product. A Pareto analysis (*see Large Insurance Losses Distributions; Ruin Probabilities: Computational Aspects*) of the defects that escape and cause field failure indicates what the screening must be designed to identify.

The use of stress levels beyond specified limits raises the possibility of the screen causing damage to products released to service; therefore, we must design the screening very carefully. The following inputs should be considered when setting stress levels:

1. Marginal conditions indicated by simulation of product performance under the expected extremes of environmental stress and product loads.
2. Formal design and process failure modes effects and corrective action (FMECA) studies.
3. Failure concerns elicited from design and manufacturing engineers and operators as they often understand the product and its potential weaknesses better than anyone.

The ideal level of stress can be better appreciated by considering the diagram (see Figure 1) where *specification limit* is the extremes of the environment specified by the customer or expected, given the application. *Design limit* is the limit within which the device has been theoretically designed to operate, giving some margin over the worst expected

conditions. *Operating limit* is the functional limit, characterized by soft failure, i.e., beyond this functionality is lost but resumes after removal of stress. *Destruct limit* is the point at which hard failure occurs, i.e., functionality does not resume after removal of stress. Ideally, the screening stress range should be as indicated but this is a theoretical diagram and the various limits are seldom known for certain and can vary with product or component batch. It is also always preferable to monitor function throughout the test as failure may occur at only certain parts of the stress regime.

The commonly used stresses are temperature, rapid change of temperature, vibration, mechanical shock, mechanical stress, voltage, and rapid on-off cycles. The means of applying stress will be product dependent and may be by operation outside normal limits. Stresses applied in combination can also be more effective, e.g., vibration at elevated or low temperatures is often more damaging as the stiffness of different materials can change significantly with temperature.

Screening Equipment

We have emphasized the need to treat every product on its own merits but the screening conditions used often involve temperature and vibration. Conventional temperature and vibration test chambers can be used, but equipment is available specifically developed for product screening and the allied step-stress development testing. The equipment manufacturers have introduced the names highly accelerated life testing (HALT) (*see Hazard and Hazard Ratio*) for such reliability testing at design and highly accelerated stress screen (HASS) (*see Product Risk Management: Testing and Warranties*) for accelerated stress screening [1]. These equipment employ rapid thermal cycling and six degrees of freedom, repetitive shock testing, akin to a multidirectional random vibration.

Screen Validation and Feedback

Even if stress levels are maintained within the operating limit, there is still a danger that significant life may be taken out of the product. Analysis using physics of failure life prediction under the applied stresses can be used to indicate if there is any reason for concern. The promoters of the HASS technique

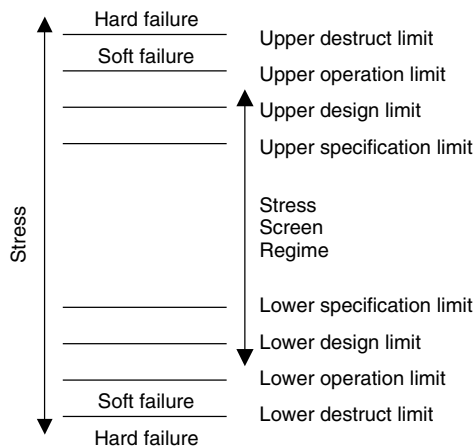


Figure 1 Schematic of stress levels in stress screening

recommend a validation test, where they submit a sample product to 20 times the proposed screen. If still working they conclude that the screen, at worst, will not take more than 5% of the life out of the product.

Whatever screening process is used, to be really effective it must be accompanied by a thorough analysis of both screening and field failures. This continuing analysis will check that the screen is effective, enable screening improvement, indicate possible product and manufacturing improvements, and confirm screening remains necessary. It is recognized that as design and manufacture matures the need for product stress screening may decline. Screening was

traditionally used in the final testing of electronic devices, but with the much improved process control and automated fabrication demanded by huge electronic volumes, screening has been largely eliminated. Manufacturers demonstrated that screening was beginning to cause more damage than good.

Reference

- [1] Gregg, K. (2000). *Hobbs Accelerated Reliability Engineering – HALT and HASS*, ISBN 0-471-97966, John Wiley & Sons.

ROBERT M. NEWMAN

Structural Models of Corporate Credit Risk

The Merton Model

The best way to introduce the structural approach starts with the Merton model (*see Credit Migration Matrices; Default Risk; Default Correlation; Credit Value at Risk*). This was the first structural model in the literature, and it contains the core idea of the approach. In fact, most recent models are extensions of the Merton model. Merton [1] considers that the asset value of a firm follows a geometric Brownian motion (GBM) (*see Insurance Pricing/Nonlife; Risk-Neutral Pricing; Importance and Relevance; Simulation in Risk Management; Weather Derivatives*) and assumes that the firm issues only one zero coupon (*see Options and Guarantees in Life Insurance; Credit Migration Matrices; Mathematical Models of Credit Risk*) corporate bond. Default is the event where the firm is unable to settle the promised payment upon the bond maturity.

Let V_t be the value of the firm's assets at time t , and K be the promised payment at maturity T , or the face value of the bond. If the firm value is higher than the face value on the maturity date, then equity holders will receive the residual value of the firm after paying back the debt; otherwise, they will receive nothing because the firm will declare bankruptcy and debt holders will own the firm. Hence, the payoff for equity holders on the debt maturity date is given by

$$\max(V_T - K, 0) \quad (1)$$

This payoff function is exactly that of the standard call option. Merton assumes a complete market and employs the Black–Scholes (BS) option pricing formula (*see Structured Products and Hybrid Securities; Default Risk; Statistical Arbitrage*) to obtain the market value of equity, S_t . More precisely, the asset price dynamics is postulated to be

$$\frac{dV}{V} = \mu dt + \sigma dW \quad (2)$$

where μ is the asset drift rate, σ is the volatility of the firm asset value, and W is the Wiener process. The BS formula is given by

$$S_t = V_t N(d_1) - K e^{-r(T-t)} N(d_2) \quad (3)$$

$$d_1 = \frac{\ln(V_t/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}},$$

$$d_2 = d_1 - \sigma\sqrt{T-t} \quad (4)$$

where r is the constant risk-free interest rate, and $N(x)$ is the cumulative distribution function of a standard normal random variable.

From the bond holders' perspective, on the maturity date, if the firm value is larger than the face value, then they will receive the face value; otherwise, they will obtain an amount equivalent to the firm's asset value. Hence, the payoff for bond holders is

$$\min(V_T, K) = K - \max(K - V_T, 0) \quad (5)$$

The present value of this corporate zero coupon bond can be fully replicated by a portfolio of a long position in a default-free bond and a short position in a put option. The default-free bond has a face value K and the put option has a strike price K and maturity T . Denote $B(t, T)$ as the corporate bond price. We have

$$B(t, T) = K e^{-r(T-t)} - K e^{-r(T-t)} N(-d_2) + V_t N(-d_1) \quad (6)$$

where d_1 and d_2 have been defined in equation (4).

Probability of Default

An objective in credit risk analysis is to measure the probability of default (PD) (*see Credit Scoring via Altman Z-Score; Credit Migration Matrices; Default Risk*) of a credit obligor. As the Merton model assumes that default only occurs on the maturity date of the debt, the default probability is the probability that the terminal asset value V_T is smaller than the face value K . This probability can be expressed in a closed form:

$$\begin{aligned} \text{PD} &= \Pr(V_T < K) \\ &= N\left(\frac{\ln\left(\frac{K}{V_t}\right) - \left(\mu - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right) \end{aligned} \quad (7)$$

As the Merton model considers the capital structure of the firm and the incentive issues of equity and

bond holders, it is known as a *structural model*. In the financial market, structural models are also known as *option-theoretic approaches* to credit risk. The Merton model shows that option pricing theory plays a key role in the formulation. We will see shortly that other structural models have similar characteristics, but relax some of the assumptions.

The major advantage of the Merton model is its analytical tractability. The trade-off is that the over-simplified assumptions limit the application of the model. First, it assumes a constant risk-free interest rate. As corporate bonds usually have long periods of maturity, a realistic model should incorporate a stochastic interest rate (see **Mathematical Models of Credit Risk**). Secondly, the Merton model only allows default to occur on the maturity of debt, but firms can default at any time in the real world. Thirdly, interim payments such as bond coupons are not considered in the analysis. Finally, the use of the GBM ignores the hazard of a sudden drop in asset value. Hence, it is important to develop models beyond Merton's idea.

First Passage Time Models

To allow firms to default at any time, Black and Cox [2] (see **Simulation in Risk Management; Default Correlation**) introduce the notion of a default barrier. When the firm value goes below a threshold, bond holders can enforce equity holders to declare bankruptcy. This indenture covenant for bond holders avoids further deterioration of the firm's asset value.

On top of the notation in the Merton model, let H be the exogenously specified default barrier (see **Credit Risk Models**) or distress level. In the Black–Cox model, equity holders receive the residual value of the firm if the terminal asset value is higher than the face value and the asset value never goes below the barrier prior to maturity. Suppose that only one zero corporate bond is issued by the underlying firm. The payoff for equity holders becomes

$$\max(V_T - K, 0)I_{\{\tau_H > T\}} \quad (8)$$

where I_A is the indicator function for the event A and τ_H is the first passage time: $\tau_H = \inf\{t : V_t < H\}$. This payoff function resembles the payoff of a down-and-out call (DOC) option. When the default-free

interest rate is a constant, there is a closed-form solution for the DOC option. Specifically,

$$\begin{aligned} S_t &= DOC(V_t, K, H) \\ &= V_t N(a) - K e^{-r(T-t)} N\left(a - \sigma\sqrt{T-t}\right) \\ &\quad - V_t (H/V_t)^{2\eta} N(b) + K e^{-r(T-t)} (H/V_t)^{2\eta-2} \\ &\quad \times N\left(b - \sigma\sqrt{T-t}\right) \end{aligned} \quad (9)$$

where

$$a = \frac{\ln(V_t/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} \quad (10)$$

$$b = \frac{\ln(H^2/V_t K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} \quad (11)$$

$$c = \frac{\ln(H/V_t) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} \quad \text{and} \quad (12)$$

$$\eta = \frac{r}{\sigma^2} + \frac{1}{2} \quad (13)$$

In equation (9), it is assumed that the strike price K is larger than the barrier H . If the barrier is higher than the strike price, then it is not reasonable for the firm to default when the barrier is hit, because the firm value is large enough to pay back the face value K .

We now consider the situation of bond holders. The face value will be received if the terminal firm value is higher than the face value and the firm value does not breach the downside default barrier. This payoff structure can be fully replicated by a portfolio of a long position in a default-free zero coupon bond with face value K , a short position in a put option with strike price K and a long position in a down-and-in call option with a barrier H and strike K .

Probability of Default

As default is triggered by the default barrier in first passage time models, the definition of the PD should be revised. Default is now viewed as the event that the firm value hits the downside barrier within a given time horizon. Mathematically, we write:

$$PD = \Pr(\tau_H < T) \quad (14)$$

If we adopt the GBM for the asset price dynamics, then a closed-form solution can be obtained for the PD:

$$N\left(\frac{\ln \frac{H}{V_t} - \left(\mu - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right) + \left(\frac{V_t}{H}\right)^{\frac{2\left(\mu - \frac{\sigma^2}{2}\right)}{\sigma^2}} \times N\left(\frac{\ln \frac{V_t}{H} - \left(\mu - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right) \quad (15)$$

However, the PD cannot be obtained in closed form in general situations, such as the GBM with time-dependent coefficients.

The KMV Approach

Moody's KMV approach is a first passage time model. Although the actual formulation of the approach is proprietary, it is well known that it views the market value of equities as a perpetual DOC option: refer to [3] for technical details. The default barrier and the strike price are both set to the default point, which is defined as the sum of short-term debt and half of the long-term debt reported in a firm's financial statement.

The KMV approach estimates the default probability rather than valuing corporate bond prices. Therefore, the market value of equities will be used as an input to estimate the market values of the firm and its volatility. The PD then is the output of the system. The methods of estimating firm values and volatility will be detailed later.

Endogenous Bankruptcy

The first passage time models discussed so far are based on an exogenous default barrier. Leland and Toft [4] propose that it is possible to obtain the barrier level endogenously. The endogenous bankruptcy should be set at the level at which both the firm value and equity value are maximized. Bankruptcy costs, taxes, asset payout, and coupon payment are incorporated in the model in [4].

Suppose that the default barrier, H , is exogenously given. Under the GBM for the asset price process, Leland and Toft [4] determine the total value of all

outstanding bonds:

$$\begin{aligned} B(V; H, T) &= \frac{C}{r} + \left(K - \frac{C}{r}\right) \left(\frac{1 - e^{-rT}}{rT} - I(T)\right) \\ &\quad + \left[(1 - \alpha)H - \frac{C}{r}\right] J(T) \\ I(T) &= \frac{1}{rT} [G(T) - e^{-rT} F(T)] \\ J(T) &= \frac{1}{z\sigma\sqrt{T}} \left[-\left(\frac{V}{H}\right)^{-a_1+z} N(q_1)q_1 \right. \\ &\quad \left. + \left(\frac{V}{H}\right)^{-a_1-z} N(q_2)q_2 \right] \\ F(T) &= N(h_1) + \left(\frac{V}{H}\right)^{-2a_1-z} N(h_2) \\ G(T) &= \left(\frac{V}{H}\right)^{-a_1+z} N(q_1) + \left(\frac{V}{H}\right)^{-a_1-z} N(q_2) \end{aligned} \quad (16)$$

where

$$\begin{aligned} q_1 &= \frac{\ln \frac{H}{V} - z\sigma^2 T}{\sigma\sqrt{T}}; \quad q_2 = \frac{\ln \frac{H}{V} + z\sigma^2 T}{\sigma\sqrt{T}} \\ h_1 &= \frac{\ln \frac{H}{V} - a_1\sigma^2 T}{\sigma\sqrt{T}}; \quad h_2 = \frac{\ln \frac{H}{V} + a_1\sigma^2 T}{\sigma\sqrt{T}} \\ a_1 &= \frac{r - \delta - \sigma^2/2}{\sigma^2}; \quad z = \frac{[(a_1\sigma^2)^2 + 2r\sigma^2]^{1/2}}{\sigma^2} \end{aligned} \quad (17)$$

and δ is the asset payout rate, K is the total principal value of all outstanding bonds, C is the total coupon paid by all outstanding bonds per year, and α is the fraction of firm's asset value lost in bankruptcy; i.e., the remaining value $(1 - \alpha)H$ is distributed to bond holders in bankruptcy.

The total market value of the firm, v , equals the asset value V plus the value of tax benefits, less the value of bankruptcy costs. Tax benefits accrue at rate τC per year as long as $V > H$, where τ is the corporate tax rate. The total firm value is given by

$$\begin{aligned} v(V; H) &= V + \frac{\tau C}{r} \left[1 - \left(\frac{V}{H}\right)^{a_1+z} \right] \\ &\quad - \alpha H \left(\frac{V}{H}\right)^{-a_1-z} \end{aligned} \quad (18)$$

Hence, the market value of equity is given by

$$S(V; H, T) = v(V; H) - B(V; H, T) \quad (19)$$

To determine the equilibrium default barrier H endogenously, Leland and Toft [4] specify that the equity valuation formula should satisfy the smooth-pasting condition, under which the value of equity and the value of the firm are maximized. Specifically, H must be solved from the following equation:

$$\left. \frac{\partial S(V; H, T)}{\partial V} \right|_{V=H} = 0 \quad (20)$$

The solution is given by

$$H = \frac{(C/r)(A/(rT) - B) - AK}{(rT) - \tau C(a_1 + z)/r} \quad (21)$$

$$1 + \alpha(a_1 + z) - (1 - \alpha)B$$

where

$$A = 2a_1 e^{rT} N(a_1 \sigma \sqrt{T}) - 2z N(z \sigma \sqrt{T}) + (z - a_1) - \frac{2}{\sigma \sqrt{T}} n(z \sigma \sqrt{T}) + \frac{2e^{-rT}}{\sigma \sqrt{T}} n(a_1 \sigma \sqrt{T}) \quad (22)$$

$$B = - \left(2z + \frac{2}{z \sigma^2 T} \right) N(z \sigma \sqrt{T}) - \frac{2}{\sigma \sqrt{T}} n(z \sigma \sqrt{T}) + (z - a_1) + \frac{1}{z \sigma^2 T} \quad (23)$$

and $n(\cdot)$ denotes the standard normal density function. Once the endogenous barrier is obtained, the equity and bond pricing formulas are respectively defined in equations (16) and (18) with the endogenous barrier substituted into the formulas.

Implementation Methods

The challenge of implementing structural models is the estimation of the value and risk of the firm's assets, both of which are not directly observable. As the market values of equities are observable for listed corporations, it is the market convention to infer the market value of a firm from its equity values.

Once a structural model has been chosen, there is a one-to-one relationship between the equity and the firm's value. To allow general discussion, we write $S = S(V; \sigma, T)$ and our target is to estimate both V and σ from a given data set of equity values:

$\mathbf{S} = \{S_0, S_1, \dots, S_n\}$, where $S_i = S(t_i)$. For instance, if the Merton model is used, then the function $S(V; \sigma, T)$ is the European call option (*see Equity-Linked Life Insurance; Options and Guarantees in Life Insurance*) formula as in equation (3). If the Black-Cox model is considered, then it becomes the DOC option formula as in equation (9).

Variance Restriction

To implement the Merton model, Ronn and Verma [5] propose a variance-restriction method that can be applied to other models as well. The idea is to apply Ito's lemma on the equity pricing formula to derive a formula for the equity volatility. The estimation then requires that the observed equity price and volatility should match the model outputs.

Because it is assumed that the firm value follows GBM as in equation (2), invoking Ito's lemma on $S(V; \sigma, T)$ yields:

$$\frac{dS}{S} = (\dots) dt + \sigma \frac{V}{S} \frac{\partial S}{\partial V} dW \quad (24)$$

This implies that the equity volatility relates to the firm's asset value and its volatility by

$$\sigma_S = \sigma \frac{V}{S} \frac{\partial S}{\partial V} \quad (25)$$

where σ_S denotes the equity volatility. Together with the equity pricing formula, it forms a system of two equations at each observed time point t_i :

$$S_i = S(V_i, \sigma_i, T) \quad \text{and} \quad \sigma_S = \sigma_i \frac{V_i}{S_i} \left[\frac{\partial S}{\partial V} \right]_{V=V_i} \quad (26)$$

For the Merton model, the term $\frac{\partial S}{\partial V}$ can be explicitly calculated as $N(d_1)$, where d_1 is defined in equation (3).

We summarize the implementation procedure as follows:

- Step 1. Calculate the rate of returns from the sample $\mathbf{S} = \{S_0, S_1, \dots, S_n\}$ using the formula $R_i = (S_i - S_{i-1})/S_{i-1}$ for $i = 1, 2, \dots, n$.
- Step 2. Compute the sample equity volatility: $\sigma_S = \sqrt{\text{var}(R)/\Delta t}$.
- Step 3. Solve the system of equations (26) for $i = 0, 1, \dots, n$.

In step 3, a numerical root-finding method should be employed. Although this variance-restriction method is very efficient, it violates the constant volatility assumption for the firm value process. If the equity volatility is a constant, then from the second equation of (26) the firm value volatility must be a stochastic variable. Once σ is stochastic, the option pricing formula is no longer true. Although one immediately recognizes that this is theoretically inconsistent, see [6], this variance-restriction method is still the most common way to implement structural models so far.

Maximum-Likelihood Estimation

A more advanced approach uses statistical techniques to estimate the value and risk of the firm. Duan [7] proposes a maximum-likelihood estimation to structural models. The idea is to derive the likelihood function for equity prices subject to the constraint that the equity price matches the model price and the firm value follows a GBM.

Duan [7] derives the log-likelihood function as,

$$L(\mu, \sigma) = \sum_{i=1}^n \ln g(V_i | V_{i-1}) - \sum_{i=1}^n \ln \left[V_i \cdot \frac{\partial S}{\partial V} \Big|_{V=V_i} \right] \quad (27)$$

where

$$g(V_i | V_{i-1}) = \frac{1}{\sigma \sqrt{2\pi}} \times \exp \left(-\frac{[\ln(V_i / V_{i-1}) - (\mu - \sigma^2/2)(t_i - t_{i-1})]^2}{2\sigma^2(t_i - t_{i-1})} \right) \quad (28)$$

The estimation is achieved by maximizing the log-likelihood function numerically subject to the constraint:

$$S_i = S(V_i, \sigma_i, T) \quad (29)$$

Hence, the implementation requires numerical optimization methods provided by software. For example, the subroutines “fminsearch” and “fmincon” provided by MATLAB can be used to numerically maximize the likelihood function. Numerical examples can be found in [6, 7].

KMV Approach

The following approach is employed by KMV to estimate parameters in structural models: see [8].

- Step 1. Set an initial value of $\sigma = \sigma^{(0)}$ and $j = 0$.
- Step 2. Solve V_i from the equation: $S_i = S(V_i; \sigma^{(j)}, T)$ for $i = 0, 1, \dots, n$.
- Step 3. Set $j = j + 1$ and $\sigma^{(j)}$ to be the sample volatility of $\{V_0, \dots, V_n\}$.
- Step 4. Repeat steps 2 and 3 until $|\sigma^{(j+1)} - \sigma^{(j)}| < \epsilon$, for some small value ϵ .

When the algorithm stops at the j th iteration, we obtain an estimate $\sigma^{(j)}$ for σ and a set of estimated firm values $\mathbf{V} = \{V_0, \dots, V_n\}$. The drift μ can then be measured as the sample expected rate of return from the sample $\mathbf{V}$. A numerical example can be found in [8].

References

- [1] Merton, R. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.
- [2] Black, F. & Cox, J. (1976). Valuing corporate securities: some effects of bond indenture provisions, *Journal of Finance* **31**, 351–367.
- [3] Crosbe, P.J. & Bohn, J.R. (1993). *Modeling Default Risk*, Report by Moody’s KMV Corporation.
- [4] Leland, H.E. & Toft, K.B. (1996). Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads, *Journal of Finance* **51**, 987–1019.
- [5] Ronn, E.I. & Verma, A.K. (1986). Pricing risk-adjusted deposit insurance: an option-based model, *Journal of Finance* **41**, 871–895.
- [6] Ericsson, J. & Reneby, J. (2005). Estimating structural bond pricing models, *Journal of Business* **78**, 707–735.
- [7] Duan, J.C. (1994). Maximum likelihood estimation using price data of the derivative contract, *Mathematical Finance* **4**(2), 155–167.
- [8] Duan, J.C., Gauthier, G. & Simonato, J.G. (2004). *On the Equivalence of the KMV and Maximum Likelihood Methods for Structural Credit Risk Models*, Working paper, University of Toronto.

Further Reading

- Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.

Structural Reliability

Structural systems (*see* **Mathematics of Risk and Reliability: A Select History**) such as buildings, bridges, water tanks, containment vessels, and communication towers are often represented as load-resistance systems. In all cases, these structures represent substantial and expensive assets that are unique or “one-off”. Even apparently, similar buildings have different occupancy loads, foundations, connection details, etc. Clearly, their reliability cannot be directly inferred from observation of failures or other experimental studies. In these circumstances, reliabilities need to be predicted from predictive models and probabilistic methods. As such, there is a clear recognition that uncertainty and variability are associated with many variables describing a structure’s performance and that this can be accounted for explicitly by the use of probability distributions and structural reliability theory (e.g. [1]).

Structural reliability theory underpins many recent advances in structural and safety engineering, namely,

- reliability-based calibration of design codes in Europe, United States, Canada, Australia, and elsewhere (e.g. [2]);
- performance-based design of new structures such as the Confederation bridge (Canada), Great Belt bridge (Denmark), and Messina Strait bridge (e.g., [3]);
- service life and safety assessment of existing structures (e.g. [4]);
- optimal maintenance of aging or deteriorating structures (e.g., [5]).

Although the emphasis of the present chapter is on structural engineering systems, the computational and probabilistic methods described herein are also appropriate for other load-resistance or demand-capacity systems, such as geotechnical, mechanical, hydraulic, electrical, and electronic systems where performance failure is deemed to occur when a predicted load/demand exceeds a resistance/capacity.

Limit States

Limit state functions (*see* **Reliability of Large Systems**) are used to define “failure”, and the definition

of failure can vary from person to person. A limit state is a boundary between desired and undesired performance. Undesired performance may take on many characteristics. For example, structural collapse of a bridge girder could be because of exceedance of flexural, shear or bearing capacities, fatigue, corrosion, excessive deflections or vibrations, and other modes of failure. Each of these modes of failure should be considered separately and can be represented by a limit state function. In general, limit state functions of interest are related to the following:

1. ultimate strength limit state – reduction in structural capacity
 - (a) strength
 - (b) stability.
2. serviceability limit state – loss of function
 - (a) deflection
 - (b) vibration
 - (c) fatigue
 - (d) cracking.

Calculation of Structural Reliability

Simple Formulation

If the limit state of interest is related to structural capacity, then failure is deemed to occur when a load effect (S) exceeds structural resistance (R). The probability of failure (p_f) is

$$\begin{aligned} p_f &= \Pr(R \leq S) = \Pr(R - S \leq 0) \\ &= \Pr(G(R, S) \leq 0) \end{aligned} \quad (1)$$

where $G()$ is termed the *limit state function*, and in the present case this is equal to $R - S$. Thus, the probability of failure is the probability of exceeding the limit state function. Note that R and S must be in the same units. The formulation given by equation (1) also holds for other limit states; for example, for excessive deflection, the “load effect” S is replaced by the actual deflection (δ), and “resistance” R represents the allowable deflection limit (Δ_L), such that equation (1) is rewritten as

$$p_f = \Pr(\delta \geq \Delta_L) = \Pr(\Delta_L - \delta \leq 0) \quad (2)$$

The following discussion will use the classic R and S notation to illustrate the key concepts of structural reliability.

The strength of identical structural components will vary from component to component because of variabilities in material properties, geometric dimensions, environmental conditions, maintenance, etc. Similarly, loadings are influenced by a variety of factors, such as environment, temperature, geographic location (wind, earthquake), etc., which are time dependent and often highly variable. Consequently, resistance and loads (and their influencing variables) should be modeled as random variables.

For the simplest case with one random variable for load (S) and another for resistance (R), the probability of failure is given by the well-known “convolution” integral

$$p_f = \int_{-\infty}^{\infty} F_R(x) f_S(x) dx \quad (3)$$

where $f_S(x)$ is the probability density function of the load S and $F_R(x)$ is the cumulative probability density function of the resistance [$F_R(x)$ is the probability that $R \leq x$]; see Figure 1. Note that failure probability may be calculated per annum, per lifetime, or for any other time period. It follows that the probabilistic models selected for loads and resistances must relate to this time period. For example, the lifetime reliability for a communications tower requires that the probabilistic model of the maximum wind speed to be experienced over the lifetime of the structure is known.

For elementary cases, the integration of equation (3) is not difficult. For the special case where R and S are both normally distributed with means μ_R and μ_S and variances σ_R^2 and σ_S^2 , the probability of failure (see **Imprecise Reliability; Reliability**

Data) is expressed simply as

$$p_f = \Pr(R - S \leq 0) = \Phi\left(\frac{-(\mu_R - \mu_S)}{\sqrt{\sigma_S^2 + \sigma_R^2}}\right) = \Phi(-\beta) \quad (4)$$

where $\Phi()$ is the standard normal distribution function (zero mean, unit variance) extensively tabulated in statistics texts and β is the “reliability index” (see **Reliability Optimization; Quantitative Reliability Assessment of Electricity Supply**). In structural reliability literature, particularly reliability-based calibration of codes, the reliability index is often used as the measure of safety. The relationship between probability of failure p_f and reliability index β is shown in Table 1.

For example, let us assume that a steel bridge girder of length 15 m is loaded with an annual peak mean uniformly distributed load (w) of 16 kN/m. The mean flexural strength of the girder is 950 kNm with coefficient of variation (COV) of 0.1. The variability of loading is higher with a COV of 0.18. From elementary structural theory, the applied moment (load effect S) for a simply supported beam is $wL^2/8$. If self-weight of the beam is ignored, the structural reliability calculations based on equation (4) are as follows:

$$\begin{aligned} \text{Resistance (R):} \quad & \mu_R = 950 \text{ kNm} \\ & \sigma_R = 0.1 \times 950 = 95 \text{ kNm} \\ \text{Load Effect (S):} \quad & \mu_S = \frac{16 \times 15^2}{8} = 450 \text{ kNm} \end{aligned}$$

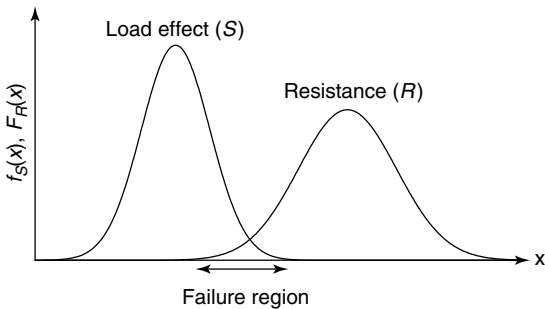


Figure 1 Basic $R - S$ problem showing $f_S(x)$ and $f_R(x)$

Table 1 Relationship between probability of failure p_f and reliability index β

p_f	β
10^{-1}	1.28
10^{-2}	2.33
10^{-3}	3.09
10^{-4}	3.71
10^{-5}	4.26
10^{-6}	4.75
10^{-7}	5.19
10^{-8}	5.62
10^{-9}	5.99
10^{-10}	6.36

$$\sigma_S = 0.18 \times 450 = 81 \text{ kNm}$$

Reliability Index:

$$\begin{aligned} \beta &= \frac{\mu_R - \mu_S}{\sqrt{\sigma_R^2 + \sigma_S^2}} \\ &= \frac{950 - 450}{\sqrt{95^2 + 81^2}} = 4.01 \end{aligned}$$

Probability of Failure: $p_f = \Phi(-\beta)$

$$= 3.05 \times 10^{-5} \text{ per year}$$

If corrosion on the bridge is observed with a 10% reduction in mean strength to 855 kNm and same statistics for loading, then $\beta = 3.44$ and the corresponding probability of failure increases by nearly 10-fold to 2.91×10^{-4} per year.

General Formulation

For many realistic problems, the simplified formulation given by equation (3) is not sufficient as the limit state function often contains more than two variables. Usually, several random variables will influence structural capacity, such as material properties, dimensions, model error, etc. Moreover, there are likely to be several load processes acting on the system at the same time. For example, these might be wind, wave, temperature, and pressure forces acting on an offshore structure. For instance, the limit state for a steel bridge girder can be expressed as

$$G(\mathbf{X}) = R - S = F_y Z - (D + L_p) \quad (5)$$

where the vector $\mathbf{X}$ represents the basic variables involved in the problem, in this case, F_y (yield strength) and Z (elastic modulus) representing the bridge girder resistance, and D and L_p representing the dead and peak vehicle live loads, respectively.

If the limit state function is expressed as $G(\mathbf{X})$, the generalized form of equation (3) becomes

$$p_f = \Pr[G(\mathbf{X}) \leq 0] = \int \dots \int_{G(\mathbf{X}) \leq 0} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \quad (6)$$

where $f_{\mathbf{X}}(\mathbf{x})$ is the joint probability density function for the n -dimensional vector $\mathbf{X} = \{X_1, \dots, X_n\}$ of random variables each representing a resistance random variable or a loading random variable acting on the system. Figure 2 shows a representation of a joint probability density function for two variables R and S , and the probability of limit state

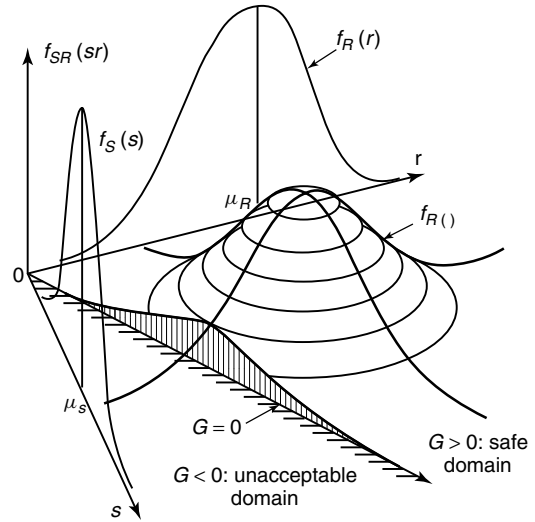


Figure 2 Region of integration for failure probability [Reproduced from [6]. © Chapman & Hall (The Thompson Group), 1997.]

exceedance (see **Bayesian Statistics in Quantitative Risk Assessment**).

The solution to equation (6) is complex, particularly because resistance and loading variables often vary in time and space (such as corrosion, fatigue, or other sources of deterioration), the variables may be correlated, and most infrastructure comprises many elements or components requiring a systems approach to infrastructure performance. Two main approaches can be taken to solve equation (6):

1. Analytical methods by transforming $f_{\mathbf{X}}(\mathbf{x})$ to a multinormal probability density function to approximate the failure region of the limit state function – second moment and transformation methods such as first-order second moment (FOSM) and first-order reliability methods (FORM) (see **Insurance Pricing/Nonlife**).
2. Numerical approximations to perform the multidimensional integration required in equation (6) – Monte Carlo simulation techniques (see **Ruin Probabilities: Computational Aspects; Simulation in Risk Management; Systems Reliability; Expert Elicitation for Risk Assessment**) that involve random sampling each random variable. The limit state $G(\mathbf{X})$ is then checked enabling p_f to be inferred from a large number of simulation runs.

These methods may be difficult to implement for all except trivial problems. For this reason, a large array of software packages are available – see Pellissetti and Schueller [7] for a review of available software.

All the methods are not of equal accuracy or applicable to all problems. Second moment and transformation methods are computationally very efficient and often very useful for most problems. They do have disadvantages, however, that nonlinear limit state functions are not easily handled and may give rise to inaccuracies, and difficulties can arise in using nonnormal random variables and dependencies. On the other hand, simulation methods are, in principle, very accurate, can handle any form of limit state function, and are not restricted to normal random variables. However, the computational times can be significant because of the high number of simulation runs often needed to produce convergent results. Nonetheless, with the availability of computers with ever increasing speed and the use of importance sampling, response surface methods, and other variance reduction techniques, the computational efficiency of simulation methods can be greatly improved.

Additional Capabilities of Reliability Analysis

The theoretical framework for structural reliability analysis is well established, and the formulation given by equation (6) can be broadened to encompass the following:

- structural systems comprising many components whose resistance or loading are correlated;
- stochastic processes
 - time-dependent process that arise from fatigue, corrosion, change in use, etc.;
 - spatial variability that arise from nonhomogeneous material, dimensional, or other properties;
- updating information from existing structures;
 - Bayes theorem;

- proof and service load information;
- acceptance criteria for new and existing structures.

It is beyond the scope of this article to describe structural reliability theory in detail. For an introductory text on fundamental probability and structural reliability theory and typical reliability-based applications, see Nowak and Collins [8]. More advanced structural reliability theory is described by Ditlevsen and Madsen [9] and Melchers [10].

References

- [1] ISO 2394 (1998). *General Principles on Reliability for Structures*, International Organization for Standards, Geneva.
- [2] Nowak, A.S. & Szersven, M.M. (2003). Calibration of design code for buildings (ACI 318): part 2 – reliability analysis and resistance factors, *ACI Structural Journal* **100**(3), 383–391.
- [3] MacGregor, J.G., Kennedy, D.J.L., Bartlett, F.M., Chernenko, D., Maes, M.A. & Dunaszegi, L. (1997). Design criteria and load and resistance factors for the confederation bridge, *Canadian Journal of Civil Engineering* **24**, 882–897.
- [4] Faber, M.H. (2000). Reliability based assessment of existing structures, *Progress in Structural Engineering and Mechanics* **2**(2), 247–253.
- [5] Stewart, M.G. (2006). Spatial variability of damage and expected maintenance costs for deteriorating RC structures, *Structure and Infrastructure Engineering* **2**(2), 79–90.
- [6] Stewart, M.G. & Melchers, R.E. (1997). *Probabilistic Risk Assessment of Engineering Systems*, Chapman & Hall, London.
- [7] Pellissetti, M.F. & Schueller, G.I. (2006). On general purpose software in structural reliability – an overview, *Structural Safety* **28**, 3–16.
- [8] Nowak, A.S. & Collins, K.R. (2000). *Reliability of Structures*, McGraw-Hill, Boston.
- [9] Ditlevsen, O. & Madsen, H.O. (1996). *Structural Reliability Methods*, John Wiley & Sons, Chichester.
- [10] Melchers, R.E. (1999). *Structural Reliability Analysis and Prediction*, John Wiley & Sons, Chichester.

MARK G. STEWART

Structured Products and Hybrid Securities

A “structured product” is a derivative financial security, whose cash flows are “structured” to depend on more basic financial values, so as to suit the parties to the arrangement. An example is a guaranteed equity bond (GEB), which might be issued with a lifetime of 5 years, at the end of which the principal invested is paid, plus say 90% of any gain, if the principal had been invested in the stock index over the 5 years and if this gain is positive. This product is structured to match the risk profile of an investor who wants to participate in any gains in the stock market but who does not want to risk losses. As is typical with structured products, the GEB is “hybrid” in nature, since it combines the upside potential of equity with the certainty of a bond.

In this survey, some of the main categories of structured products and the motivation for their design are described, and the valuation and risk management of these products are briefly discussed.

Structured Notes

A structured note allows an issuer to borrow at a very low rate, by including a structure that suits the risk appetite of the investor. Crabbe and Argilagos [1] give an overview of this market. For example, an investor might have the view that interest rates will fall over the next 5 years. A 5-year inverse floating rate note (IFRN) with interest payments being a fixed rate F minus 6-month LIBOR (the LONDON INTERBANK OFFERED RATE) allows him to benefit, if his view is correct. LIBOR is essentially the riskless rate for borrowing for a short period, which is 6 months in this case. LIBOR is set at the beginning of each 6-month interest period, and the investor believes that this rate will fall over the next 5 years.

The investor asks an investment bank to find a borrower who is prepared to issue such a note. An issuer might be able to issue straight debt with coupon of 8%, and if the 5-year swap rate is say 6%, then he will benefit if he can set $F = 13.9\%$ above. To explain this, swapping cash flows at the fixed swap rate, for LIBOR, is a fair (zero net present value)

transaction, and so the issuer should expect to pay $14\% (= 6 + 8\%)$ minus LIBOR with this structure. However, the investor is prepared to receive just 13.9%, since this structure allows him to implement his prediction on interest rates.

The issuer can reverse out of this structure, by separately entering a swap agreement to pay LIBOR and receive 6%. So the issuer is not in the end exposed to interest rate risk. Also, it should be noted that the investor will only benefit to the extent that he outpredicts the market itself, since the swap rate of 6% incorporates the markets view of the movements of LIBOR over the coming 5 years.

Many similar examples are given in Das [2]; for example, a 5-year note might make coupon payments every 6 months, based on the value of the 5-year CMT (constant maturity treasury) rate at each interest date. This CMT rate is a published benchmark rate for borrowing for 5 years. It would be more natural to have the interest rate determined by 6-month LIBOR in this note to match the interest period. This structure might be motivated as a play on the slope of the yield curve, as represented by the difference between the 6-month and the 5-year rates.

The commodity linked structured products market has grown significantly in recent years. Commodity linked notes provide investors a return that is linked to the performance of a basket of commodities. Typically these products are 100% principal protected; the return is linked to the positive return of the basket and investors are protected from negative returns. The underlying baskets for these products can be quite varied – the basket may contain just a few key energies and metals, for example crude oil, natural gas, aluminum, copper, nickel; or they may be a very broad basket of soft commodities such as wheat, corn, soybeans, cotton, sugar, coffee, cocoa, cattle, and hogs.

Perhaps the most notorious structured product was that agreed upon between Bankers Trust (BT) and Proctor and Gamble (P&G), in November 1993. This has been analyzed in detail in Smith [3], and the details have become public only because it became the subject of legal action involving these counterparties.

Basically, the arrangement was agreed to run for 5 years with a principal of US\$200 million, and involved a “swap” and a “spread”. Under the swap, P&G would pay to BT every 6 months the commercial paper (CP) rate minus 75 basis points

(i.e., 0.75%), and receive a fixed rate of 5.3%. The CP rate is a floating interest rate that high-grade borrowers might expect to pay. Assuming that P&G could have borrowed at 5.3% over this 5-year period, this swap is advantageous to them, with the value given by the 75 basis point interest adjustment for 5 years. Bundled with P&G borrowing at a fixed rate, this arrangement is equivalent to a structured note (see **Statistical Arbitrage**).

The spread was given by the formula $\max \left\{ 0, (98.5 \times \frac{5\text{-year CMT\%}}{5.78\%} - (30\text{-year TSY}))/100 \right\}$, in which (30-year TSY) is the price of the treasury bond with a coupon of 6.25% and maturity of August 2023. P&G would pay this every 6 months, starting from the first year, i.e., November 1994, although its value would be fixed according to rates at 6 months after issue, i.e., in May 1994.

This spread can be understood as an option written by P&G, equivalent to a bet that interest rates would not rise, since they would have to pay if either the 5-year rate or the 30-year rate rose. Their premium for this resides in their favorable terms on the swap.

What made this deal so notorious was that it was so aggressive, and blew up so spectacularly. Interest rates rose before May 1994, with the 5-year CMT rate rising from 5.02 to 6.71%, and the 30-year bond yield rising from 6.06 to 7.35% (price falling from \$102.57811 to \$86.84375). Putting these numbers into the formula, P&G has to make interest payments of 27.50% for $4\frac{1}{2}$ years, on a principal of \$200 million, putting them underwater by about \$200 million. P&G initiated legal action against BT but they settled out of court. Arguably, this episode fatally wounded the reputation of BT, and they were taken over by Deutsche Bank in 1999. Smith subjects this deal to a rigorous evaluation, and concludes that, apart from the exposure entailed, the value of the swap was not enough to compensate P&G for the option they were writing.

Callable and Convertible Bonds, US Style Prepayable Mortgages

A convertible bond (CB) is issued by a company to raise capital, and it is a bond, but the holder has the option to convert the bond into new equity, which will then be issued by the company, at a ratio that is prespecified in the CB contract.

With structured securities, often the question of motivation for creating the security is a challenge, and with respect to the CB this has been given a fascinating answer by finance theorists. At first glance, a CB offers cheap financing: a firm might have a share price of \$70, and the ability to issue straight bonds with coupon of 8%. Such a firm might be able to issue CBs with a coupon of 4%, convertible into shares corresponding to a price of \$100, i.e., if the CB has principal of \$1000, then it is convertible into 10 shares (see Brennan and Schwartz [4]). Thus, the company is paying a low interest rate if the bonds are not converted, and it has issued equity at a premium, if the bonds are converted. However, this way of reasoning ignores the value of the optionality in the CB. This CB can be more accurately understood as being 50% bond and 50% an option to buy 10 shares for \$100 each. This option has value, although it is initially out of the money, i.e., it would not be rational to exercise the option immediately.

Current thinking is that CBs are in general correctly priced in the market, and they allow the firm to issue bonds, but to retain flexibility over its business strategy (see Green [5]). Such flexibility in conjunction with issuing a straight (i.e., not convertible) bond could lead to an “asset substitution problem” (see Jensen and Meckling [6]). If the firm issues straight debt and then chooses a more risky investment strategy, the benefits arising from success will accrue mostly to equity, because payments to debt are limited, but the losses arising from failure will fall disproportionately on the debt, if the firm becomes bankrupt, since the debt might not be paid back. Anticipating these outcomes, choosing a more risky strategy will then have the immediate effect of increasing the equity value at the expense of the debt value. And anticipating that the firm might increase the riskiness of their investment, the market will in any case demand a higher interest rate on the debt. Issuing a CB can avoid this problem, because the CB can participate in the upside potential of the investment, *via* being converted. This line of reasoning can explain why CBs are mostly issued by high-risk, high-growth firms.

Corporate bonds are sometimes “callable”, i.e., they can be redeemed before maturity (“called”) by the issuer, at a prespecified price, often their par value. In the past few years, the rationale for this feature might have been to allow for the orderly

retirement of the debt, but this rationale is increasingly moot as bond markets become more liquid, so that companies can buy their bonds in the market without causing any price distortions. A more current rationale might be to allow the issuer to benefit if interest rates fall, in which case the issuer can call the bonds, and concurrently issue more bonds at a lower rate.

A variation of the CB is the liquid yield option note (LYON), which is puttable by the investor, and callable by the issuer, as well as being convertible. McConnell and Schwartz [7] analyze this structure in detail, and show that some actual issues were reasonably priced in the market at initiation.

US style mortgages are fixed rate, and the mortgagee has the option to “prepay”, i.e., to pay off the mortgage at any time, by paying the amount outstanding according to the amortization schedule. If it is considered that taking out a mortgage is similar to issuing a bond (both are borrowing money), then the prepayment option is similar to the call feature on a bond. Also, a similar rationale applies, in that the mortgagee can prepay and refinance, if the interest rate falls. Other reasons for prepaying a mortgage include selling the mortgaged property or having a windfall gain.

Structured Retail Products

Equity based products that are structured to appeal to retail investors are broadly termed equity linked notes (ELNs). Das [2] categorizes these into principal protection vehicles and yield enhancement vehicles. The GEB mentioned above is an example the former. Das [2, pp. 526–527] gives an example of the latter. This is a 1-year note with a high coupon of 16.5%, but with maturity payment given by the principal, minus the gain that the principal would have made if it had been invested in the Nikkei 225 index, if this is positive. Thus, the investor has sold a call on the Nikkei index, and is being paid for this in the shape of the high coupon.

Many variations on these structures are possible. For example, an ELN might pay the return earned by investing a certain principal in the stock index, but with the return capped at some limit. In return for accepting a limited upside potential, the investor can buy the ELN at a discount from the principal. ELNs following these designs now trade on the Chicago Board Options Exchange and other exchanges.

An interesting structured retail product is the Swing Guaranteed Fund, designed by Societe Generale and issued in September 2002. This makes dividend payments at 1.0, 2.0, 3.0, and 4.5 years, based on a globally diversified portfolio of 18 stocks, and also pays back the principal invested after 4.5 years. The dividends are proportional to the minimum absolute return of the stocks in the portfolio, over each dividend period. Thus, the investor will benefit if the global markets are volatile, whether they rise or fall. Similarly complex products, involving many equities, are described by Quessette [8] and Overhaus [9].

The following discussion shows that valuation and risk management of such structures is more challenging than all the structures mentioned previously, because this depends on the individual movements of many prices.

Valuation and Risk Management

Many structured products are easy to separate into standard products. For example, the GEB discussed above can be decomposed into a bond, and a call option: if say \$100 is invested in the GEB, the bond pays \$100 at the 5-year maturity and no coupons, and the option pays $\$90 \times \max \left\{ \frac{S_t - S_0}{S_0}, 0 \right\}$, where S_t is the value of the index at time t . The \$100 principal here can be decomposed into the prices of the bond, and of the option, plus the profits received by the issuer, and the “participation rate” of 90% will be fixed, so that these do not exceed \$100.

For pricing and risk management of this structure, the main challenge is the option component. Since the maturity of 5 years is very long, there is no liquid market for such an option, and so the issuer cannot simply buy this option in the market, and must “synthesize” it *via* dynamically maintaining a position in the underlying index or corresponding futures contract, and borrowing/lending (see Hull [10]). This strategy might naively be done using the Black and Scholes [11] framework, but this will not be sufficiently accurate, because of the very long maturity. The issuer will have to synthesize the option in a manner that takes into account the interest rate risk, volatility risk, and the “smile effect” (see **Volatility Smile**), which is a well-recognized distortion in equity option markets.

Valuing and hedging (risk management), generally in financial markets, which can include equities, interest rates, currencies and commodities, is presented in a unified framework by Duffie [12]. This framework can be regarded as an extension of the Black–Scholes model, and similar partial differential equations (PDEs) hold to value derivatives, with each source of risk being a variable in the equation. To solve the PDE, the modeler might use an analytic formula, or a numerical procedure, or a Monte Carlo (MC) simulation. Of these, MC is the most cumbersome, but it is the only feasible one if there are more than about three-state variables.

Within this framework, interest rate modeling presents its own challenges, because there are many interest rates, and these have to be reduced to a manageable number of state variables. See Ho and Lee [13] for a survey of interest rate models. To price and hedge P&Gs structure, a two-factor interest rate model would probably be adequate, calibrated to the 5- and 30-year maturities, and the model would also enable the user to hedge the risks *via* a dynamic strategy using say the 5-year T Note and the T Bond futures contracts. To deal with a callable bond or prepayable mortgage, a one-factor model might be adequate, and to deal with a CB, an extra factor, representing the underlying stock value, should be included.

Of all the structures described above, the Swing Guaranteed Fund is perhaps the most difficult to price and hedge. This is because the underlying stocks represent 18 state variables, and these may not be reduced to a smaller number. MC simulation must be employed, and to hedge this structure, the modeler must maintain dynamic positions corresponding to the 18 stocks.

References

- [1] Crabbe, L.E. & Argilagos, J.D. (1994). Anatomy of the structured note market, *Journal of Applied Corporate Finance* 7(3), 85–98.

- [2] Das, S. (2001). *Structured Products and Hybrid Securities*, 2nd Edition, John Wiley & Sons, New York.
- [3] Smith, D.J. (1997). Aggressive corporate finance: a close look at the proctor and gamble & bankers trust leveraged swap, *Journal of Derivatives* 4(4), 67–79.
- [4] Brennan, M. & Schwartz, E.S. (1986). The case for convertibles, in *The Revolution in Corporate Finance*, J. Stern & D. Chew Jr, eds, Basil Blackwell, New York.
- [5] Green, R. (1984). Investment incentives, debt and warrants, *Journal of Financial Economics* 14, 115–136.
- [6] Jensen, M. & Meckling, W. (1976). Theory of the firm: managerial behavior, agency costs and ownership structure, *Journal of Financial Economics* 3(4), 305–360.
- [7] McConnell, J.J. & Schwartz, E.S. (1986). LYON taming, *Journal of Finance* 41(2), 561–577.
- [8] Quessette, R. (2002). New products, new risks, *RISK Magazine* 15(3), 97–100.
- [9] Overhaus, M. (2002). Himalaya options, *RISK Magazine* 15(3), 101–103.
- [10] Hull, J. (2005). *Fundamentals of Futures and Options Markets*, 5th Edition, Pearson, Prentice Hall.
- [11] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* 81(3), 637–654.
- [12] Duffie, D. (2001). *Dynamic Asset Pricing*, 3rd Edition, Princeton University Press.
- [13] Ho, T.S.-Y. & Lee, S.-B. (2004). *The Oxford Guide to Financial Modelling*, Oxford University Press.

Related Articles

Credit Migration Matrices

Model Risk

Risk in Credit Granting and Lending Decisions: Credit Scoring

ANDREW CARVERHILL

Subjective Expected Utility

For choosing a course of action, it is paramount that a “score” is assigned to each possible alternative, if not explicitly, at least implicitly. *Subjective expected utility* aims at determining how such a representation of our choices is possible, given the type and structure of our preferences. The earliest attempts started with the least difficult case and assumed that some probability distribution of the possible consequences existed beforehand. The “objective” or “frequentist” concepts of probability were invoked [1, 2]. We refer below briefly to these earliest attempts, for the concepts ironed out there have conditioned the later developments of SEUs and also because they are often of practical interest. Yet, the first *genuine* SEU generalizations [3–5] allowed to depart from situations of risk (i.e., known probabilities) and to face situations of uncertainty.

This article follows the above development path in the four first sections and ends with a discussion and some more recent advances in the last section.

Early Contributions

Less than 60 years after Pascal had formally expressed *expected gain* as a rule of evaluation in risky issues, Nicolas Bernoulli spread all over Europe the word, in the letter dated September 13, 1709, that Pascal’s rule was being violated by the subjects he had tested at St Petersburg’s imperial court: to participate in a parlor game, the expected gain of which was infinite, members of the court did not want to pay more than a few gold coins (“St Petersburg paradox”). Many solutions were offered in the late 1720s and until 1777, but the one that has been postulated by Daniel Bernoulli has been popular: when amounts at stake are large with respect to the decision maker’s assets, gains and assets should not be evaluated in plain money terms, but by a psychological appraisal of the accrual to the assets, named *emolumentum* [1] and later translated as *utility* (see **Utility Function**). The “paradox” was solved and a first version of the *expected utility* rule was postulated.

When looking for an evaluation of lotteries in a risky world, von Neumann (von Neumann and Morgenstern (henceforth vNM)) exhumed Bernoulli’s work, but gave it a more formal foundation [2]. In what follows, the term *prospect* (or *lottery with a finite number of consequences*) is taken to be the formal representation of a risky bet, including in technology, marketing, etc. von Neumann and Morgenstern argued that these prospects are inscribed in a “mixture space”, meaning that, from two lotteries A and B, it is always possible to construct a third one, written as $\alpha A + (1 - \alpha) B$, which we shall call a “convex combination” of lottery A and lottery B. The combined lottery belongs, in the probability space (simplex), to the segment of line linking the two points representing A and B. The virtue of assuming a mixture space is twofold: (a) In the simplex, the set of lotteries is convex and (b) if one lottery has as an outcome a ticket opening the possibility to play another lottery, this “two-stage” lottery can be represented by a “one-stage” lottery (“compounding lotteries”). vNM’s three axioms can now be presented as follows:

vNM1 (weak ordering)

The individual’s preferences can be represented by a complete, reflexive, and transitive relation between all lotteries, denoted by $\succsim$ and read as “not preferred or indifferent to”. If some engineer, for example, does not judge some risky technique T° as preferable to another one, say T^* , while seeing it as preferable to T , then $T \succsim T^\circ \succsim T^*$ and this implies $T \succsim T^*$ (transitivity). In proper mathematical terms, this binary relation is a complete preorder.

The next question refers to the existence of *indifference loci* of lotteries. Such an existence is postulated through a peculiar *continuity* property of preferences (axiom vNM2).

vNM2 (Archimedean property)

For any three lotteries such as L_1, L_2, L_3 , s.t. $L_1 \prec L_2 \prec L_3$, there always exists a real number λ between 0 and 1, such that $L_2 [\lambda L_3 + (1 - \lambda)L_1]$. In other words, a lottery like L_2 is always indifferent to some convex combination of a preferred lottery like L_3 and of a less preferred lottery like L_1 . Of course, the numerical value of λ depends not only on L_3 and L_1 but also on the *behavior toward risk* of the individual (see **Risk Attitude**).

Under these hypotheses and axioms, it can be shown that some real-valued index (a score function, defined up to any monotonically increasing transformation) $U(\cdot)$ exists on the set of lotteries, such that, from the point of view of the individual

$$L_1 \preceq L_2 \iff U(L_1) \leq U(L_2) \quad (1)$$

We have to make this result operational and find some way to determine the numerical values of $U(\cdot)$ i.e., to “encode” the U function, through some additional restrictions. Axiom vNM3 does the job.

vNM3 (independence)

Given any lottery M , and two lotteries L_1 and L_2 , such that $L_1 \preceq L_2$, any $\alpha \in [0, 1]$ will yield

$$\alpha L_1 + (1 - \alpha)M \preceq \alpha L_2 + (1 - \alpha)M \quad (2)$$

vNM3 can be interpreted as a separability axiom, for one can see that the outcomes of M do not play any role in equation (2). It has turned to be the most controversial axiom of all three above and is at the root of Allais’s paradox. It is closely related to the “sure thing principle” formulated by Savage (see the section titled “Savage’s Contribution”). vNM’s theorem then reads as follows [2].

Consider any lottery $L = (x_1, p_1; \dots; x_n, p_n)$. Denote by $u(x_i)(i = 1, \dots, n)$ the restriction of $U(\cdot)$ to degenerate lotteries like $(x_1, 1), \dots, (x_i, 1), \dots, (x_n, 1)$. Then, under vNM1–vNM3, whatever the feasible set of lotteries, the individual chooses the lottery L for which

$$U(L) = \sum_{i=1}^n p_i \cdot u(x_i) = E[u(\tilde{x})] \quad (3)$$

takes its maximal value in the feasible set. The $u(\cdot)$ function is called a *vNM utility function*, and is defined up to a positive affine transformation. This last point is essential: Neumannian expected utility (EU) scores are not “measures” in the strong mathematical sense (like absolute temperature in physics) but are *interval scales* (like the Fahrenheit or Celsius temperature scales). This plays an important role in management applications to reach some economic efficiency. Indeed, ratios of EU differences are well-defined numbers, which allows for an easy optimal allocation of resources among potential actions (including in enterprise risk management; (see **Enterprise Risk Management (ERM)**)).

Applications to *portfolio theory* have developed since the 1950s through a particular specification of the utility function (see **Utility Function**): the quadratic utility. It is then straightforward to show that EU is a function of $E(\tilde{R})$ and of $\sigma(\tilde{R})$, where $\tilde{R}$ denotes the return of financial shares [6]. Since the late 1970s, experimental methods have been considerably improved to encode vNM utility functions through individual elicitation methods [7]. This has considerably developed potential applications in *industrial management*, for which objective probabilities rarely exist.

Savage’s Contribution

Early in the nineteenth century, Keynes [8] and Knight [9] had forcefully argued that the *kind* of uncertainty facing *entrepreneurs* was different from the one facing a roulette gambler. Ramsey [10] expressed the idea that probabilities could be inferred from watching an agent’s behavior, while de Finetti [3] provided us with an operational theorem to do so, and with a definition of subjective probability. Assume that I own some 20Mio\$, and I am receiving an additional sum of 1000\$ if my horse wins the next *Grand Prix*. By then, my subjective probability of my horse winning is my *ex ante* conditional price of receiving the 1000\$.

Savage [4] achieved a synthesis between vNM and the above quoted authors, with all the weight of the difficulty put on the axioms of *individual* rationality. Then, let S be the (infinite) set of *states of the world* s (a state of the world entails everything that might have an impact on the consequences of an act, conditional on the state of the world that is obtained, and is thus a complete resolution of uncertainty). *Events* are subsets of S and C is the set of consequences. We now define an *act* as a *function* $f(s)$ from the set of states S to the set of consequences C , with no probability on S . The set of acts F is the set of all possible $f(s)$.

When we consider two different acts, say $f(s)$ and $g(s)$, and some event $E \subset S$, we can always think of a new act $f_E g$ with the following definition:

$$\text{For all } s \in E, f_E g(s) = f(s) \quad (4)$$

$$\text{For all } s \notin E, f_E g(s) = g(s) \quad (5)$$

For example, E is the event “horse H wins”. If the act $f(\cdot)$ is “betting on H” and the act $g(\cdot)$ is “betting

on horse K”, then $f(s)$ would yield the best prize only if $s \in E$, $g(s)$ could do so only for some state of the worlds such that $s \notin E$, while $f_E g(s)$ would do like $f(s)$ if $s \in E$ but also like $g(s)$ if $s \notin E$.

A null event E° is then an event that is so unlikely that it would not make any difference between $f_E g(\cdot)$ and $f'_{E'} g(\cdot)$, for all f, f' . Clearly, the empty set $\emptyset$ is a null event.

A constant act $\chi(c)$ is an act that assigns the same consequence, namely c , to every state of the world.

Savage’s contribution [4] has been to show that, under the conditions below described by postulates, the following holds:

(CS1) Under $P1$ – $P5$, there exists on the set of all possible events a transitive and complete binary relation $\geq$, which can read “at least as likely to obtain as”. This binary relation is a qualitative probability, for it preserves the additivity property, i.e.,

$$E \geq E' \iff E \cup D \geq E' \cup D, \quad \text{provided } E \cap D = E' \cap D = \emptyset \quad (6)$$

(CS2) By the continuity properties implied by the axioms below (in particular, by axiom $P6$), the qualitative probability relation of (CS1) becomes a unique subjective probability distribution.

(CS3) Preferences on acts generate preferences over probability distributions that are analogous to vNM preferences under risk, which were examined in the introductory section. Such preferences lead to the expected utility rule of rational choice under risk.

Note that (CS1) is intimately related to a previous work by de Finetti [3] and that (CS3) pertains to the work by von Neumann [2]. Part (CS2) is therefore the great achievement of Savage’s contribution, which provided the theory of subjective probability (see **Subjective Probability**) with the first strong foundation in the history of thought since subjective probabilities were first evoked [11].

By skipping some purely technical conditions here, we can formulate and give intuitive meaning to Savage’s axioms in our own way as under.

P1 (weak ordering on acts)

The individual we consider here has a complete, reflexive, and transitive binary relation $\succsim$ on acts.

P2 (sure thing principle)

For all acts f, f', h , and h' and every event E :

$$f_E h \succsim f'_E h \iff f_E h' \succsim f'_E h' \quad (7)$$

This amounts to saying that the modification that changes f into f' (which is obviously the same on the right- and the left-hand side of equation (6) and which impacts only the consequences of $f_E h$ on event E , consequences that change into the ones of $f'_E h$) explains *all* the reasons why $f_E h \succsim f'_E h$. Thus, there seems to be an argument that the valuation of the consequences of any act for some event E is independent of the valuation of the consequences of that act for the complementary event(s), which refers to mathematical separability. The reader could compare this with vNM3 (*independence axiom*) above and conclude that the Sure Thing Principle of Savage implies logically in vNM. The reader should compare with vNM3 (*independence axiom*) above and admit that the sure thing principle in Savage implies the independence axiom in vNM. As for practical implications, let us remark that $P2$ is a sufficient condition to justify that we may “roll back” a decision tree.

This postulate has, nevertheless, triggered persistent criticisms, following *Allais’ paradox* (see [12]). *Ellsberg’s paradox* can also be presented as a case of violation of this postulate [13].

P3 (conditional preferences)

For every nonnull event E and for all constant acts $\chi(c)$ and $\chi(c')$:

$$\chi(c)_E f \succsim \chi(c')_E f \iff c \succsim c' \quad (8)$$

This axiom embodies several ideas. In equation (8) we mean that the individual has preferences “on the consequences of E ”, i.e., *conditional preferences*, because, in the light of $P2$ above, the preference on constant acts can be associated to the preference “if E obtains”, E being nonnull. We also mean that this preference on constant acts is equivalent to a preference on consequences, which can be appraised *under certainty* as well, namely, c and c' , as expressed by the right-hand side of equation (8). This last point has been questioned by the authors of more general theories (state-dependent preferences), for it implies that the rank order of the consequences is independent from the underlying event [14]. However, $P3$ quite clearly suggests that if an act has for

an event E consequences that are in the view of the decision maker preferable than the ones in the E^c , selecting that act should be interpreted as a “bet on E ”, i.e., E is seen as more likely than E^c . Such a bet should neither depend on the absolute nor on the relative levels of c and c' . This is paramount to saying that preferences on acts induce judgments on the probabilities of events.

P4 (independence of tastes and beliefs)

For all events E and E' and constant acts $\chi(c)$, $\chi(d)$, $\chi(c')$, $\chi(d')$, with $c > d$, $c' > d'$:

$$\begin{aligned} \chi(c)_E \chi(d) &\succeq \chi(c)_E \chi(d) \\ \iff \chi(c')_E \chi(d') &\succeq \chi(c')_E \chi(d') \end{aligned} \quad (9)$$

Again, the authors of state-dependent preferences do not accept that axiom as is, because it implies that risk attitude (see **Risk Attitude**) is state independent. But *P4* clearly implies that tastes are independent of beliefs.

P5 (nontriviality)

For some constant act $\chi(c)$, there is at least one c' such that

$$\chi(c') > \chi(c) \text{ or } \chi(c) > \chi(c') \quad (10)$$

Under *P1–P5*, part (CS1) of Savage’s contribution holds. Going beyond this qualitative result requires some continuity condition on the preferences. It would seem reasonable to postulate that if two acts are very close to one another, they should be ranked the same with respect to a third act and that there exist partitions of events into as many null events as we want, which will rule out any “atomicity” (concentration on one point) of the probability measure on events. Formally,

P6 (continuity of preferences)

For all acts f , g , and h with $f > g$, there exists a finite partition $\{E_i, i = 1 \dots n\}$ of the set of states such that, for all i , $f >_{E_i} h$ and $h_{E_i} f > g$.

Under *P1–P6*, there exists a nonatomic probability measure on the set of states, which is defined on the set of events, i.e., finitely additive. This is the (CS2) part of Savage’s work, the most important one, as already mentioned.

Finally, Savage requires just one more axiom to confirm (under a much more general settings,

admittedly) the von Neumann–Morgenstern results of the introductory section. This last axiom says that, if we strictly prefer some act f to some other act g , conditionally to some event E and whatever the consequence $g(s)$ the act g yields in E , then we should strictly prefer act f to act g conditionally to event E . Formally,

P7 (conditional dominance)

For every event E and all acts f and f' , if $f >_E g(s)$ for all s in E , then $f >_E g$.

Under *P1–P7*, Savage’s second main theorem proves that the individual should maximize the expectation of a utility function, namely, a real-valued function defined on the set of consequences, with respect to the *unique* probability measure established as a result of *P1–P6* (see above). The utility function is unique up to a positive affine transformation and bounded. Formally, for all acts f and g

$$f \succeq g \iff \int_S u[f(s)] dp(s) \succeq \int_S u[g(s)] dp(s) \quad (11)$$

A consequence of Savage’s results is that the distinction between risk and uncertainty loses its meaning very much under *P1–P6*. Yet, the framework described above is quite complex. This is why Anscombe and Aumann provided us with a different setting, leading to a less complex treatment [5].

Anscombe and Aumann’s Contribution

Here we follow the notations of the section titled “Early Contributions” unless otherwise stated. Anscombe and Aumann [5] distinguish between “horse lotteries” (h-lotteries), which correspond to a case of uncertainty, and “roulette lotteries” (r-lotteries), which correspond to risky situations. Under their representation, the amounts at stake when betting on horses are not money prizes, but rather tickets to some r-lotteries. Thus, an act in the Savagian sense above is here a function from S (set of states of the world) into the set of r-lotteries, say L . Note that S is a finite set here. Subjective probabilities of h-lotteries will then result from observed choices’ comparisons between lotteries that are compounded from h- and r-lotteries. To ease matters, the axiom Anscombe and Aumann iron out is their “reversal of order in compound lotteries”, which is akin to the vNM “mixture

space” hypothesis *extended to h-lotteries*. We may therefore mix two acts f and g as we have mixed above two vNM lotteries (here: r-lotteries), i.e., $\alpha f + (1 - \alpha)g \in F$, with the definition that, whatever the possible outcomes $f(s)$ and $g(s)$, $[\alpha f + (1 - \alpha)g](s) = \alpha f(s) + (1 - \alpha)g(s)$. F is therefore convex and we essentially return to the kind of question solved by vNM. Aumann and Anscombe considerably simplify Savage’s framework, but at substantial expense.

Discussion and Conclusion

Many readers may find that the mathematical apparel necessary to define subjective probabilities as resulting from observed (real or experimental) bets is very heavy. This follows from the fact that the authors above want their base axioms to (a) make no reference to any of the collective links (institutional, psychosociological, etc.) between individual’s rationalities and (b) bear on sets of acts, states of the world and consequences, all endowed with the lightest possible structure, as in Savage, so that general conclusions may ensue. The second aspect is the source of an endless quest of mathematical psychology research; the first has been less dealt with by social mathematics research [3, 15, 16]. It allows less restrictive assumptions on individual rationality. Perhaps the continuity postulate $P6$ should be softened or perhaps even dispensed with, the present form being quite extreme [17].

The criticism of $P3$ and $P4$ with respect to state-dependent preferences has been alluded to in the text above. The sure thing principle and the independence axiom have been heavily criticized, notably in the experimental literature, with the result that alternative utility and decision models have been suggested. The impact of these generalizations on the very question of subjective probability is only beginning to be understood.

References

- [1] Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis, *Commentarii Academiae Scientiarum Imperialis Petropolitanae* **5**, 175–192. (Translated as Exposition of a new theory on the measurement of risk, *Econometrica* **22**(1954), 23–36).
- [2] von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*, Princeton University Press.
- [3] de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives, *Annales de l’Institut Poincaré* **7**, 1–68. Translated in H.E. Kyburg & H.E. Smoklers (eds) (1964). *Studies in Subjective Probabilities*, John Wiley & Sons, New York.
- [4] Savage, L.J. (1954). *The Foundations of Statistics*, John Wiley & Sons, New York.
- [5] Anscombe, F.J. & Aumann, R.J. (1963). A definition of subjective probability, *The Annals of Mathematics and Statistics* **34**, 199–205.
- [6] Markowitz, H.M. (1952). Portfolio selection, *The Journal of Finance* **7**(2), 77–91.
- [7] Abdellaoui, M. (2000). Parameter free elicitation of utility and probability weighting functions, *Management Science* **46**, 1497–1512.
- [8] Keynes, J.M. (1921). *A Treatise on Probability*, Macmillan, London.
- [9] Knight, F. (1921). *Risk, Uncertainty and Profit*, Kelley, Boston.
- [10] Ramsey, F.P. (1931). Truth and probability, in *The Foundations of Mathematics and Other Logical Essays*, R.B. Braithwaite & F. Plumpton, eds, Paul, Trench, Trubner, London.
- [11] Bernoulli, J. (1713). *Ars Conjectandi*, Impensis Thurnisiorum Fratrum, Basel.
- [12] Allais, M. (1953). Critique des postulats et axiomes de l’Ecole Américaine, *Econometrica* **21**, 503–546.
- [13] Ellsberg, D. (1961). Risk, ambiguity and the savage axioms, *Quarterly Journal of Economics* **75**, 643–659.
- [14] Karni, E. (2007). State-dependent utility, in *The Handbook of Rational and Social Choice*, P. Pattanaik, C. Puppe & P. Anand, eds, Oxford University Press.
- [15] Munier, B.R. (1991). Market uncertainty and the process of belief formation, *Journal of Risk and Uncertainty* **4**, 233–250.
- [16] Nau, R.J. (1999). Arbitrage, incomplete models and other people’s brains, in *Beliefs, Interactions and Preferences in Decision Making*, M.J. Machina & B.R. Munier, eds, Kluwer Academic Publishers, Dordrecht/Boston, pp. 217–236.
- [17] Munier, B.R. (1995). Complexity and strategic decision under uncertainty: How can we adapt the theory? in *Markets, Risk and Money*, B.R. Munier, ed, Kluwer Academic Publishers, Boston/Dordrecht, pp. 209–232.

BERTRAND R. MUNIER

Subjective Probability

The fundamental reality that motivates interest in any form of probability is variability. There is variability over instances, space, and time. As a consequence, uncertainty is experienced: a lack of surety in the face of variability. Although uncertain, it is intuitively recognized that some things are more uncertain than others and variability itself is variable. That uncertainty has degrees also supports the attempts at its measurement. Thus, the overarching definition of subjective probability is that it is a measure that captures a degree of belief, a degree of judged certainty. Surrounding this general notion are several major variant approaches to subjective probability.

The first approach is to treat subjective probabilities as an interpretation of mathematical probability theory. Thus, the section titled “Subjective Probability as an Interpretation of Mathematical Probability Theory” of this article will begin with probability theory as a mathematical system and derive subjective probability as an interpretation of a correspondence between this theory and reality. Subjective probabilities, being personal, need to come from the person. There are two general variants for assessing these degrees of belief: the direct expression of likelihoods and the inference of likelihoods from expressed preferences. The section titled “Subjective Probabilities as Assessed Degrees of Belief” derives these variants from an axiomatic system for a qualitative probability structure designed to connect probability more directly to the underlying behavior it is intended to represent. In the section titled “Probabilities as Inherently Subjective” a broader argument is presented as a second general approach to subjective probability. The claim is that subjective probability is not just one of many interpretations, but, rather, subjectivity is an inherent aspect of all uses of probability, i.e. all probability statements are subjective. Bayesian inference is described as an outgrowth of this approach (*see* **Bayesian Statistics in Quantitative Risk Assessment**). The section titled “Justification-Based Measures of Subjective Probability” describes a third general approach to subjective probability, one that disconnects from traditional mathematical probability theory. Treated as a measure of a degree of belief or of subjective uncertainty, it is noted that uncertainty has not always been seen as a unitary construct. Specifically, from the early

days of probability theory, a distinction has been claimed between uncertainty arising from the weight of evidence and uncertainty arising from the balance of evidence. Dempster–Shafer degrees of belief are outlined to exemplify meaningful measures of uncertainty that are distinct from probability theory.

A good resource for many of the topics and issues raised in this article is the book edited by Wright and Ayton [1]. The book provides a good next step for those interested in pursuing subjective probabilities further.

Subjective Probability as an Interpretation of Mathematical Probability Theory

Mathematical Probability Theory

Probability theory is mathematical. Like all branches of mathematics, e.g., algebra or geometry, the theory is an abstract, self-contained representation, with no necessary connection to the reality in which we live. However, although this theory can be viewed separately from reality, like other mathematical theories, probability theory derives its strength and interest from its correspondence to reality that can be formed. In that we can draw a parallel between elements of geometry (points, lines, and planes) and elements of the real world (objects, edges, and surfaces), geometry is critical to good architecture and construction. Before addressing the correspondences for probability theory, the theory itself needs to be outlined.

As a mathematical theory, probability theory starts with basic principles and definitions and, from these, other laws and theorems are derived. Kolmogorov [2] is generally credited with the mathematical formulation of modern probability theory. He defined the field of probability as a system of sets and an assignment of numbers that satisfy the following axioms:

1. E is a collection of all elementary events [the complete set of possibilities].
2. All subsets of E , and their sums $[A + B]$, i.e., A or B , products $[AB]$, i.e., A and B and differences $[A - B]$, i.e., A but not B are defined.
3. Each subset A can be assigned a probability $P(A) \geq 0$. [Sets the lower bound at 0, which gets assigned to the impossible event.]
4. $P(E) = 1$. [Sets the upper bound at 1; that one of the elementary events will occur is presumed certain.]

5. If A and B are mutually exclusive (have no elementary events in common), then $P(A + B) = P(A) + P(B)$.

These axioms, along with the definitions of a conditional probability [probability of A given B , i.e., conditional on the assumption that B occurs for certain]:

$$P(A | B) = \frac{P(AB)}{P(B)} \quad (1)$$

and of independent events:

A , B are independent if $P(AB) = P(A) \times P(B)$

are the key specifications of mathematical probability theory.

Views of Theory Correspondence

How then to apply this theory in the real world? First two nonsubjective interpretations of probability are presented to provide the context for a third, subjective view that is the focus of this article.

Classical View. *What is the probability of a coin turning up heads in a single toss?*

In the history of probability theory (e.g., as described by Hacking [3]), games of chance provided an early motivation for the theory's development. The view begins with the specification of the elementary events of Axiom 1. Further, the situation provides no differentiating knowledge about the elementary events, and a perceived symmetry is observed among them. For example, the usual six-sided die provides no information as to any side being favored, and each side is roughly equivalent. This state of affairs leads to an application of the principle of insufficient reason, whereby all elementary events are accorded the same likelihood (for the die, each face has probability $1/6$ of appearing).

This view has been most usefully applied in using probability as the metric for games of chance, e.g., in casinos and lotteries. Under this interpretation, the emphasis is upon counting: translating the numbers of elementary events to numbers of more complex events using the relationships derived within mathematical probability theory. For example, if drawing each individual card in a standard deck of 52 cards is an elementary event having a $1/52$ chance, what

is the probability of getting exactly three matching card values out of five cards (a three-of-a-kind)? The mathematics of probability theory combines with that of combinatorics to derive complex probability statements like these.

Frequentist View. *What proportion of days in the past 10 years has there been a traffic accident involving at least one fatality, within the city limits of Minneapolis, Minnesota (USA)?*

What is the probability of a traffic accident involving at least one fatality tomorrow, within the city limits of Minneapolis, Minnesota (USA)?

Probability is also recognized as describing proportions, as in the first traffic example above. Proportions clearly correspond with the tenets of probability theory; indeed, probability theory is well characterized as the mathematics of proportions. So, the proportion of days with exactly one fatality can be added to the proportion of days with two or more fatalities and this sum will exactly equal the proportion of days with one or more fatalities (an application of Axiom 5).

The frequentist interpretation takes this correspondence and extends it to the realm of uncertainty. As a measure of likelihood, the frequentist interpretation views probabilities as idealized proportions (as the number of observations approaches infinity). Thus, the proportion of the first traffic example could be idealized to a representation of a process producing potentially indefinite traffic fatality values over time, according to some idealized proportion. To the extent that tomorrow's value is drawn from this process, the idealized proportion provides a likelihood estimate for tomorrow. Reversing the logic, as the number of tomorrows increases (approaching the unreachable, idealized value of infinity), the observed proportion of days with fatalities converges to a value that is then accepted as the probability.

In general, this view is treated as useful in situations where there is (a) repeatability (identical events are recurring so that some large number of occasions can be observed) and (b) stability (the same process is operating over time to generate these occasions, so that the events are identically distributed). Unlike the classical view, the elementary events are not assumed to be equally likely. However, they could be, making this interpretation a more general case of the classical view, with an empirical grounding.

Subjective View. *What is the probability of the euro being adopted by the United Kingdom within the next 10 years?*

It is hard to imagine any set of similar, equivalent events for which a frequentist interpretation of this probability would make sense. And yet, making such a judgment does make sense to most people. The subjective view partly arose from such considerations.

In the subjective view, probabilities are characterized as degrees of belief, more specifically as judged likelihoods. Probabilities under this interpretation are presumed to derive from the evidence that underlies the qualified belief. This evidence could include the presumption of equal likelihoods or evidence of proportions, making this interpretation a more general case of the classical and frequentist views, while stepping back from the more direct empiricism of the frequentist view.

Thus, one approach to subjective probability is as an interpretation of the mathematical theory of probability. In this approach, a subjective probability is a correspondence between the theory and a reality defined as a subjective degree of belief based on evidence. Before turning in the section titled “Subjective Probabilities as Assessed Degrees of Belief” to the development of two variants of this correspondence, the next section discusses the implications of this view for the important issue in risk assessment of the quality of subjective probability measures.

Quality: Internal Coherence

Under this perspective, what does it mean for a subjective probability to be good or at least better than another? What standard of quality can be meaningfully applied? One difficulty with the subjective view as a generalization of the frequentist view, as described above, is the disconnect that is created from empirical evidence. Whereas frequentist probabilities can be meaningfully compared with observed proportions (e.g., see the section titled “Quality: Scoring Rules”), subjective probabilities do not necessarily afford such a comparison. If one person judges that the $P(\text{Rain tomorrow}) = 0.7$ based on the evidence he or she has and another person judges $P(\text{Rain tomorrow}) = 0.6$ based on his or her evidence and if it rains tomorrow, is the former’s judgment better? Since the judgment is personal and based on different evidence, we might be hesitant

to draw this conclusion, particularly for unrepeatable events.

The primary means of assessing quality that has developed from the view of subjective probability as an interpretation of mathematical probability is the criterion of internal coherence. This criterion is not applicable to a single judgment like the above; no quality assessment of a single subjective probability has been convincingly argued. Internal coherence does provide a quality standard for certain groups of judgments, however. The mathematical theory, e.g., as illustrated by Axiom 5 and by the definition of conditional probability (equation (1)), makes statements about how probabilities are related to each other. So, if independent assessments of $P(A)$, $P(B)$, and $P(A + B)$ are made for events A and B that are disjoint, then they should relate to each other according to the specifications of Axiom 5: $P(A) + P(B) = P(A + B)$. For subjective probabilities to satisfy this condition represents internal coherence. The three probabilities cohere as a group to the required relationship set forth by the mathematical theory.

Numerous studies have tested different relationships required by the theory. Generally, the results are not positive; subjective probabilities typically do not cohere. For the relationship in Axiom 5, individuals tend to be superadditive, underjudging the likelihood of a disjunction, i.e., stating judgments such that $P(A) + P(B) > P(A + B)$ for mutually exclusive events A and B . For example, Barclay and Beach [4] asked students to imagine someone they know getting a car for graduation. They assessed the probabilities of the car being a Ford, being a Chevrolet, and being a Ford or a Chevrolet. The average responses showed $P(\text{Ford}) + P(\text{Chevrolet}) > P(\text{Ford or Chevrolet})$. A special case of the effect is for partitions of the collection of all possibilities: $\{A_i\} = E$. Wright and Walley [5] showed partitions having five to six constituent elements to have a probability sum of about 2 (vs $P(E) = 1$); and, for partitions of 16 elements, the sum rose to 3.

One of the earliest lines of research (e.g. [6]) on the coherence of subjective probability judgments demonstrated conservatism in the updating of probabilities with evidence, as compared to Bayes’ theorem (see **Bayes’ Theorem and Updating of Belief**). The theorem, which can be derived from the definition of a conditional probability given in equation (1), can

be stated as

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)} \quad (2)$$

where H is interpreted as some hypothesis and D as data relevant to the hypothesis. In this equation, $P(H)$ is the prior probability of the hypothesis before the data are received and $P(H|D)$ is the posterior probability after the data are received. $P(D|H)$ expresses the likelihood of the data for this hypothesis and $P(D)$ normalizes the calculation so that the posterior probability remains within the interval $[0, 1]$. The usual finding was that judgments of $P(H|D)$ were less than those calculated using equation (2) from the judgments of the components on the right-hand side of the equation, i.e., the updated probability is too conservative.

An even more dramatic example of a failure of internal incoherence is the conjunction fallacy identified by Tversky and Kahneman [7]. The mathematical theorem tested in this case is that

$$P(AB) \leq P(A) \quad (3)$$

i.e., the joint occurrence of two events cannot be more likely than either event alone. For example, a head on the first toss of a coin cannot be less likely than a head on both of the first two tosses. As intuitive as this principle appears, systematic violations of the relationship have been demonstrated in a number of studies. The Linda problem, one of those used by Tversky and Kahneman, provides an oft-cited example. Consider the following description:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

Then, presented with several statements including the following that

- [A] Linda is a bank teller
- [B] Linda is active in the feminist movement
- [AB] Linda is a bank teller and is active in the feminist movement

the vast majority of respondents judged the likelihood of A as less than the likelihood of AB .

(For additional information, including reasons for such behavior, Yates [8] has an accessible discussion

of findings with respect to the internal coherence standard.)

Subjective Probabilities as Assessed Degrees of Belief

The mathematical approach of the previous section is outcome oriented. It focuses upon the resulting probabilities attached to events without considering how the degrees of belief arise. Addressing the latter leads to a more behaviorally oriented approach. There are two general means by which degrees of belief are obtained: by direct assessment of likelihoods and by inference from assessed preferences. And, an important theoretical foundation in either case is the use of an axiomatic theory that analyzes the conditions required of these assessments to derive probabilities from them. This section describes an axiomatic system for a qualitative probability structure designed to connect subjective probability more directly to the underlying behavior it is intended to represent.

Qualitative Probability Structures

In terms of an axiomatic subjective theory, de Finetti [9] provided the starting point. The idea is to provide minimal constraints on the judgments that would be used in assessing the probabilities and from which a coherent set of quantitative probabilities will be derived. In other words, if I want to have a set of numerical probabilities that behave in a coherent mathematical way, what constraints must be placed on the judgments? In the axiom system, the judgment is described as a qualitative comparative relation, $A \succsim B$, between sets (events) A and B . Typically, this relation translates to either “is at least as likely as” or “is at least as preferred as”. These two interpretations form the bases of the next two subsections. The axiomatic structure below follows that described by Luce and Suppes [10].

1. All subsets of E , and their sums $[A + B$, i.e., A or $B]$, products $[AB$, i.e., A and $B]$ and differences $[A - B$, i.e., A but not $B]$ are defined [equivalent to Axiom 2 above for mathematical probability theory].
2. $A \succsim B$ and $B \succsim C$ implies $A \succsim C$ [transitivity].
3. Either $A \succsim B$ or $B \succsim A$ [or both; 2 and 3 signify that there is a weak ordering of all subsets, “weak” in that ties are allowed].

4. If A and C are mutually exclusive (have nothing in common) and B and C are mutually exclusive, then $A \succeq B$ if and only if $A + C \succeq B + C$ [common elements, e.g., C , can be ignored or included without effect, a principle of the irrelevance of common alternatives].
5. $A \succeq \emptyset$ [i.e., nothing is less likely than the impossible event; sets a lower bound].
6. Not $\emptyset \succeq E$ [the entire set of possibilities is more likely than the impossible event; eliminates the trivial case in which everything is impossible].

To these axioms we can add the following definition:

$A \succ B$ if and only if $B \not\succeq A$,

defining the strict ordering relation (without ties). A conditional probability can also be defined; however, this is more complicated and will be left aside here. Suppes and Zanotti [11] provide an example. For our purposes, it is sufficient to note that constraints on relations involving conditionality are similar to those above, and support the existence of a quantitative definition like that of equation (1).

The qualitative axioms present a series of constraints on subjective probability judgments. For example, Axiom 2 claims transitivity: If rain tomorrow is judged to be at least as likely as snow and snow is at least as likely as fair weather, then the axioms require that rain be judged to be at least as likely as fair weather. As straightforward as transitivity seems, there are situations for which preferences have been demonstrated to violate transitivity [12]. Such behavior presents a caution to derivations of subjective probability.

In terms of these axioms, probably the most controversial is Axiom 4. As an example, Ellsberg [13], in introducing the concept of ambiguity and positing consistent ambiguity aversion, focused on this axiom. The challenge can be demonstrated by the following

example (Table 1). An urn contains 150 balls: 50 of these are known to be blue and the remaining 100 are red and/or white in some unknown proportion. All 100 could be red, white, or anything in between. Several options, A–D, are defined with respect to this urn with the payoffs associated with color draws in each game shown by Table 1.

Two choices are offered: a choice between option A and option B and a choice between option C and option D. In the first choice, the modal selection is option A, avoiding the additional uncertainty, termed *ambiguity*, of option B. (How many winning balls are there?) In the second choice, the modal selection is option D, similarly avoiding the ambiguity of option C. However, this pattern of choices is inconsistent with a qualitative probability structure, particularly with Axiom 4, the irrelevance of common alternatives.

To show this, we can set the utility or value $u(\$100) = 1$ and $u(\$0) = 0$, since value is generally recognized as being scaled intervally, i.e., any linear transformation of value retains its meaning. Then, the preference for option A implies

$$P(\text{Blue}) > P(\text{Red})$$

whereas the preference for option C implies

$$P(\text{Red}) + P(\text{White}) > P(\text{Blue}) + P(\text{White})$$

or, canceling the common alternatives

$$P(\text{Red}) > P(\text{Blue})$$

leading to incompatible inequalities.

The problem highlighted by this particular example leads to an expansion of the concept of subjective probability that is taken up in the section titled “Justification-Based Measures of Subjective Probability”. In addition, the example shows a potential discrepancy between two forms of assessing subjective probabilities. These are identified next.

Assessing Subjective Probabilities

By far, the most common way of assessing subjective probabilities in practice is to do so directly. The relation $A \succeq B$ is generally interpreted as “ A is at least as likely as B ”. By this method, degrees of belief as sensitive to likelihood are taken as the primitives that assessors can judge directly. Thus, a weather forecaster or a physician, providing a forecast or diagnosis, directly reports the numerical measures. As demonstrated by the example in the previous section,

Table 1 Sample options demonstrating a version of Ellsberg’s paradox

Option	50 balls		100 balls
	Blue	Red	White
A	\$100	\$0	\$0
B	\$0	\$100	\$0
C	\$100	\$0	\$100
D	\$0	\$100	\$100

an alternative is to derive subjective probabilities from preferences. In this case, the relation $A \succeq B$ is generally interpreted as “A is at least as preferred as B”.

Decision scientists have tended to favor this latter approach, grounded on preference behavior. Savage’s [14] landmark treatise particularly connected probability and utility theory, providing a theoretical and axiomatic basis for what has come to be known as *subjective expected utility (SEU) theory*. This theory incorporates a subjective (personal) probability measure into preference as modeled by utility theory. Another manifestation of the decision analytic ties between probability and utility is the probability elicitation method for assessing utilities. In this method, utilities are derived from judgments like the following: for what probability P are you indifferent between receiving \$50 for certain; or, receiving \$100 with probability P , otherwise receiving nothing? This method essentially measures utilities as an assessment of preference along a probability scale.

Under a strict decision analytic view (*see Decision Analysis*), judgments are only of significance if they are connected to consequences. Direct assessments that are not grounded in outcomes do not qualify. Only preference-based methods, tied to real outcomes, lead to usable subjective probability assessments. Despite this argument, in practice, direct assessments predominate.

Quality: Convergence

Considering subjective probabilities as arising from an assessor based on the evidence available, another view of decision quality can be proposed. However, for whatever reasons, this approach is rarely used. From this view, quality can be defined as correspondence with a converging value with shared evidence. This idea has at least three components. First is that multiple judges of an event will tend to converge on a common probability judgment as they increasingly share their knowledge on which the judgment is based. Second is that, if two judgments are based on sets of information where one set of information is a subset of the other, then the judgment based on the more inclusive set of evidence is not worse than the other. Third is that quality can be measured as the degree of correspondence with the convergent value under the condition that the shared information is complete, i.e., as the information is fully shared. This

notion of quality is more speculative than the others identified in this article, as this view has not been applied experimentally to the evaluation of subjective probabilities. However, authors have identified convergence as a potential expectation and, as such, it provides a potential standard of quality.

Probabilities as Inherently Subjective

All the above characteristics define subjective probability as against other means of interpreting probability, e.g., against classical or frequentist interpretations. An alternative approach is to accept that probabilities in practice are all subjective. The characteristics that lead to the interpretations are just different forms of evidence.

Consider rolling a six-sided die. The probability of a single roll leading to an ace is generally judged to be 1/6. Such an assessment, from the classical view, arises from there being six elemental events, each of which is treated as equally likely. But, how does one determine that they are equally likely? Not all elemental events are so. If one is assessing the $P(\text{Rain in Minneapolis MN, USA, tomorrow})$, one can define Rain and Not Rain as elemental events, but they are not equally likely. To define equally likely events in this situation is probably not possible (a limitation of the classical view that led to its falling from favor). So, again, how does one determine that the six outcomes of a die roll are equally likely? One cannot prove this logically. One cannot prove this empirically either. Even under the frequentist view of a probability as a long-run frequency, the long run is not an empirical possibility but a theoretical concept: the frequency as the number of rolls approaches an unattainable infinite number of rolls.

Finally, the answer is that one cannot prove that the events are equally likely. Ultimately it is a subjective assessment. Consequently, the principle of insufficient reason is understood, not as a statement of how to treat symmetrical events but, as a subjective recommendation of how to assign probabilities to elemental events when no other evidence is available.

Similarly, probability is not equivalent to frequency, even in an idealized sense. To the extent that frequencies underlie a probability, wholly or in part, this also is ultimately a subjective judgment arising from the assessor’s knowledge. If it is just analyzing how proportions are related to each other, no

subjective assessment is necessary. But, the use of proportions as a basis for a probability is necessarily going beyond the available evidence to treat the proportion as an idealized long-run frequency. What justifies such a statement? The frequentist interpretation of probability is argued here as grounded in subjective assessments of repeatability and stability. Consider assessing the likelihood of a traffic fatality in Minneapolis tomorrow by using past frequency data. It is assumed that the events underlying the proportion are comparable to each other (e.g., each day is comparable to an identifiable set of other days) and that the likelihood is unchanging (e.g., the chances of an accident tomorrow are the same as on each of those other comparable days). How can we know either of these conditions? We cannot. It is a subjective assessment based on available evidence that leads us to operate on these assumptions.

The argument, therefore, is that subjective probability is not one of several possible interpretations of probability, but rather that all real-world applications of the mathematical probability are necessarily subjective. Elaborating, the uncertainty being captured in a probability judgment is recognized as depending not only on the event being assessed, but also necessarily upon the knowledge that the assessor brings to the judgment.

Bayesian Inference

The recognition of a subjective aspect in using any probability extends to the use of probabilities in statistical inference, most notably in the Bayesian approach to inference (see **Bayesian Statistics in Quantitative Risk Assessment**). Traditional statistical inference, as manifest in null hypothesis testing and the use of p values is frequentist. A p value is interpreted as the probability of a sample result, in an idealized long-run number of experiments, if the null hypothesis is true. In this, the p value is most comparable to the likelihood in equation (2), $P(D|H)$, with a frequentist interpretation of this probability.

The Bayesian inference approach considers inference from a more subjective viewpoint. The approach is to consider probability statements as arising from the consideration of evidence, from a process of updating assessments based on accumulating evidence. Bayes' theorem, which describes the updating of an hypothesis in light of data, for example as

expressed in equation (2), is the conceptual centerpiece to the subjective approach to statistical inference. First, the focus is upon $P(H|D)$ rather than $P(D|H)$. Congruently, the interpretation of the posterior $P(H|D)$ is as a subjective degree of belief in light of the evidence of D .

As to the first point, the Bayesian approach is accommodating a more natural way of viewing the situation. Anyone who has taught an introductory statistics course can testify to the nonintuitive nature of the p value as $P(D|H)$. Despite all warnings to the contrary, students tend to treat the p value as $P(H|D)$. This confusion of the two conditionals has also been observed notably by Eddy [15] in his analysis of reports in the medical literature. Eddy observed the same tendency among researchers interpreting medical data as observed in the introductory class, interpreting $P(D|H)$ as $P(H|D)$. Bayes' theorem explicitly acknowledges that it is ultimately $P(H|D)$ that is of interest and models this explicitly.

The second complementary aspect is the interpretation of the resulting probability, the posterior $P(H|D)$, as subjective. In particular, the probability is interpreted as a degree of belief in response to evidence. In this light, the principle of insufficient reason of the classical view is simply a proposal for the starting point, the prior probability $P(H)$ before any evidence is available. And, frequencies are just a particular form of evidence D from which posterior probabilities are derived, along with any other non-frequentist evidence that may be available. In the end, all probabilities are viewed as subjective statements based on evidence.

Quality: Scoring Rules

Subjective probability as an encompassing category, treating all probabilities as inherently subjective, creates a connection back to the frequentist approach that has been exploited in evaluating the quality of subjective probability assessments. The approach removes the distinction between subjective and frequentist probabilities, treating frequencies or base rates, as just part of the information that can be incorporated subjectively. Thus, this evidence also can be used to evaluate subjective probabilities without contradiction.

The most direct applications of this idea are calibration analyses. A subjective probability judgment of 0.70 attached to some target event is said to be

well calibrated if the target event occurs 70% of the time that such a probability is stated. If this is true for all the probability numbers that one uses, then he or she is a well-calibrated judge. The usual index used to measure calibration is the weighted squared differences between the actual frequencies and the subjective probability judgments.

Typically, researchers try to select events, in the calculation of the calibration index, that are comparable, i.e., that can arguably be classified as repeatable events of the same type. In different situations though, this comparability may be closer than in others. In weather forecasting, for example, each daily forecast can be considered a repeatable event, forming a reasonably coherent category of events. Alternatively, researchers have requested probability assessments using general knowledge questions on a variety of topics (How certain are you that your answer is correct?), a less coherent set of events.

This variability in the use of calibration has not gone unnoticed. Most notably, Gigerenzer *et al.* [16] made a similar point in questioning the ecological validity of questions used in calibration studies, i.e., challenging to what extent the distribution of questions reflects what is encountered in everyday life. Here, the somewhat curious aspect of using frequentist data to evaluate subjective probabilities is simply observed. From a view where these are distinct interpretations of mathematical probability, this is nonsensical; however, the rationale is supported by a view that encompasses all probability interpretations within the subjective rubric.

Expanding off of this idea, calibration is just one aspect of performance that can be assessed by comparing judgments to observed frequencies. Other measures can be identified beginning with the idea of scoring rules. The most common scoring rule used is the quadratic rule. Defining p_j as the subjective probability for situation j and d_j as an outcome index such that

$$d_j = 1 \text{ if the target event being assessed occurs} \\ 0 \text{ if the target event does not occur}$$

the quadratic rule is an overall index given by

$$\sum_j (p_j - d_j)^2 \quad (4)$$

The quadratic rule is a proper scoring rule. A *proper* scoring rule is one that, if used to reward performance, leads to maximal payoff when assessors report their judgments honestly and accurately without bias. Another reason for the widespread use of the quadratic rule is the ability to decompose the overall score into component scores that are diagnostic of different aspects of performance quality, including the calibration index discussed above. Yates [17] provides a good discussion of properness – demonstrating the property with several rules, proper and improper – and of scoring rule decompositions of the quadratic rule.

Justification-Based Measures of Subjective Probability

In the earlier sections, there was a progression of generality from considering subjective probability as one interpretation among several of the mathematical theory of probability to considering subjective probability as encompassing all behavioral applications of probability. In this section, the generality is pushed further to consider subjective probability as a general approach to uncertainty whether grounded in the traditional mathematical theory of probability or not. Two basic variations can be identified.

The first uses measures that allow imprecision in the subjective probability judgments. Instead of requiring a precise judgment, however assessed, some vagueness is allowed. Examples are uses of upper/lower probabilities, so that individuals need to only specify a range of values. Similar in spirit, though even less precise, are uses of verbal phrases of likelihood, e.g., “possibly”, “probably”, or “high chance”.

The second variation of measuring uncertainty, outside of mathematical probability theory, is as old as probability theory itself, tracing to the earliest days of the development of the mathematical theory. The basic idea is that there are two clearly distinguishable aspects of uncertainty that have subjective correlates, each of which can be, and has been, expressed as a degree of belief, though distinct. Following the evidential approach of the previous section, the two relate to the strength of evidence and the weight of evidence that underlie the subjective assessment. Alternatively, if knowledge is classified as justified true belief (e.g. [18]), then there are two criteria

implied. Subjective probabilities as likelihoods capture the degree of belief relative to the criterion of truth; but, subjective probabilities could equally be developed relative to the criterion of justification. Shafer [19] provides an excellent, comprehensive history of the difference.

The distinction is well demonstrated by the concept of ambiguity as described and illustrated by Ellsberg [13]. Reconsider options A–D in Table 1 involving the urn containing 150 balls: 50 blue and 100 red and/or white balls in an unknown proportion. As posited by Ellsberg, people will react to the ambiguity in the choices between options A and B and between options C and D. The behavior is typically one of avoiding ambiguity and reactions to ambiguity have been widely verified (e.g. [20–22]). The example exemplifies two types of uncertainty. Relative to the truth criterion, all the options involve risk in that the outcome (the true state of affairs that will occur), \$100 or nothing, is uncertain. If asked for the likelihood of winning, the modal response is consistent: 1/3 for options A and B and 2/3 for options C and D. These likelihoods capture the balance of the evidence. There being no additional evidence, the balance of the evidence falls equally toward the three elemental events of the three colors, i.e., $P(\text{Each color being drawn}) = 1/3$. Options B and C, however, also contain additional uncertainty, termed *ambiguity* by Ellsberg, in that the chances of winning are more uncertain for these options. Relative to the justification criterion, the difference is in the weight of the evidence with these options. There is more information about options A and D, more justification or support for conclusions drawn about them, as the exact proportions of winning and losing colors are known. The justification is less for options B and C.

The goal of incorporating weight of evidence into a subjective measure of uncertainty motivates the development of a number of calculi as alternatives to standard mathematical theory of probability as outlined above, e.g., the use of certainty factors or possibilistic measures, particularly in artificial intelligence [23, 24]. As an example, consider the functions within Dempster–Shafer theory as measures of support. A full account of the rationale and mathematical formulation can be found in Shafer [25]. An operationalization of the theory as a subjective probability measure using direct assessments was done by Curley [26] in a study of the evaluation of evidence in a

legal setting. In traditional, direct probability assessment, numbers are assigned to a partition of the set of all possibilities (S) such that the numbers sum to 1. Alternatively, reserve functions allow judges to assign numbers to any subset of S so that they sum to 1. Assigning $m\%$ of one’s belief to a subset is interpreted as follows: “On the basis of the evidence, I believe with $m\%$ of my belief that the possibilities in this set are supported; however, I cannot distinguish between the elements in the set individually.” When there is no evidence, all the belief is reserved to S , $m(S) = 1$. As evidence accumulates, the evidence moves into subsets to the degree that the evidence implicates the contained elements.

For example, suppose a piece of evidence implicates a left-handed suspect and persons A and B among the possible suspects are left handed. Initially, before any evidence is available, $S = \text{set of possible suspects}$ and $m(S) = 1$. With the piece of evidence implicating persons A and B, the support moves so that $m(\{A, B\}) = k$ and $m(S) = (1 - k)$, where k is the degree of implication provided by the evidence. This creates two differences from probability theory. First, with traditional probability if the likelihood of $\{A, B\}$ following this evidence is P , then the likelihood of all other suspects excluding $\{A, B\}$ is $(1 - P)$. This is not the case with justification (as S includes all suspects, including A and B), and the mathematics of reserve functions does not force this requirement. Second, in probability theory, for A and B distinct, $P(\{A, B\}) = P(\{A\}) + P(\{B\})$: in considering which is true, the likelihood of A or B is the sum of their relative likelihoods. In considering which is justified, this is not required. In the above example, $m(\{A\}) = m(\{B\}) = 0$, while $m(\{A, B\}) = k$. A support function does not require that the justification be able to be partitioned into the constituent elements if the evidence does not support doing so.

Being more novel and less used, the approaches in this section have not addressed the issue of quality. Ultimately, internal coherence (*cf.* the section titled “Quality: Internal Coherence”) and convergence (*cf.* the section titled “Quality: Convergence”) criteria seem most promising.

Conclusion

Subjective probability, generally characterized as a degree of belief in the face of uncertainty arising from

variability in our environment, has proven a robust construct. Several major variant approaches to subjective probability have developed. One view frames subjective probability as one of several interpretations whereby the mathematical theory of probability can be applied in real life. The probabilities are either assessed directly or inferred from stated preferences. Alternatively, it is argued that all probabilities are subjective, derived from evidence which may include frequency or process information. Bayesian inference best connects to this view. Finally, alternatives to the mathematical theory of probability have been proposed. Degrees of belief, as capturing the justification provided by evidence as opposed to the likelihood of the resulting events occurring, are described as an exemplar.

References

- [1] Wright, G. & Ayton, P. (eds) (1994). *Subjective Probability*, John Wiley & Sons, Chichester.
- [2] Kolmogorov, A.N. (1933/1950). *Foundations of the Theory of Probability* (translation by N Morrison), Chelsea Publishing, New York.
- [3] Hacking, I. (1975). *The Emergence of Probability*, Cambridge University Press, Cambridge.
- [4] Barclay, S. & Beach, L.R. (1972). Combinatorial properties of personal probabilities, *Organizational Behavior and Human Performance* **8**, 176–183.
- [5] Wright, G. & Walley, P. (1983). The supra-additivity of subjective probability, in *Foundations of Utility and Risk Theory with Applications*, B.P. Stigum & F. Wenstop, eds, Reidel, Dordrecht, pp. 233–244.
- [6] Phillips, L.D. & Edwards, W. (1966). Conservatism in a simple probability inference task, *Journal of Experimental Psychology* **72**, 346–354.
- [7] Tversky, A. & Kahneman, D. (1982). Judgments of and by representativeness, in *Judgment under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic & A. Tversky, eds, Cambridge University Press, Cambridge, pp. 84–98.
- [8] Yates, J.F. (1990). *Judgment and Decision Making*, Prentice-Hall, Englewood Cliffs.
- [9] de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives, *Annales de L'Institut Henri Poincaré* **7**, 1–68; English translation in Kyburg Jr, H.E. & Smokler, H.E. (eds) (1964). *Studies in Subjective Probability*, Krieger, Huntington, pp. 93–158.
- [10] Luce, R.D. & Suppes, P. (1965). Preference, utility and subjective probability, in *Handbook of Mathematical Psychology*, R.D. Luce, R.R. Bush & E. Galanter, eds, John Wiley & Sons, New York, Vol. 3, pp. 249–410.
- [11] Suppes, M. & Zanotti, P. (1982). Necessary and sufficient qualitative axioms for conditional probability, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **60**, 163–169.
- [12] Tversky, A. (1969). Intransitivity of preferences, *Psychological Review* **76**, 31–48.
- [13] Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms, *Quarterly Journal of Economics* **75**, 643–669.
- [14] Savage, L.J. (1954). *The Foundations of Statistics*, John Wiley & Sons, New York.
- [15] Eddy, D.M. (1982). Probabilistic reasoning in clinical medicine: problems and opportunities, in *Judgment under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic & A. Tversky, eds, Cambridge University Press, Cambridge, pp. 249–267.
- [16] Gigerenzer, G., Hoffrage, U. & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence, *Psychological Review* **98**, 506–528.
- [17] Yates, J.F. (1982). External correspondence: decompositions of the mean probability score, *Organizational Behavior and Human Performance* **30**, 132–156.
- [18] Shope, R.K. (1983). *The Analysis of Knowing*, Princeton University Press, Princeton.
- [19] Shafer, G. (1978). Non-additive probabilities in the work of Bernoulli and Lambert, *Archive for History of Exact Sciences* **19**, 309–370.
- [20] Camerer, C.F. & Weber, M. (1992). Recent developments in modeling preferences: uncertainty and ambiguity, *Journal of Risk and Uncertainty* **5**, 325–370.
- [21] Curley, S.P. & Yates, J.F. (1989). An empirical evaluation of descriptive models of ambiguity reactions in choice situations, *Journal of Mathematical Psychology* **33**, 397–427.
- [22] Keren, G. & Gerritsen, L.E.M. (1999). On the robustness and possible accounts of ambiguity aversion, *Acta Psychologica* **103**, 149–172.
- [23] Dubois, D. & Prade, H. (1988). An introduction to fuzzy and possibilistic logics, in *Non-Standard Logics for Automated Reasoning*, P. Smets, E.H. Mamdani, D. Dubois & H. Prade, eds, Academic Press, London.
- [24] Shortliffe, E.H. (1976). *Computer Based Medical Consultations: MYCIN*, Elsevier Science, New York.
- [25] Shafer, G. (1976). *A Mathematical Theory of Evidence*, Princeton University Press, Princeton.
- [26] Curley, S.P. (2007). The application of Dempster-Shafer theory demonstrated with justification provided by legal evidence, *Judgment and Decision Making* **2**, 257–276.

Related Articles

Expert Judgment

Group Decision

Players in a Decision

Uncertainty Analysis and Dependence Modeling

SHAWN P. CURLEY

Supra Decision Maker

Decisions about the regulation of risks typically impact on a number of different sectors of society, and so have to take cognizance of the views of a variety of parties. In a system of representative democracy, such decisions often fall within the remit of independent regulators or agencies, for example, an Environment Agency. In the case of particularly consequential decisions, such as the disposal of nuclear waste, governments may retain decision making authority centrally, but may request a broadly based *ad hoc* committee to study the alternatives, and make the decision on the basis of the committee's recommendations (see **Stakeholder Participation in Risk Management Decision Making**).

One way to conceptualize such decisions is to think of the parties who are impacted by the decision – the “stakeholders” – as having preferences which are representable by value functions $v_i(\cdot)$ over an alternative space X (where i indexes n stakeholders). In this case, we can think of the regulatory or governmental decision maker as a “supra decision maker” (supra DM), who wishes to take account of the various $v_i(\cdot)$ in their own value function $V(\cdot)$.

Essentially, the supra DM concept is a way of thinking about the aggregation of individual preferences, a topic that has been explored extensively in the social choice, welfare economics and game theory literatures. However, the notion of a “supra DM” represents a distinctively decision-theoretic take on this topic, turning the focus on concrete choices made by an actual agent, rather than seeking to characterize properties of particular aggregation rules in the abstract.

As an example of how this perspective can be useful, consider the parallel problem of aggregating probability distributions elicited from a number of experts. An important question is: should it matter whether probabilities are updated *via* Bayes theorem (see **Bayes' Theorem and Updating of Belief**) with data before or subsequent to aggregation? As it happens, with popular approaches to aggregation such as linear opinion pools, it does matter (these aggregation procedures do not exhibit “external Bayesianity”). Here, the introduction of a (Bayesian) supra assessor, who takes as input data the probability distributions assessed by a number of individual assessors, has been used as a thought experiment to clarify

ideas about the desirability or otherwise of external Bayesianity [1].

Is it helpful to entertain the idea of an *expected utility maximizing supra DM* analogous to this Bayesian supra assessor? A compelling argument against this is provided by Broome [2]. The Sure Thing Principle of Utility Theory (see **Utility Function**) requires that if a supra DM is indifferent between two alternatives, she is also indifferent between these alternatives and their probabilistic mixture. Yet, in choosing which of the two indistinguishable individuals is to receive an indivisible good or bad (such as a green card or a Vietnam call-up), a supra DM might prefer to hold a lottery rather than make a deterministic allocation to one person or another.

Because this role for randomizing devices in the fair allocation of goods seems to undercut utility theory, the focus here, is only on the theory of the supra DM in the case of certainty. The formalization of this theory presented in Keeney and Raiffa [3, pp. 524–526] involves the following three assumptions: first, an assumption that the supra DM's preferences are completely determined by the stakeholders' preferences; secondly, an assumption of preferential independence on the part of the supra DM; and thirdly, a Pareto-like assumption that if an alternative can be unilaterally improved for one stakeholder, without worsening it for any of the other stakeholders, the improved alternative will be preferred by the supra DM.

Given these assumptions, it can be shown that the supra DM's value function must take the form $V(x) = \sum_{i=1}^n f_i(v_i(x))$. Further specialization of this function is possible: for example, the Bergson additive power form, $V(x) = \sum_{i=1}^n \alpha_i (v_i(x))^p$, has some attractive properties [4, Section 6]. The parameters of this function have a natural interpretation, as the multiplicand α_i is a person weight and the power parameter p reflects the supra DM's equity attitude, with $0 < p < 1$ corresponding to inequity aversion.

Yet this formal presentation avoids two crucial problems, which relate to the essentially private, personal nature of preferences. The first problem is the difficulty of knowing whether the stakeholders have reported their preferences honestly. Theorists in the field of social choice have agonized over this question, even before Gibbard [5] demonstrated that all voting systems of any interest encourage

strategic voting behavior. Devices for incentivizing the revelation of true preferences do exist, but are unlikely to be implementable in a risk analysis setting. Practically speaking, the supra DM's only recourse is to make a judgment about whether preference information is being given in good faith, and such judgments, inevitably, are fallible.

The second, more philosophically profound, problem arises from the impossibility of knowing with a high level of precision what it feels like to be someone else. Assessing the person's weights, α_i requires the supra DM to imagine *how strong her preferences for one alternative over another would be if she were stakeholder i* relative to *how strong her preferences would be for some other pair of alternatives if she were stakeholder i'* , where i and i' are distinct. (This leaves aside the further complicating possibility that the supra DM may care more strongly for one stakeholder than another.) Requiring both a high level of empathy with stakeholders and a facility in characterizing one's own internal states quantitatively, this is a remarkably demanding mental task to set for anyone.

Perhaps because of these problems, actual decision and risk analyses tend not to structure problems using the supra DM formulation, although some examples do exist (e.g. [6]). Instead, conflicts in values tend to be dealt with in other ways. For example, one approach is to collect decision criteria relevant to all stakeholders and evaluate alternatives on these criteria. Different systems of preferences can be explored by parameterizing the model with sets of criteria weights reflecting the value judgments of different stakeholders, as the UK's Committee on Radioactive Waste Management (CoRWM) did in their analysis of disposal options for the UK's radioactive waste [7]. This approach facilitates the identification of robust solutions and sheds light on why holistic preferences might differ.

The supra DM formulation may be particularly unhelpful in applications where active involvement is sought from stakeholders in the analysis. The reasons are not far to seek – the explicit representation of individual stakeholder preferences may lead to

positions hardening, and the starkness of the interpersonal trade-offs required of the supra DM seem unlikely to facilitate acceptance of whatever solution emerges from the process. Yet, many risk-related decisions *do* impact different stakeholders differently, however much analysts may want to evade this, and the idea of a supra DM is a useful frame for thinking about, and perhaps for private analysis of, such decisions.

References

- [1] French, S. (1985). Group consensus probability distributions: a critical survey, in *Bayesian Statistics 2*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith, eds, Elsevier Science, Amsterdam, pp. 183–202.
- [2] Broome, J. (1982). Equity in risk bearing, *Operations Research* **30**(2), 412–414.
- [3] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley & Sons, Chichester.
- [4] Samuelson, P.A. (1965). Using full duality to show that simultaneously additive direct and indirect utilities implies unitary price elasticity of demand, *Econometrica* **33**, 781–796.
- [5] Gibbard, A. (1973). Manipulation of voting schemes: a general result, *Econometrica* **41**, 587–601.
- [6] Butterworth, N.J. (1989). Giving up 'The Smoke': a major institution investigates alternatives to being sited in the city, *Journal of the Operational Research Society* **40**, 711–717.
- [7] Catalyze (2006). *CORWM: MCDA Decision Conference*, 28–30 March 2006, Technical Report COR006, Catalyze Ltd, Winchester.

Related Articles

Decision Analysis

Evaluation of Risk Communication Efforts

Expert Judgment

Societal Decision Making

ALEC MORTON

Survival Analysis

In many studies, the main outcome is the time from a well-defined *time origin* to the occurrence of a particular event or *endpoint*. If the endpoint is the death of a patient, the resulting data are literally *survival times*. However, other endpoints are possible, for example, the time to relief of symptoms or to recurrence of a particular condition, or simply to the completion of an experimental task. Such observations are often referred to as *time to event data* although the generic term *survival data* is commonly used to indicate any time to event data.

Standard statistical methodology is not usually appropriate for survival data, for two main reasons:

1. The distribution of survival time, in general, is likely to be *positively skewed* and so assuming normality for an analysis (as done, for example, by a *t*-test or a regression) is probably not reasonable.
2. More critical than doubts about normality, however, is the presence of *censored* observations, where the survival time of an individual is referred to as *censored* when the endpoint of interest has not yet been reached (more precisely *right-censored*). For true survival times this might be because the data from a study are analyzed at a time point when some participants are still alive. Another reason for censored event times is that an individual might have been *lost to follow-up* for reasons unrelated to the event of interest, for example, due to moving to a location that cannot be traced. When censoring occurs, all that is known is that the actual, but unknown, survival time is larger than the censored survival time.

Specialized statistical techniques that have been developed to analyze such censored and possibly skewed outcomes are known as *survival analysis*. An important assumption made in standard survival analysis is that the censoring is *noninformative*, that is, the actual survival time of an individual is independent of any mechanism that causes that individual's survival time to be censored. For simplicity, this description also concentrates on techniques for continuous survival times; for the analysis of discrete survival times see [1, 2].

A Survival Data Example

As an example, consider the data in Table 1 that contain the times heroin addicts remained in a clinic for methadone maintenance treatment [3]. The study ($n = 238$) recorded the duration spent in the clinic, and whether the recorded time corresponds to the time the patient leaves the program or the end of the observation period (for reasons other than the patient deciding that the treatment is “complete”). In this study, the time of origin is the date on which the addicts first attended the methadone maintenance clinic and the endpoint is methadone treatment cessation (whether by patient's or doctor's choice). The durations of patients who were lost during the follow-up process are regarded as right-censored. The “status” variable in Table 1 takes the value unity if methadone treatment was stopped, and zero if the patient was lost to follow-up. In addition, a number of prognostic variables were recorded

1. one of two clinics
2. presence of a prison record
3. and maximum methadone dose prescribed.

The main aim of the study was to identify predictors of the length of the methadone maintenance period.

Survival Analysis Concepts

To describe survival, two functions of time are of central interest. The *survival function* $S(t)$ is defined as the probability that an individual's survival time, T , is greater than or equal to time t , that is,

$$S(t) = \text{Prob}(T \geq t) \quad (1)$$

The graph of $S(t)$ against t is known as the *survival curve*. The survival curve can be thought of as a particular way of displaying the frequency distribution of the event times, rather than by, say, a histogram.

In the analysis of survival data, it is often of some interest to assess which periods have the highest and which have the lowest chance of death (or whatever the event of interest happens to be), amongst those people at risk at the time. The appropriate quantity for such risks is the *hazard function*, $h(t)$, defined as the (scaled) probability that an individual experiences the event in a small time interval δt , given

Table 1 Durations of heroin addicts remaining in a methadone treatment program (only five patients from each clinic are shown)

Patient ID	Clinic	Status	Time (days)	Prison record 1 = present (0 = 'absent')	Maximum methadone (mg day ⁻¹)
1	1	1	428	0	50
2	1	1	275	1	55
3	1	1	262	0	55
4	1	1	183	0	30
5	1	1	259	1	65
...					
103	2	1	708	1	60
104	2	0	713	0	50
105	2	0	146	0	50
106	2	1	450	0	55
109	2	0	555	0	80
...					

that the individual has not experienced the event up to the beginning of the interval. The hazard function therefore represents the *instantaneous event rate* for an individual at risk at time t . It is a measure of how likely an individual is to experience an event as a function of the age of the individual. The hazard function may remain constant, increase or decrease with time, or take some more complex form. The hazard function of death in human beings, for example, has a “bath tub” shape. It is relatively high immediately after birth, declines rapidly in the early years, and then remains pretty much constant until beginning to rise during late middle age.

In formal, mathematical terms, the hazard function is defined as the following limiting value:

$$h(t) = \lim_{\delta t \rightarrow 0} \left[\frac{\text{Prob}(t \leq T < t + \delta t | T \geq t)}{\delta t} \right] \quad (2)$$

The conditional probability is expressed as a probability per unit time and therefore converted into a rate by dividing by the size of the time interval, δt .

A further function that features widely in survival analysis is the *integrated or cumulative hazard function*, $H(t)$, defined as

$$H(t) = \int_0^t h(u) du \quad (3)$$

The hazard function is mathematically related to the survivor functions. Hence, once a hazard function is specified so is the survivor function and *vice versa*.

Nonparametric Procedures

An initial step in the analysis of a set of survival data is the numerical or graphical description of the survival times. However, owing to the censoring this is not readily achieved using conventional descriptive methods such as box plots and summary statistics. Instead, survival data are conveniently summarized through estimates of the survivor or hazard function obtained from the sample.

When there are no censored observations in the sample of survival times, the survival function can be estimated by the *empirical survivor function*

$$\hat{S}(t) = \frac{\text{Number of individuals with event times} \geq t}{\text{Number of individuals in the data set}} \quad (4)$$

Since every subject is “alive” at the beginning of the study and no one is observed to survive longer than the largest of the observed survival times, then

$$\hat{S}(0) = 1 \text{ and } \hat{S}(t) = 0 \text{ for } t > t_{\max} \quad (5)$$

Furthermore, the estimated survivor function is assumed constant between two adjacent death times so that a plot of $\hat{S}(t)$ against t is a step function that decreases immediately after each “death”. However, this simple method cannot be used when there are censored observations since the method does not allow for information provided by an individual whose survival time is censored before time t to be used in the computing of the estimate at t .

The most commonly used method for estimating the survival function for survival data containing censored observations is the *Kaplan–Meier* or *product-limit-estimator* [4]. The essence of this approach is the use of a product of a series of conditional probabilities. This involves ordering the r sample event times from the smallest to the largest such that

$$t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(r)} \quad (6)$$

Then the survival curve is estimated from the formula

$$\hat{S}(t) = \prod_{j|t_{(j)} \leq t} \left(1 - \frac{d_j}{r_j}\right) \quad (7)$$

where r_j is the number of individuals at risk at $t_{(j)}$ and d_j is the number experiencing the event of interest at $t_{(j)}$ (individuals censored at $t_{(j)}$ are included in r_j). For example, the estimated survivor function at the second event time $t_{(2)}$ is equal to the estimated probability of not experiencing the event at time $t_{(1)}$ times the estimated probability, given that the individual is still at risk at time $t_{(2)}$, of not experiencing it at time $t_{(2)}$.

The Kaplan–Meier estimators of the survivor curves for the two methadone clinics are displayed in Figure 1. The survivor curves are step functions with decrease at the time points when patients ceased methadone treatment. The censored observations in

the data are indicated by the “cross” marks on the curves.

The variance of the Kaplan–Meier estimator of the survival curve can itself be estimated from *Greenwood’s formula* and once the standard error has been determined, pointwise symmetric confidence intervals can be found by assuming a normal distribution on the original scale or asymmetric intervals can be constructed after transforming $\hat{S}(t)$ to a value on the continuous scale; for details, see [1, 2].

A *Kaplan–Meier type estimator* of the hazard function is given by the proportion of individuals experiencing an event in an interval per unit time, given that they are at risk at the beginning of the interval, that is,

$$\tilde{h}(t) = \frac{d_j}{r_j (t_{(j+1)} - t_{(j)})} \quad (8)$$

Integration leads to the *Nelson–Aalen* or *Altshuler’s* estimator of the cumulative hazard function, $\tilde{H}(t)$, and employing the theoretical relationship between the survivor function and the cumulative hazard function to the Nelson–Aalen estimator of the survivor function. Finally, it needs to be noted that relevant functions can be estimated using the so-called *life table* or *actuarial estimator*. This approach is, however, sensitive to the choice of intervals used in its construction and therefore not generally recommended for continuous survival data (readers are referred to [1, 2]).

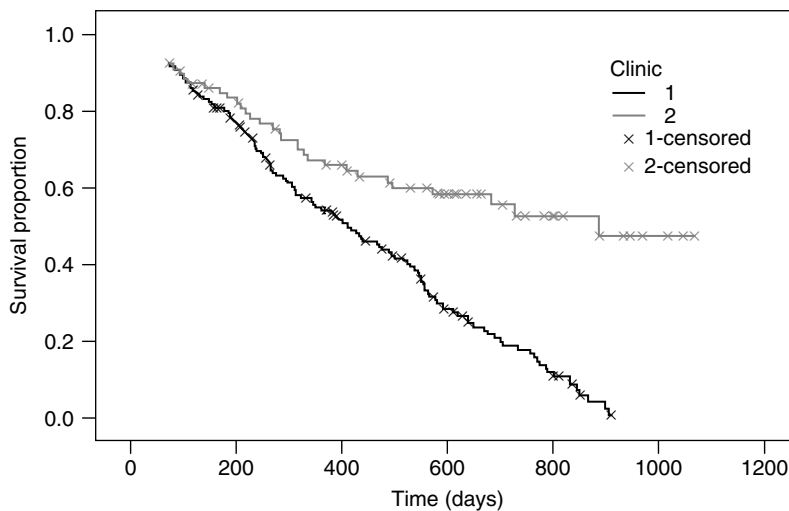


Figure 1 Kaplan–Meier survivor curves for the heroin addicts data

Standard errors and confidence intervals can be constructed for all three functions although the estimated hazard function is generally considered “too noisy” for practical use. The Nelson–Aalen estimator is typically used to describe the cumulative hazard function while the Kaplan–Meier estimator is used for the survival function.

Since the distribution of survival times tends to be positively skewed, the median is the preferred summary measure of location. The *median event time* is the time beyond which 50% of the individuals in the population under study are expected to “survive”, and, once the survivor function has been estimated by $\hat{S}(t)$, can be estimated by the smallest observed survival time, t_{50} , for which the value of the estimated survivor function is less than 0.5. The estimated median survival time can be read from the survival curve by finding the smallest value on the x axis for which the survival proportion reaches less than 0.5. Figure 1 shows that the median methadone treatment duration in clinic 1 group can be estimated as 428 days while an estimate is not available for clinic 2 since more than 50% of the patients continued treatment throughout the study period. A similar procedure can be used to estimate other percentiles of the distribution of the survival times and approximate confidence intervals can be found once the variance of the estimated percentile has been derived from the variance of the estimator of the survivor function.

In addition to comparing survivor functions graphically, a more formal statistical test for a group difference is often required. In the absence of censoring, a nonparametric test like the Mann–Whitney test could be used. In the presence of censoring, the *log-rank* or *Mantel–Haenszel test* [5] is the most commonly used nonparametric test. It tests the null hypothesis that the population survival functions $S_1(t), S_2(t), \dots, S_k(t)$ are the same in k groups.

Briefly, the test is based on computing the expected number of events for each observed event time in the data set, assuming that the chances of the event, given that subjects are at risk, are the same in the groups. The total number of expected events is then computed for each group by adding the expected number of events for each event time. The test finally compares the observed number of events in each group with the expected number of events using a χ^2 test with $k - 1$ degrees of freedom; see [1, 2].

The log-rank test statistic, X^2 , weights contributions from all failure times equally. Several alternative test statistics have been proposed that give differential weights to the failure times. For example, the *generalized Wilcoxon test* (or *Breslow test*) uses weights equal to the number at risk. For the heroin addicts data in Table 1 the log-rank test ($X^2 = 27.9$ on one degree of freedom, $p < 0.0001$) detects a significant clinic difference in favor of longer treatment durations in clinic 2. The Wilcoxon test puts relatively more weight on differences between the survival curves at earlier times but also reaches significance ($X^2 = 11.6$ on one degree of freedom, $p = 0.0007$).

Modeling Survival Times

Modeling survival times is useful especially when there are several explanatory variables of interest. For example, the methadone treatment durations of the heroin addicts might be affected by the prognostic variables maximum methadone dose and prison record, as well as the clinic attended. The main approaches used for modeling the effects of covariates on survival can be divided roughly into two classes – *models for the hazard function* and *models for the survival times themselves*. In essence, these models act as analogies of multiple linear regression for survival times containing censored observations, for which regression itself is clearly not suitable.

Proportional Hazards Models

The main technique is due to Cox [6] and known as the *proportional hazards model* or, more simply, *Cox’s regression*. The approach stipulates a model for the hazard function. Central to the procedure is the assumption that the hazard functions for two individuals at any point in time are proportional, the so-called *proportional hazards assumption*. In other words, if an individual has a risk of the event at some initial time point that is twice as high as another individual, then at all later times, the risk of the event remains twice as high. Cox’s model is made up of an unspecified *baseline hazard function*, $h_0(t)$, which is then multiplied by a suitable function of an individual’s explanatory variable values, to give the

individual's hazard function. Formally, for a set of p explanatory variables, $x_1, x_2, \dots, x_p$, the model is

$$h(t) = h_0(t) \exp \left(\sum_{i=1}^p \beta_i x_i \right) \quad (9)$$

where the terms $\beta_1, \dots, \beta_p$ are the parameters of the model that have to be estimated from sample data. Under this model, the *hazard* or *incidence rate ratio*, h_{12} , for two individuals, with covariate values $x_{11}, x_{12}, \dots, x_{1p}$ and $x_{21}, x_{22}, \dots, x_{2p}$ given by

$$h_{12} = \frac{h_1(t)}{h_2(t)} = \exp \left[\sum_{i=1}^p \beta_i (x_{1i} - x_{2i}) \right] \quad (10)$$

does not depend on t . The interpretation of the parameter β_i is that $\exp(\beta_i)$ gives the incidence rate change associated with an increase of one unit in x_i , all other explanatory variables remaining constant. Specifically, in the simple case of comparing hazards between two groups, $\exp(\beta)$, measures the hazard ratio between the two groups. The effect of the covariates is assumed to be multiplicative.

Cox's regression is considered a *semiparametric* procedure because the baseline hazard function, $h_0(t)$, and by implication the probability distribution of the survival times, does not have to be specified. The baseline hazard is left unspecified; a different parameter is essentially included for each unique survival time. These parameters can be thought of as nuisance parameters whose purpose is merely to control the parameters of interest for any changes in the hazard over time. Cox's regression model can also be extended to allow the baseline hazard function to vary with the levels of a stratification variable. Such a *stratified proportional hazards model* is useful in situations where the stratifier is thought to affect the hazard function, but the effect itself is not of primary interest.

A Cox regression can be used to model the methadone treatment times from Table 1. The model uses prison record and methadone dose as explanatory variables, whereas the variable clinic, whose effect was not of interest, merely needed to be taken account of and did not fulfill the proportional hazards assumption, was used as a stratifier. The estimated regression coefficients are shown in Table 2. The coefficient of the prison record indicator variable is 0.389 with a standard error of 0.17. This translates into a hazard ratio of $\exp(0.389) = 1.475$ with a 95% confidence interval ranging from 1.059 to 2.054. In other words, a prison record is estimated to increase the hazard of immediate treatment cessation by 47.5%. Similarly, the hazard of treatment cessation was estimated to be reduced by 3.5% for every extra milligram per day of methadone prescribed.

Statistical software packages typically report three different tests for testing regression coefficients, the *likelihood ratio (LR)* test, the *score test* (which is equivalent to the log-rank test for Cox's proportional hazards model), and the *Wald test*. The test statistic of each of the tests can be compared with a χ^2 distribution to derive a P value. The three tests are asymptotically equivalent but differ in finite samples. The LR test is generally considered the most reliable, and the Wald test, the least. Here, presence of a prison record tests statistically significant after adjusting for clinic and methadone dose (LR test: $X^2 = 5.2$ on one degree of freedom, $p = 0.022$) and so does methadone dose after adjusting for clinic and prison record (LR test: $X^2 = 30.0$ on one degree of freedom, $p < 0.0001$).

Cox's model does not require specification of the probability distribution of the survival times. The hazard function is not restricted to a specific form and, as a result, the semiparametric model has flexibility and is widely used. However, if the assumption of a particular probability distribution for the data is valid, inferences based on such an assumption

Table 2 Parameter estimates from Cox regression of treatment duration on maximum methadone dose prescribed and presence of a prison record stratified by clinic

Predictor variable	Effect estimate			95% CI for $\exp(\beta)$	
	Regression coefficient ($\hat{\beta}$)	Standard error ($\sqrt{\text{var}(\hat{\beta})}$)	Hazard ratio ($\exp(\hat{\beta})$)	Lower limit	Upper limit
Prison record	0.389	0.17	1.475	1.059	2.054
Maximum methadone dose	-0.035	0.006	0.965	0.953	0.978

are more precise. For example, estimates of hazard ratios or median survival times will have smaller standard errors. A *fully parametric proportional hazards model* makes the same assumptions as Cox's regression, but in addition also assumes that the baseline hazard function, $h_0(t)$, can be parameterized according to a specific model for the distribution of the survival times. Survival time distributions that can be used for this purpose, that is, those that have the *proportional hazards property*, are principally the *exponential*, *Weibull*, and *Gompertz* distributions. Different distributions imply different shapes of the hazard function, and, in practice the distribution that best describes the functional form of the observed hazard function is chosen – for details see [1, 2].

Models for Direct Effects on Survival Times

Accelerated failure time models form a family of fully parametric models that assume direct multiplicative effects of covariates on survival times and hence do not rely on proportional hazards. A wider range of survival time distributions possesses the *accelerated failure time property*, principally, the *exponential*, *Weibull*, *log-logistic*, *generalized gamma*, or *lognormal* distributions. In addition, this family of parametric models includes distributions (e.g., the log-logistic distribution) that model unimodal hazard functions, while all distributions suitable for the proportional hazards model imply hazard functions that increase or decrease monotonically. The latter property might be limiting, for example, for modeling the hazard of dying after a complicated operation that peaks in the postoperative period.

The general accelerated failure time model for the effects of p explanatory variables, $x_1, x_2, \dots, x_p$, can be represented as a loglinear model for survival time,

T , namely,

$$\ln(T) = \alpha_0 + \sum_{i=1}^p \alpha_i x_i + \text{error} \quad (11)$$

where $\alpha_1, \dots, \alpha_p$ are the unknown coefficients of the explanatory variables and α_0 an intercept parameter. The parameter α_i reflects the effect that the i th covariate has on log-survival time with positive values indicating that the survival time increases with increasing values of the covariate and *vice versa*. In terms of the original time scale, the model implies that the explanatory variables measured on an individual act multiplicatively, and so affect the speed of progression to the event of interest.

The interpretation of the parameter α_i , then, is that $\exp(\alpha_i)$ gives the factor by which any survival time percentile (e.g., the median survival time) changes per unit increase in x_i , all other explanatory variables remaining constant. Expressed differently, the probability, that an individual with covariate value $x_i + 1$ survives beyond t , is equal to the probability, that an individual with value x_i survives beyond $\exp(-\alpha_i)t$. Hence $\exp(-\alpha_i)$ determines the change in the speed with which individuals proceed along the timescale and the coefficient is known as the *acceleration factor* of the i th covariate.

Software packages typically use the loglinear formulation. The regression coefficients from fitting a log-logistic accelerated failure time model to the methadone treatment durations using prison record and methadone dose as explanatory variables and clinic as a stratifier are shown in Table 3. The negative regression coefficient for prison suggests that the treatment durations tend to be shorter for those with a prison record. The positive regression coefficient for dose suggests that treatment durations tend to be prolonged for those on larger methadone doses. The estimated acceleration factor for an individual with

Table 3 Parameter estimates from log-logistic accelerated failure time model of treatment duration on maximum methadone dose prescribed and presence of a prison record stratified by clinic

Predictor variable	Effect estimate			95% CI for $\exp(\alpha)$	
	Regression coefficient ($\hat{\alpha}$)	Standard error ($\sqrt{\text{var}(\hat{\alpha})}$)	Acceleration factor ($\exp(-\hat{\alpha})$)	Lower limit	Upper limit
Prison record	-0.328	0.140	1.388	1.054	1.827
Maximum methadone dose	0.0315	0.0055	0.969	0.959	0.979

a prison record compared with one without such a record is $\exp(0.328) = 1.388$, that is, a prison record is estimated to accelerate the progression to treatment cessation by a factor of about 1.4. Both explanatory variables, prison record (LR test: $X^2 = 5.4$ on one degree of freedom, $p = 0.021$) and maximum methadone dose prescribed (LR test: $X^2 = 31.9$ on one degree of freedom, $p < 0.0001$) are found to have statistically significant effects on treatment duration according to the log-logistic accelerated failure time model.

Summary

Survival analysis is a powerful tool for analyzing time to event data. The classical techniques Kaplan–Meier estimation, proportional hazards, and accelerated failure time modeling are implemented in most general purpose statistical packages, with the S-PLUS and R packages having particularly extensive facilities for fitting and checking nonstandard Cox models, see [7]. The area is complex and one of active current research. For additional topics such as other forms of censoring and *truncation* (delayed entry), *recurrent events*, models for *competing risks*, *multistate models* for different transition rates, and *frailty models* to include random effects the reader is referred to [7–10].

References

- [1] Collett, D. (2003). *Modelling Survival Data in Medical Research*, 2nd Edition, Chapman & Hall/CRC, London.
- [2] Hosmer, D.W. & Lemeshow, S. (1999). *Applied Survival Analysis*, John Wiley & Sons, New York.
- [3] Caplethorn, J. (1991). Methadone dose and retention of patients in maintenance treatment, *Medical Journal of Australia* **154**, 195–199.
- [4] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation for incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.
- [5] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariance test procedures (with discussion), *Journal of the Royal Statistical Society, Series A* **135**, 185–206.
- [6] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **74**, 187–220.
- [7] Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer, New York.
- [8] Andersen, P.K. (ed) (2002). *Multistate models, Statistical Methods in Medical Research*, Arnold Publishing, Vol. 11, London.
- [9] Crowder, K.J. (2001). *Classical Competing Risks*, Chapman & Hall/CRC, Boca Raton.
- [10] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer, New York.

SABINE LANDAU

Article originally published in Encyclopedia of Statistics in Behavioral Science (2005, John Wiley & Sons, Ltd).

Syndromic Surveillance

Syndromic surveillance (see **Managing Infrastructure Reliability, Safety, and Security; Game Theoretic Methods; Digital Governance, Hotspot Detection, and Homeland Security; Sampling and Inspection for Monitoring Threats to Homeland Security**) has been defined as “the ongoing, systematic collection, analysis, interpretation, and application of real-time (or near-real-time) indicators of diseases and outbreaks that allow for their detection before public health authorities would otherwise note them” [1]. It has also been defined as “...surveillance using health-related data that precedes diagnosis and signals a sufficient probability of a case or an outbreak to warrant further public health response” [2]. These definitions focus on a number of concepts important to syndromic surveillance:

1. Syndromic surveillance is public health surveillance, not military, regulatory, or intelligence surveillance.
2. The surveillance is based on the ongoing, routine collection of “health-related” data, such as counts of individuals coming into medical facilities, over-the-counter medication sales, and aggregate laboratory test results, that is generally collected for some purpose other than surveillance.
3. The data and associated surveillance are intended to provide leading indicators that “precede diagnosis” or “case” confirmation with the goal of providing early warning of an outbreak.
4. The system must provide timely signals that trigger “further public health response” while simultaneously minimizing the number of false alarms [2].

A *syndrome* is “A set of symptoms or conditions that occur together and suggest the presence of a certain disease or an increased chance of developing the disease” [3]. In the context of syndromic surveillance, a syndrome is a set of nonspecific pre-diagnosis medical and other information that may indicate the release of a bioterrorism agent or natural disease outbreak (see, for example, Syndrome Definitions for Diseases Associated with Critical

Bioterrorism-associated Agents [4]). The data in syndromic surveillance systems may be clinically well-defined and linked to specific types of outbreaks, such as groupings of ICD-9 codes from emergency room “chief complaint” data, or only vaguely defined and perhaps only weakly linked to specific types of outbreaks, such as over-the-counter sales of cough and cold medication or absenteeism rates.

Since its inception, syndromic surveillance has focused mainly on *early event detection*: (see **Fault Detection and Diagnosis; Behavioral Decision Studies; Public Health Surveillance**) gathering and analyzing data in advance of diagnostic case confirmation to give early warning of a possible outbreak. Such early event detection is not supposed to provide a definitive determination that an outbreak is occurring. Rather, it is supposed to signal that an outbreak *may* be occurring, indicating a need for further evidence or triggering an investigation by public health officials (i.e., the Centers for Disease Control and Prevention (CDC) or a local or state public health department).

More recently, the purpose of syndromic surveillance has been expanded to include using “existing health data in real time to provide immediate analysis and feedback to those charged with investigation and follow-up of potential outbreaks” [5]. This broader focus on *electronic biosurveillance* includes both early event detection and *situational awareness* [2]. Situational awareness is the real-time analysis and display of health data to monitor the location, magnitude, and spread of an outbreak, as well as the availability and application of public health and medical resources in response to the outbreak. As Bravata *et al.* [6] said, “...an essential component of preparations for illnesses and syndromes potentially related to bioterrorism includes the deployment of surveillance systems that can rapidly detect *and monitor* [emphasis added] the course of an outbreak and thus minimize associated morbidity and mortality”.

Syndromic surveillance is one form of *public health surveillance* (see **Lead; Cohort Studies; Digital Governance, Hotspot Detection, and Homeland Security**), which is the “ongoing, systematic collection, analysis, interpretation, and dissemination of data regarding a health-related event for use in public health action to reduce morbidity and mortality and to improve health” [7]. In terms of bioterrorism, syndromic surveillance uses medical and health-related data, which is distinct from other types of detection

methods such as air or water sampling for pathogens (e.g., the BioWatch program [8, 9]). Some syndromic surveillance systems may be useful for epidemic monitoring, though such monitoring can be and is currently conducted *via* a variety of methods.

Syndromic surveillance is an emerging tool in epidemiology – the study of the distribution, determinants, and occurrence of disease and health-related conditions in populations [10, p. 1]. However, there are a number of characteristics that distinguish syndromic surveillance from public health surveillance and epidemiology as they have traditionally been practiced. For example, syndromic surveillance uses nonspecific health-related data. Traditional notifiable disease reporting is based on suspected or confirmed cases. Syndromic surveillance systems are intended to actively search for evidence of possible outbreaks, where the type of outbreak is not specified in advance. In conventional public health surveillance, it is unusual to initiate active surveillance without a suspected outbreak.

Traditional epidemiological methods tend to focus on retrospective studies initiated to address a specific population risk event. In comparison, syndromic surveillance systems designed for early event detection do not have a specific risk event focus. This lack of a specific focus leads to difficult questions about how to aggregate data – syndromically, spatially, and temporally – and to questions of what types of models and detection methods are most appropriate. Given that the data is usually in-hand, in classic retrospective epidemiological studies these types of questions can be directly resolved. Yet, conducting a retrospective study to try to understand the cause of a disease or outbreak can still be quite challenging. Implementing and operating a system intended to effectively detect a wide variety of possible outbreaks that have not yet occurred, from naturally occurring diseases to bioterrorism attacks, is a daunting task.

Surveillance Systems

The CDC and many state and local health departments around the United States are actively developing and fielding electronic biosurveillance systems, such as the CDC's BioSense system. Sosin [11] states that approximately 100 state and local health jurisdictions were conducting some form of syndromic

surveillance in 2003. In 2004, Bravata *et al.* [6] conducted a systematic review of the publicly available literature and various websites from which they identified 115 surveillance systems, of which they found 29 that were designed specifically for detecting bioterrorism.

Examples of Systems Currently in Use

Below are brief descriptions of three surveillance systems chosen to illustrate large-scale systems currently in operation. The first two are true systems, in the sense that they are comprised of both dedicated computer hardware and software. The third is more properly described as a set of software programs that can be freely downloaded and implemented by any public health organization.

- **BioSense** (www.cdc.gov/biosense)
Developed and operated by the CDC, BioSense is intended to be a United States-wide electronic biosurveillance system. Begun in 2003, BioSense initially used Department of Defense and Department of Veterans Affairs outpatient data along with medical laboratory test results from a nationwide commercial laboratory. In 2006, BioSense began incorporating data from civilian hospitals as well. The primary objective of BioSense is to “expedite event recognition and response coordination among federal, state, and local public health and healthcare organizations” (J. Tokars, Personal communication, 2006) [12, 13].
- **ESSENCE** (www.geis.fhp.osd.mil/GEIS/SurveillanceActivities/ESSENCE/ESSENCE.asp)
ESSENCE is an acronym for *Electronic Surveillance System for the Early Notification of Community-based Epidemics*. Developed by the Department of Defense in 1999, ESSENCE IV now monitors for infectious disease outbreaks at more than 300 military treatment facilities worldwide on a daily basis using data from patient visits to the facilities and from pharmacy data. For the Washington DC area, ESSENCE II monitors military and civilian outpatient visit data as well as over-the-counter pharmacy sales and school absenteeism [14–16]. Though ESSENCE has gone through a number of iterations, the subsequent versions have been strongly influenced by the original military project (H. Burkom, Personal communication, 2006).

- **EARS** (www.bt.cdc.gov/surveillance/ears/) EARS is an acronym for *Early Aberration Reporting System*.

Developed by the CDC, EARS was designed for monitoring for bioterrorism during large-scale events that often have little or no baseline data (i.e., as a short-term *drop-in* surveillance method) [17]. For example, the EARS system was used in the aftermath of Hurricane Katrina to monitor communicable diseases in Louisiana [18], for syndromic surveillance at the 2001 Super Bowl and World Series, as well as at the Democratic National Convention in 2000 [19]. Though developed as a drop-in surveillance method, EARS is now being used on an ongoing basis in many syndromic surveillance systems.

Descriptions of some state and local syndromic surveillance systems (as well as other information) can be found in [20, 21].

Components of Electronic Biosurveillance Systems

Electronic biosurveillance systems comprise a series of functional components. Expanding on Rolka [22], an ideal system will contain all of the components listed below.

- The original data, of which access is gained only after appropriately addressing legal and regulatory requirements, as well as personal privacy and proprietary issues.
- Computer hardware and information technology for (near) real-time assembly, recording, transfer, and preprocessing of data.
- Data management subject matter experts, software, and techniques for processing incoming data into analytic databases, as well as processes and procedures for managing and maintaining these databases.
- Statistical algorithms to analyze the data for possible outbreaks over space and time that are of sufficient sensitivity to provide signals within an actionable timeframe while simultaneously limiting false positive signals to a tolerable level.
- Public health experts with sufficient statistical expertise that can appropriately choose and apply the algorithms most relevant to their jurisdiction and appropriately interpret the signals when they occur.

- Data display and query software, as well as the necessary underlying data, that facilitates rapid and easy investigation and adjudication of signals by public health experts, including the ability to “trace back” from signal to likely source.
- Other data displays, combined with decision support and communication tools, to support situational awareness during an outbreak to facilitate effective and efficient public health response.

Mandl *et al.* [23], in *Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience*, provides a detailed discussion of what is required and guidance about how to implement syndromic surveillance systems.

Quantitative Methods Used in Biosurveillance Systems for Early Event Detection

Biosurveillance systems use a variety of temporal and spatial methods for early event detection. As is discussed in more detail below, most of these have been adapted from other fields and applications. Buckeridge *et al.* [24] provides a useful classification of common surveillance scenarios and a mapping of some early event detection methods to those scenarios.

Situational awareness is an emerging concept in electronic biosurveillance and, as such, specific methods for assessing and displaying relevant information has not yet been incorporated into the systems. Currently, situational awareness is limited to displaying the spatial distribution of data *via* geographic information system (GIS) software (see, for example, Li *et al.* [25]).

Temporal Methods

Most syndromic surveillance systems apply variants of the standard univariate statistical process control (SPC) methods (*see Credit Scoring via Altman Z-Score; Reliability Integrated Engineering Using Physics of Failure; Sampling and Inspection for Monitoring Threats to Homeland Security*): Shewhart, cumulative sum (CUSUM), and/or exponentially weighted moving average (EWMA) charts. Woodall [26] provides a comprehensive overview of the application of control charts to health surveillance. Montgomery [27] is an excellent introduction

to these methods in a SPC setting. Shmueli and Fienberg [28] and Shmueli [29] give a review of these and other methods potentially applicable to early event detection in a biosurveillance setting.

The challenge in applying these methods is that syndromic surveillance generally violates classical SPC assumptions, particularly the assumptions of normality and independent and identically distributed observations. Biosurveillance data is generally autocorrelated and frequently has seasonal periodicities. Further, given the goal of quick detection, methods are usually run on individual observations for which an assumption of normality generally does not apply.

In spite of this, the standard SPC methods are sometimes applied with little modification (see, for example, Fricker [30] or Stoto *et al.* [31]) and in some cases the methods are modified to attempt to account for the autocorrelation. For example, EARS [17] applies variants of the Shewhart chart (see equations (1) and (2)) and a cumulative method motivated by the CUSUM (see equation (3)). These methods use various moving windows of data to estimate the process mean and standard deviation [19] and they are intended to be used when little historic (“baseline”) information is available.

The EARS methods are called C_1 , C_2 , and C_3 and are defined as follows [32] (L. Hutwagner, Personal communication, 2006). Let $X(t)$ be an observation for period t , for example the number of individuals presenting to a particular hospital with a specific syndrome on day t . The C_1 method calculates the statistic $C_1(t)$ for day t as

$$C_1(t) = \frac{X(t) - \bar{X}_1(t)}{s_1(t)} \quad (1)$$

where $\bar{X}_1(t) = \sum_{i=t-1}^{t-7} X(i)/7$ and $s_1(t) = \sqrt{\sum_{i=t-1}^{t-7} (X(i) - \bar{X}_1(t))^2/6}$. It signals an alarm at time t when the C_1 statistic exceeds a *threshold* which is fixed at three standard deviations above the moving sample average: $C_1(t) > 3$.

The C_2 method is similar to the C_1 method, but incorporates a 2-day lag in the mean and standard deviation calculations. It calculates

$$C_2(t) = \frac{X(t) - \bar{X}_3(t)}{s_3(t)} \quad (2)$$

where $\bar{X}_3(t) = \sum_{i=t-3}^{t-9} X(i)/7$ and $s_3(t) = \sqrt{\sum_{i=t-3}^{t-9} (X(i) - \bar{X}_3(t))^2/6}$, and signals an alarm when $C_2(t) > 3$.

Finally, the C_3 method uses the C_2 statistics from the past 3 days, calculating the statistic $C_3(t)$ for day t as

$$C_3(t) = \sum_{i=t-3}^{t-2} \max[0, C_2(i) - 1] \quad (3)$$

It signals an alarm when $C_3(t) > 2$.

BioSense originally implemented the C_1 , C_2 , and C_3 methods, but has since modified the C_2 . Calling the new method “W2”, it calculates the mean and standard deviation separately for weekdays and weekends (using the relevant last 7 days with a 2-day lag). Following the suggestion of Buckeridge *et al.* [24], it uses an empiric method to define the threshold and users of the system can select the threshold as an option [33].

To date, multivariate SPC methods have not been incorporated into operational electronic biosurveillance systems. Some research has been conducted to define and evaluate directional multivariate methods, including Fricker *et al.* [34], Fricker [30], Stoto *et al.* [31], and Joner *et al.* [35]. Whether or not multivariate methods provide additional sensitivity to detect outbreaks compared to the current practice of using multiple simultaneous univariate methods is yet to be conclusively demonstrated. Various methods have also been proposed to combine and/or adjust for the application of multiple univariate methods in multivariate settings. See the discussion in Rolka *et al.* [36] about parallel and consensus monitoring methods and Stoto *et al.* [31, 37] for a discussion and an evaluation of some of these methods.

Regression and time series methods have also been proposed to explicitly model, and hence account for, seasonality. The basic idea is to model the disease incidence process, perhaps including terms in the model for annual seasonal variations, monthly variations, and even day of the week and holiday variations. The model is then used to either: (a) predict the expected number of events and the difference between the expected number and observed number is monitored for excessive deviations or, (b) model and then remove the explainable/known effects (sometimes called *preconditioning*) and then the residuals are monitored using traditional SPC

methods. This is consistent with recommendations by Montgomery [27] for autocorrelated data.

Examples in the literature include Fricker *et al.* [38], Brillman *et al.* [39], who apply the CUSUM to the prediction errors, the CDC's cyclical regression models discussed in Hutwagner *et al.* [19], log-linear regression models in Farrington *et al.* [40], and time series models in Reis and Mandl [41]. See Shmueli [29] for additional discussion of the use of regression and time series methods for syndromic surveillance and Burkom *et al.* [42] for a comparison of two regression-based methods and an exponential smoothing method applied to biosurveillance forecasting. Also see Lotze *et al.* [43] for a detailed discussion of preconditioning applied to syndromic surveillance data.

Other temporal methods that have been proposed or are in use for syndromic surveillance include: wavelets (see Goldberg *et al.* [44], Zhang *et al.* [45], Shmueli [46], and the discussion in Shmueli [29]); Bayesian networks (see Wong *et al.* [47], Rolka *et al.* [36]); hidden Markov models (see Le Strat and Carrat [48]); Bayesian dynamic models (see, for example, Sebastiani *et al.* [49]), and rule-based methods (see, for example, Wong *et al.* [50]).

Spatial and Spatiotemporal Methods

Kleinman *et al.* [51] and Lazarus *et al.* [52] proposed a generalized linear mixed model (GLMM) (see **Bayesian Statistics in Quantitative Risk Assessment**) to simultaneously monitor disease counts over time in a region divided into smaller subareas (zip codes). It is statistically attractive because it uses information across the entire region while appropriately adjusting for the smaller areas (see **Statistics for Environmental Toxicity; Environmental Remediation**).

As described in Kleinman *et al.* [51], there are two forms of the model depending on whether individual data and covariates are available versus aggregated counts and covariates by zip code. In the former case, the model is

$$E(y_{ijt}|b_i) = p_{ijt} \quad \text{and} \quad \text{logit}(p_{ijt}) = \mathbf{x}_{ijt}\boldsymbol{\beta} + b_i \quad (4)$$

where y_{ijt} is an indicator for whether or not person j in area i is a case on day t , p_{ijt} is the probability he or she is a case, $\mathbf{x}_{ijt}$ is a vector of observed covariates on person j and/or area i over time up to and including

day t , $\boldsymbol{\beta}$ is a vector of fixed effects, and b_i is a random effect for area i . When no individual level covariate information is available, the most likely situation, the model is

$$E(y_{it}|b_i) = n_{it} p_{it} \quad \text{and} \quad \text{logit}(p_{it}) = \mathbf{x}_{it}\boldsymbol{\beta} + b_i \quad (5)$$

where $y_{it} = \sum_{j=1}^{n_{it}} y_{ijt}$. In this model, p_{it} can be thought of as the probability that each individual in area i will be a case on day t .

Having fit the model in equation (5), z cases are observed on day $t + 1$. The rarity of the observed count is assessed by calculating

$$\Pr(Z \geq z \text{ cases}) = 1 - \sum_{k=1}^{z-1} \binom{n_{it}}{k} \hat{p}_{it}^k (1 - \hat{p}_{it})^{n_{it}-k} \quad (6)$$

where $\hat{p}_{it}$ is calculated from the estimated coefficients in the usual way for logistic regression and $1/(\hat{p}_{it} \times \text{number of tests conducted})$ is proposed as the *recurrence interval*: the number of time periods for which the expected number of counts of z or more cases is one (Woodall *et al.* [53]). Waller [54] recommends an alternate calculation for the recurrence interval and Woodall *et al.* [53] take issue with both the use of and recommended calculations for the recurrence interval.

The small area regression and testing ("SMART") method in BioSense (see the BioSense User Guide [55]) is based on the Kleinman *et al.* [51] and Lazarus *et al.* [52] GLMM approach. However, as implemented in BioSense it only uses spatial information to bin data into separate time series, the output of which are subsequently combined using a Bonferroni correction. Hence, the BioSense SMART method is properly classified as a temporal method (H. Burkom, Personal communication, 2006).

The most commonly used spatial method is the scan statistic, particularly as implemented in the SaTScan software (www.satscan.org). Originally developed to retrospectively identify disease clusters (see Kulldorff [56]), the method is now regularly used prospectively in electronic biosurveillance systems (see Kulldorff [57]). For example, it was used as part of a drop-in syndromic surveillance system in New York City after the 9/11 attack (Ackelsberg *et al.* [58]). While it has been studied by the BioSense program, it has not yet been implemented in the BioSense system interface (see Bradley [59] and the BioSense User Guide [55]).

The basic idea in SaTScan is to count the number of cases that occur in a cylinder, where the circle is the geographic base and the height of the cylinder corresponds to time. The cylinder is passed over space, varying the radius of the circle (up to a maximum radius that includes 50% of the monitored population) and the height of the cylinder and counts of cases for those geographic regions whose centroids fall within the circle for the period of time specified by the height of the cylinder are summed. When used for prospective biosurveillance, the start date of the height of the cylinder is varied but the end date is fixed at the most current time period. Conditioning on the expected spatial distribution of observations, SaTScan reports the most likely cluster (in both space and time) and its p value.

Though widely used, some aspects of the prospective application of the SaTScan methodology have been questioned, particularly the use of recurrence intervals and performance comparisons between SaTScan and other methods. See Woodall *et al.* [53] for further details. Also, see Kulldorff [57] for other methods for disease mapping and for testing whether an observed pattern of disease is due to chance.

Olson *et al.* [60] and Forsberg *et al.* [61] take an alternate approach to assessing possible disease clusters using M-statistics based on the distribution of pairwise distances between cases. That is, let $\mathbf{X} = \{X_1, \dots, X_n\}$ represent the locations of n cases on the plane and let $\mathbf{d} = \{d_1, \dots, d_{\binom{n}{2}}\}$ be the $\binom{n}{2}$ interpoint distances, then the M-statistic is

$$M = (\mathbf{o} - \mathbf{e})^T \hat{S}^{-1} (\mathbf{o} - \mathbf{e}) \quad (7)$$

where $\mathbf{o}$ and $\mathbf{e}$ are vectors of observed and expected counts of binned interpoint distances and $\hat{S}^{-1}$ is the inverse of the estimate of the covariance matrix.

Other approaches to spatial and spatiotemporal biosurveillance methods include the repeated two-sample rank procedure of Fricker and Chang [62], the automated epidemiologic geotemporal integrated surveillance system (AEGIS) by Olson *et al.* [63] and the application of CUSUM methods to the spatial distribution of cases (see Rogerson and Yamada [64]). See Lawson and Kleinman [65] for additional exposition and methods, and Mandl *et al.* [23] for further discussion on spatial and spatiotemporal modeling issues. For spatial methods with application to more traditional public health data and problems, see Waller and Gotway [66].

Issues and Challenges in Biosurveillance

A number of papers discuss the various issues, challenges, and important research needs associated with effective implementation and operation of electronic biosurveillance systems. These include Sosin [1], Fricker and Rolka [2], Bravata *et al.* [6], Rolka [22], Shmueli [29], Stoto *et al.* [32], and Buehler *et al.* [67]. Research challenges span many disciplines and problems, from legal and regulatory challenges that must be resolved to gain access to data, to the technological challenges of designing the computer hardware and software for collecting and assembling the data, to the ethical and procedural issues for managing and safeguarding the data, to the quantitative challenges of analyzing the data for potential outbreaks and displaying the data to enhance situational awareness, to the managerial challenges of effectively assembling and operating the entire system.

Much of the controversy surrounding syndromic surveillance stems from the initial focus on early event detection, a use that requires a number of (as yet unproven) assumptions, including:

- Leading indicators of outbreaks exist in prediagnosis health-related data of sufficient strength that they are statistically detectable.
- The leading indicators occur sufficiently far in advance of clinical diagnoses so that, when found, they provide the public health community with sufficient advance notice to take action.

Reingold [68] has suggested that a compelling case for the implementation of syndromic surveillance systems is yet to be made and Stoto *et al.* [37] have questioned whether syndromic surveillance systems can achieve an effective early detection capability.

However, a myopic focus only on early event detection for bioterrorism in syndromic surveillance systems misses other important benefits electronic biosurveillance can provide, particularly the potential to significantly advance and modernize the practice of public health surveillance. For example, whether or not electronic biosurveillance systems prove effective at the early detection of bioterrorism, they are likely to have a significant and continuing role in the detection of seasonal and pandemic flu, as well as other naturally occurring disease outbreaks. This

latter function is echoed in an Institute of Medicine report on *Microbial Threats to Health* [69]:

“[S]yndromic surveillance is likely to be increasingly helpful in the detection and monitoring of epidemics, as well as the evaluation of healthcare utilization for infectious diseases.”

In terms of bioterrorism, Stoto [70] states that electronic biosurveillance systems build links between public health and healthcare providers – links that could prove to be critical for consequence management should a bioterrorism attack occur. Furthermore, Sosin [1] points out that electronic biosurveillance systems can:

- act as a safety net, should the existing avenues of detection fail to detect an outbreak, so countermeasures can be taken swiftly;
- provide additional lead time to public health authorities so that they can take more effective public health actions.

The safety net idea says that detection in a biosurveillance system does not necessarily have to be earlier than the first clinical diagnosis to be useful. To illustrate, a Dutch biologist conducting automated salmonella surveillance related that a surveillance system detected outbreaks whose occurrence was somehow missed by sentinel physicians (H. Burkom, Personal communication, 2006). And the idea of additional lead time is that unusual indicators in an electronic biosurveillance system (not necessarily a signal from an early event algorithm) may give public health organizations time to begin organizing and marshaling resources in advance of a confirmed case and/or provide critical information about how and where to apply the resources.

Conclusion

Electronic biosurveillance systems are being developed and implemented around the world. They are motivated by a need for improved public health surveillance, not only for bioterrorism, but also to improve detection and responsiveness to natural disease outbreaks such as avian influenza and severe acute respiratory syndrome (SARS), and as such they hold great promise as public health tools.

For those interested in additional information, the International Society for Disease Surveillance (<http://syndromic.org>) is a professional society devoted to advancing the field of disease surveillance, and *Advance in Disease Surveillance* (www.isdsjournal.org/) is its journal.

Acknowledgments

I would like to thank Howard Burkom, Lori Hutwagner, Michael Jackson, Galit Shmueli, Mike Stoto, Jerry Tokars, Lance Waller, and Bill Woodall for their helpful and thoughtful comments on an early version of this article. It has been significantly improved as a result of their input. Any errors or omissions that may remain are solely my responsibility.

References

- [1] Sosin, D.M. (2003). Syndromic surveillance: the case for skillful investment view, *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* **1**, 247–253, at <http://www.medscape.com/viewarticle/466780> (accessed Nov 2006).
- [2] Fricker Jr, R.D. & Rolka, H. (2006). Protecting against biological terrorism: statistical issues in electronic biosurveillance, *Chance* **19**, 4–13.
- [3] International Foundation for Functional Gastrointestinal Disorders, at <http://www.iffgd.org/GIDisorders/glossary.html> (accessed Nov 2006).
- [4] Centers for Disease Control and Prevention (2003). *Syndrome Definitions for Diseases Associated with Critical Bioterrorism-Associated Agents*, October 23, 2003, at <http://www.bt.cdc.gov/surveillance/syndromedef/> (accessed Nov 2006).
- [5] Hennings, K.J. & Centers for Disease Control and Prevention (2004). What is syndromic surveillance? Syndromic surveillance: reports from a National Conference, 2003, *Morbidity and Mortality Weekly Report* **53**, 7–11.
- [6] Bravata, D.M., McDonald, K.M., Smithe, W.M., Rydzak, C., Szeto, H., Buckeridge, D.L., Haberland, C. & Owens, D.K. (2004). Systematic review: surveillance systems for early detection of bioterrorism-related diseases, *Annals of Internal Medicine* **140**(11), 910–922.
- [7] The Centers for Disease Control and Prevention (2001). Updated guidelines for evaluating public health surveillance systems, *Morbidity and Mortality Weekly Report* **50**(RR13), 1–35, at <http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm> (accessed Nov 2006).
- [8] Chertoff, M. (2006). *Testimony Before the Senate Committee on Homeland Security and Governmental Affairs*, released date September 12, 2006, at http://www.dhs.gov/xnews/testimony/testimony_1158336548990.shtm (accessed Nov 2006).

- [9] Shea, D.A. & Lister, S.A. (2003). The BioWatch program: detection of bioterrorism, Report No. RL 32152, Congressional Research Service, November 19, 2003, at <http://www.fas.org/sgp/crs/terror/RL32152.html> (accessed Nov 2006).
- [10] Gerstman, B.B. (1998). *Epidemiology Kept Simple: An Introduction to Classic and Modern Epidemiology*, John Wiley & Sons, New York.
- [11] Sosin, D. & Centers for Disease Control and Prevention (2005). Evaluation challenges for syndromic surveillance—making incremental progress, *Morbidity and Mortality Weekly Report* **53**, 125–129.
- [12] Tokars, J. (2006). The BioSense application, *Presentation at the 2006 PHIN Conference*, Atlanta, at <http://www.cdc.gov/biosense/presentations.htm> (accessed Dec 2007).
- [13] The Centers for Disease Control and Prevention. *Bio Sense*, at <http://www.cdc.gov/biosense/> (accessed Nov 2006).
- [14] Department of Defense, at <http://www.geis.fhp.osd.mil/GEIS/SurveillanceActivities/ESSENCE/ESSENCE.asp> (accessed Nov 2006).
- [15] Lombardo, J.S., Burkom, H. & Pavlin, J. (2004). ESSENCE II and the framework for evaluating syndromic surveillance systems, *Morbidity and Mortality Weekly Report* **53**, 159–165, at http://www.cdc.gov/mmwr/mmwr_sup.html (accessed Nov 2006).
- [16] Office of the Secretary of Defense. *ESSENCE IV Improves Nation's Bio-Surveillance Capability*, at http://deploymentlink.osd.mil/news/jan05/news_20050125_001.shtml (accessed Nov 2006).
- [17] The Centers for Disease Control and Prevention (2006). *Early Aberration Reporting System Website*, at <http://www.bt.cdc.gov/surveillance/ears/> (accessed Nov 2006).
- [18] Toprani, A., Ratard, R., Straif-Bourgeois, S., Sokol, T., Averhoff, F., Brady, J., Staten, D., Sullivan, M., Brooks, J.T., Rowe, A.K., Johnson, K., Vranken, P., Sergienko, E. & Centers for Disease Control and Prevention (2006). Surveillance in hurricane evacuation centers—Louisiana, September–October 2005, *Morbidity and Mortality Weekly Report* **55**, 32–35.
- [19] Hutwagner, L., Thompson, W., Seeman, G.M. & Treadwell, T. (2003). The bioterrorism preparedness and response Early Aberration Reporting System (EARS), *Journal of Urban Health: Bulletin of the New York Academy of Medicine* **80**(2 Suppl.1), 89i–96i.
- [20] The Centers for Disease Control and Prevention (2004). *Morbidity and Mortality Weekly Report* **53**, at http://www.cdc.gov/mmwr/preview/ind2004_su.html (accessed 2004) (supplement).
- [21] The Centers for Disease Control and Prevention (2006). *Annotated Bibliography for Syndromic Surveillance*, at <http://www.cdc.gov/EPO/dphsi/syndromic/evaluation.htm> (accessed Nov 2006).
- [22] Rolka, H. (2006). Data analysis research issues and emerging public health biosurveillance directions, in *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, A. Wilson, G. Wilson & D.H. Olwell, eds, Springer, New York, pp. 101–107.
- [23] Mandl, K.D., Overhage, J.M., Wagner, M.W., Lober, W.B., Sebastiani, P., Mostashari, F., Pavlin, J.A., Geste-land, P.H., Treadwell, T., Koski, E., Hutwagner, L., Buckeridge, D.L., Aller, R.D. & Grannis, S. (2004). Implementing syndromic surveillance: a practical guide informed by the early experience, *The Journal of the American Medical Informatics Association* **11**, 141–150, at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=353021> (accessed Nov 2006).
- [24] Buckeridge, D.L., Burkom, H., Campbell, M., Hogan, W.R. & Moore, A.W. (2004). Algorithms for rapid outbreak detection: a research synthesis, *Journal of Biomedical Informatics* **38**, 99–113.
- [25] Li, H., Faruque, F., Williams, W. & Finley, R. (2006). Real-time syndromic surveillance, *ArcUser: The Magazine for ESRI Software Users*, 17–19, January–March issue at <http://www.esri.com/news/arcuser/0206/winter2006.html>.
- [26] Woodall, W.H. (2006). The use of control charts in health-care and public-health surveillance, *Journal of Quality Technology* **38**, 1–16, at <http://www.asq.org/pub/jqt>.
- [27] Montgomery, D.C. (2001). *Introduction to Statistical Quality Control*, 4th Edition, John Wiley & Sons, New York.
- [28] Shmueli, G. & Fienberg, S.E. (2006). Current and potential statistical methods for monitoring multiple data streams for biosurveillance, in *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, A. Wilson, G. Wilson & D.H. Olwell, eds, Springer, New York, pp. 109–140.
- [29] Shmueli, G. (2006). Statistical challenges in modern biosurveillance, *Technometrics* (in submission).
- [30] Fricker, Jr, R.D. (2007). Directionally sensitive multivariate statistical process control methods with application to syndromic surveillance, *Advances in Disease Surveillance* **3**(1), at <http://www.isdsjournal.org/>.
- [31] Stoto, M.A., Fricker Jr, R.D., Jain, A., Diamond, A., Davies-Cole, J.O., Glymph, C., Kidane, G., Lum, G., Jones, L., Dehan, K. & Yuan, C. (2006). Evaluating statistical methods for syndromic surveillance, *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, A. Wilson, G. Wilson & D.H. Olwell, eds, Springer, New York, pp. 141–172.
- [32] Hutwagner, L., Browne, T., Seeman, G.M. & Fleischauer, A.T. (2005). Comparing aberration detection methods with simulated data, *Emerging Infectious Diseases* **11**, at <http://www.cdc.gov/ncidod/EID/vol11no02/04-0587.htm> (accessed Nov 2006).
- [33] The Centers for Disease Control and Prevention (2006). *BioSense Bulletin*, at <http://www.cdc.gov/biosense/communications.htm>.
- [34] Fricker, Jr, R.D., Knitt, M.C. & Hu, C.X. Directionally sensitive MCUSUM and MEWMA procedures with

- application to biosurveillance, in submission to *Quality Engineering*.
- [35] Joner Jr, M.D., Woodall, W.H., Reynolds Jr, M.R. & Fricker Jr, R.D. (2006). *The Use of Multivariate Control Charts in Public Health Surveillance*.
- [36] Rolka, R., Burkom, H., Cooper, G.F., Kulldorff, M., Madigan, D. & Wong, W. (2007). Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: research needs, *Statistics in Medicine* **26**(8), 1834–1856.
- [37] Stoto, M.A., Schonlau, M. & Mariano, L.T. (2004). Syndromic surveillance: Is it worth the effort? *Chance* **17**, 19–24.
- [38] Fricker Jr, R.D., Hegler, B.L. & Dunfee, D.A., (2008). Comparing syndromic surveillance detection methods: EARS' versus a CUSUM-based methodology, *Statistics in Medicine*. DOI:10.1002/sim3197.
- [39] Brillman, J.C., Burr, T., Forslund, D., Joyce, E., Picard, R. & Umland, E. (2005). Modeling emergency department visit patterns for infectious disease complaints: results and application to disease surveillance, *BMC Medical Informatics and Decision Making* **5**, at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=555597> (accessed Nov 2006).
- [40] Farrington, C.P., Andrews, N.J., Beale, A.D. & Catchpole, M.A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease, *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **159**, 547–563.
- [41] Reis, B.Y. & Mandl, K.D. (2003). Time series modeling for syndromic surveillance, *BMC Medical Informatics for Decision Making* **3**, at <http://www.biomedcentral.com/1472-6947/3/2> (accessed Nov 2006).
- [42] Burkom, H.S., Murphy, S.P. & Shmueli, G. (2007). Automated time series forecasting for biosurveillance (draft), *Statistics in Medicine* **26**(22), 4202–4218.
- [43] Lotze, T., Murphy, S.P. & Shmueli, G. (2006). Preparing biosurveillance data for classic monitoring (draft), *Advances in Disease Surveillance* (in submission).
- [44] Goldberg, A., Shmueli, G., Caruana, R.A. & Fienberg, S.E. (2002). Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales, *Proceeding of the National Academy of Sciences* **99**, 5237–5240, at <http://www.pnas.org/cgi/reprint/99/8/5237> (accessed Nov 2006).
- [45] Zhang, J., Tsui, F., Wagner, M. & Hogan, W. (2003). Detection of outbreaks from time series data using wavelet transform, *AMIA Annual Symposium Proceedings* 748–752, at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=14728273> (accessed Nov 2006).
- [46] Shmueli, G. (2005). Wavelet-based monitoring for modern biosurveillance, Technical report RHS-06-002, University of Maryland, Robert H. Smith School of Business, at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=902878 (accessed Nov 2006).
- [47] Wong, W., Cooper, G., Dash, D., Levander, J., Dowling, J., Hogan, W. & Wagner, M. (2005). Use of multiple data streams to conduct Bayesian biologic surveillance, *Morbidity and Mortality Weekly Report* **54**(suppl.), 63–69, at <http://www.cdc.gov/MMWR/preview/mmwr.html/su5401a12.htm> (accessed Dec 2007).
- [48] Le Strat, Y. & Carrat, F. (1999). Monitoring epidemiologic surveillance data using hidden Markov models, *Statistics in Medicine* **18**, 3463–3478, at <http://www3.interscience.wiley.com/cgi-bin/fulltext/68502848/PDFSTART> (accessed Nov 2006).
- [49] Sebastiani, P., Mandl, K.D., Szolovits, P., Kohane, I.S. & Ramoni, M.F. (2006). A Bayesian dynamic model for influenza (with discussion), *Statistics in Medicine* **25**, 1803–1825.
- [50] Wong, W., Moore, A., Cooper, G. & Wagner, M. (2003). WSARE: what's strange about recent events? *Journal of Urban Health: Bulletin of the New York Academy of Medicine* **80**, 66i–75i, at <http://rods.health.pitt.edu/LIBRARY/WSARE%20What's%20Strange%20%20J%20Urban%20Health%20Cooper%202003.pdf> (accessed Nov 2006).
- [51] Kleinman, K., Lazarus, R. & Platt, R. (2004). A generalized mixed model approach for detecting incident clusters of disease in small areas, with an application to biological terrorism, *American Journal of Epidemiology* **159**, 217–224.
- [52] Lazarus, R., Kleinman, K., Dashevsky, I., Adams, C., Kludt, P., DeMaria Jr, A. & Platt, R. (2002). Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events, *Emerging Infectious Diseases* **8**, 753–760, at http://www.medscape.com/viewarticle/440756_print (accessed Nov 2006).
- [53] Woodall, W.H., Marshall, J.B., Joner Jr, M.D., Fraker, S.E. & Abdel-Salam, A.G. (2007). On the use of scan methods in prospective public health surveillance, *Journal of the Royal Statistical Society, Series A (Statistics in Society)* (to appear).
- [54] Waller, L.A. (2004). Invited commentary: syndromic surveillance – some statistical comments, *American Journal of Epidemiology* **159**, 225–227.
- [55] The Centers for Disease Control and Prevention (2006). *BioSense User Guide*, Version 2.0, August 2006, at http://www.cdc.gov/biosense/files/CDC_BioSense_User_Guide_VA_DoD_LabCorp_v2.05.pdf (accessed Dec 2007).
- [56] Kulldorff, M. (1997). A spatial scan statistic, *Communications in Statistics, Theory and Methods* **26**, 1481–1496, at <http://www.satscan.org/papers/k-cstm1997.pdf> (accessed Nov 2006).
- [57] Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **164**, 61–72, at <http://www.satscan.org/papers/k-jrssa2001.pdf> (accessed Nov 2006).
- [58] Ackelsberg, J., Balter, S., Bornschelgel, K., Carubis, E., Cherry, C., Das, D., Fine, A., Karpati, A., Layton, M.,

- Mostashari, F., Nivin, B., Reddy, V., Weiss, D., Hutwagner, L., Seeman, G.M., McQuiston, J., Treadwell, T., Rhodes, J. & Centers for Disease Control and Prevention (2002). Syndromic surveillance for bioterrorism following the attacks on the world trade center – New York city, 2001, *Morbidity and Mortality Weekly Report (Special Issue)* **51**, 13–15.
- [59] Bradley, C. (2005). Visualizing and monitoring data in BioSense using SaTScan, *Presentation at the 2005 Syndromic Surveillance Conference*, Seattle at <http://www.cdc.gov/biosense/presentations.htm> (accessed Nov 2006).
- [60] Olson, K.L., Bonetti, M., Pagano, M. & Mandl, K.D. (2005). Real time spatial cluster detection using inter-point distances among precise patient locations, *BMC Medical Informatics and Decision Making* **5**, 19, at <http://www.biomedcentral.com/1472-6947/5/19> (accessed Dec 2006).
- [61] Forsberg, L., Jeffery, C., Ozonoff, A. & Pagano, M. (2006). A spatiotemporal analysis of syndromic data for biosurveillance, in *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, A. Wilson, G. Wilson & D.H. Olwell, eds, Springer, New York, pp. 173–191.
- [62] Fricker Jr, R.D., & Chang, J.T., A spatio-temporal methodology for real-time biosurveillance, in submission to *Quality Engineering*.
- [63] Olson, K.L., Bonetti, M., Pagano, M. & Mandl, K.D. (2002). Enhanced power to detect bioterrorism with spatial clustering (abstract), *Pediatric Research* **51**, 94.
- [64] Rogerson, P.A. & Yamada, I. (2004). Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches, *Statistics in Medicine* **23**, 2195–2214.
- [65] Lawson, A.B. & Kleinman, K. (eds) (2005). *Spatial and Syndromic Surveillance for Public Health*, John Wiley & Sons.
- [66] Waller, L.A. & Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*, John Wiley & Sons.
- [67] Buehler, J.W., Berkelman, R.L., Hartley, D.M. & Peters, C.J. (2003). Syndromic surveillance and bioterrorism-related epidemics, *Emerging Infectious Diseases* **9**, 1197–1204.
- [68] Reingold, A. (2003). If syndromic surveillance is the answer, what is the question? *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* **1**, 1–5, at <http://www.medscape.com/viewarticle/458654> (accessed Nov 2006).
- [69] Smolinski, M.S., Hamburg, M.A. & Lederberg, J. (eds) (2003). *Microbial Threats to Health: Emergence, Detection, and Response*, National Academies Press.
- [70] Stoto, M.A. (2006). *Syndromic Surveillance in Public Health Practice*, presentation to Institute of Medicine Forum on Microbial Threats.

RONALD D. FRICKER

Systems Reliability

Systems reliability assessment (see **Reliability Growth Testing; Reliability Data; Mathematics of Risk and Reliability: A Select History**) is the term commonly applied to the prediction of system performance parameters such as the likelihood or frequency of system failure in a specified mode. Depending on the task the system performs, the relevant system failure parameter could be the probability of failure at a point in time (unavailability), the chance that the system has not functioned without failure over a period of time (unreliability), or the rate at which the system fails (failure rate). System quantification can also produce importance measures, that indicate the contribution that each component makes to the system failure mode. These measures provide a numerical indicator between 0 and 1 and account for the component failure likelihood and the system structure. If the component can, on its own, cause the system failure, it is usually ranked higher than if it were, for example, part of a redundant structure and would need to fail in conjunction with other components to cause system malfunction.

Systems are designed to perform a specific function and when they fail they can fail in different ways known as modes. For example, a safety system designed to shut a process down when a hazard occurs can fail in a way where it is incapable of detecting and responding to the hazard or it can fail spuriously and produce a shutdown when the hazard does not exist. The system assessment needs to focus on one particular failure mode rather than a generic failure. The system is made up of components that can also fail in different ways. The state of the system is a function of the state of the components. The system failure modes are caused by component failures either occurring individually or in combinations.

Many systems are safety critical and have the potential to fail in a way that results in fatalities. For this type of system, its reliability or failure frequency is commonly assessed as part of a risk assessment undertaken to ensure that an acceptable level of performance can be expected. The system analysis is best carried out as part of the system design process, when there is the greatest flexibility to effect change. When used as part of a risk assessment, the probability or frequency of a hazardous event

is judged in conjunction with the consequences of the event. Risk [1] is defined as $\text{risk} = \text{frequency} \times \text{consequences}$.

There may be several potential hazards for a system, and its total risk is the sum of the risks posed by each hazard. This measure is then used to judge the acceptability of the system performance.

For new system designs, while the system in its entirety does not exist, the components used in its construction, or similar ones, are likely to be part of other systems. The failure characteristics of the components can be studied and, using the historical data, failure time distributions can be constructed. Taking account of the maintenance process to give times to repair, the failure likelihood of each component can be predicted. Systems reliability uses the historical performance of the components and the proposed structure or architecture of the system (defining how the components are linked).

There are a variety of methods that can be used to predict the system failure probability or frequency. They make different assumptions about the way the system is operated and maintained. The selection of one method rather than another to perform the analysis is dependent upon how closely the assumptions of the method match the operating condition for the system. There can be two outcomes from the analysis, qualitative and quantitative. The qualitative analysis yields the combinations of component events that cause the system failure mode of concern. These are termed minimal cut sets and are formally defined and discussed later in this article. Quantification produces the system failure likelihood and frequency measures.

This article describes the alternative modeling techniques for systems reliability assessment. An overview of the methods is provided, along with the assumptions implicit in each method. Guidelines on when it is appropriate to select a particular method for a system assessment are then given. A detailed presentation of the methods presented here can be found in [2, 3].

Combinatorial Methods

Combinatorial methods are a category of techniques which determine the combinations of component failures which can cause the system failure as an intermediate step in producing system performance predictions. Appropriate models are used to calculate

the likelihood of each component failure mode. Then, assuming all components fail independently (the failure of one component does not affect the likelihood of failure of any other) the component failure parameters (see **Simulation in Risk Management; Reliability Data; Markov Modeling in Reliability**) are used in conjunction with the list of the failure combinations to evaluate the likelihood of system failure. Within this class of techniques are reliability networks and the fault tree method. The assumption of component failure independence is central to these techniques, which differ in the way they represent the system architecture.

The models used to calculate the component failure probabilities are common to both methods. Usually, as in the models given in equations (1) and (2), they introduce another assumption of a constant failure rate, though this is not a limitation of the combinatorial methods and alternative models can be used.

Revealed Component Failures

When the system operation is such that a component failure is revealed, then its repair can be immediately instigated. For components that have a constant failure rate, λ , and repair rate, ν , the probability, $q(t)$, that it has failed at time t is given by

$$q(t) = \frac{\lambda}{\lambda + \nu} (1 - e^{-(\lambda + \nu)t}) \quad (1)$$

For nonrepairable components, $\nu = 0$.

Dormant Component Failures

When the component is not actively involved in the system operation at the point that it fails, for example, part of a standby system, it may not be noticed that the failure has occurred. The fact that this type of component has failed is usually discovered when a demand is placed on it to function or when a test identifies the failure. For such components tested at regular time intervals of θ with mean time to repair τ , the average component unavailability q_{AV} is approximated by

$$q_{AV} = \lambda \left(\frac{\theta}{2} + \tau \right) \quad (2)$$

Reliability Networks (Reliability Block Diagrams)

The Reliability Network approach, sometime known as the *reliability block diagram* (see **Multistate Systems; Probabilistic Risk Assessment**) or RBD [4] method was the first technique used to assess system reliability. The network consists of nodes, which represent the component failure modes, and edges, which link the component failure events according to the structure of the system.

A typical reliability network is shown in Figure 1. A “start node” is placed in the far left of the network and an “end node”, at the far right. Nodes are then incorporated in the network according to how the component failure mode that they represent influences the system functionality. If a component functions then there is a path through the component node. Paths through the network go in a direction from left to right. If vertical edges appear in the network, flow through them can be in either direction. The occurrence of the component failure mode means that there can be no path through the associated node. If there is a path through the network that connects the start node to the end node, then the system is considered to function. System failure occurs when the component failures in the system cause a cut through the network that isolates the start node from the end node. For fully redundant systems where it requires all components to fail in order to cause failure of the system, the reliability network is a parallel structure. Nonredundant systems that cannot tolerate the failure of any component result in a series system. For the network shown in Figure 1, components A and B are in parallel, and this section is in series with component C. Should components A and B both fail or component C fail on its own, there would be no path through the network, and under either of these conditions, the system can be considered to have failed.

For quantification of the system failure probability (see **Repair, Inspection, and Replacement Models**;

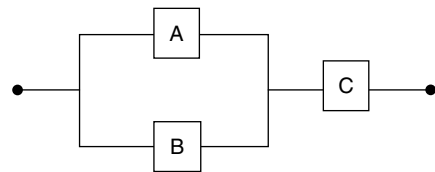


Figure 1 Simple reliability network

Imprecise Reliability; Markov Modeling in Reliability), the network can be categorized as either “simple” or “complex”. Simple networks are those whose structure is made from independent series and parallel sections. “Complex” networks are those that do not fall into this category.

The unreliability of simple systems can be obtained by taking each series or parallel section (of, say, n components) and replacing it by a “virtual” component having the same reliability characteristics as the section. So, for series sections, which are nonredundant, the section unreliability Q_{ser} can be calculated from

$$Q_{\text{ser}} = 1 - \prod_{i=1}^n (1 - Q_i) \quad (3)$$

Parallel, fully redundant sections, have unreliability Q_{par} given by

$$Q_{\text{par}} = \prod_{i=1}^n Q_i \quad (4)$$

where Q_i is the failure probability of each of the n components in the section.

The process continues until there is a single virtual component between the start and end nodes of the network. The reliability of this virtual component is the reliability of the system. For example, in the network shown in Figure 1, the reduction is performed in two stages. The first calculates the unreliability of the parallel structure containing components A and B. In the second step, this new virtual component is treated as being in series with component C. This yields the system unreliability for the network. While this method works for simple network configurations, a generalized method is required that should work whatever the structure of the network. This is based on finding the causes of system failure.

The network can be analyzed to produce the system failure modes known as *minimal cut sets*. The concept of a *cut set* is defined as a set of component failure modes which, if they occur together, would cause the system to fail in the mode of concern. The cut set is, however, of little use as it may exceed the required conditions for the system to fail. For example, {A,C} is a cut set of the network in Figure 1. As soon as component C has failed, however, the system has already failed, and the failure of component A is not a necessary condition

for system failure. The more useful concept is that of the *minimal cut set*; this is a minimal (necessary and sufficient) combination of component failures, which, if they occur, would result in system failure. For the network of Figure 1, there are two minimal cut sets {A,B} and {C}. The name cut set is used because these combinations cause a cut through the network.

The reliability network also has the capability of being viewed from the complementary viewpoint of system failure, the system success. It is possible to determine the working component conditions required for system functionality. This gives rise to the concept of a *path set* which is a combination of working component states that guarantee that the system works. A *minimal path set* is a minimal set of functioning components that will cause the system to function. For the system shown in Figure 1 the minimal path sets are {A,C} and {B,C}.

Consider the well-known “bridge” network structure, which is shown in Figure 2. This network cannot be sequentially reduced in series and parallel sections to determine the system reliability.

Any network in this category is classed as “complex”, and the approach is to first determine the list of minimal cut sets, $C_i, i = 1, \dots, N_c$, where N_c is the total number of minimal cut sets and + represents the logical OR operator. The causes of system failure, SYS, can then be expressed as

$$SYS = C_1 + C_2 + \dots + C_{N_c} \quad (5)$$

The likelihood for system failure is then given in by the inclusion–exclusion expansion:

$$\begin{aligned} Q_{\text{sys}} = & \sum_{i=1}^{N_c} P(C_i) - \sum_{i=2}^{N_c} \sum_{j=1}^{i-1} P(C_i \cap C_j) \\ & + \sum_{i=3}^{N_c} \sum_{j=2}^{i-1} \sum_{k=1}^{j-1} P(C_i \cap C_j \cap C_k) - \dots \\ & + (-1)^{N_c+1} P(C_1 \cap C_2 \dots \cap C_{N_c}) \end{aligned} \quad (6)$$

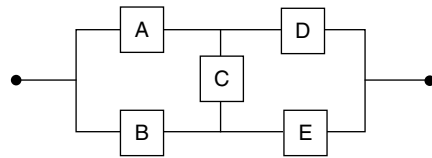


Figure 2 The classical “bridge” network

For the bridge system shown in Figure 2, there are four minimal cut sets: {A,B}, {A,C,E}, {B,C,D}, and {D,E}. If the failure probability of component i is q_i , then the system failure probability is (after simplification)

$$\begin{aligned} Q_{\text{SYS}} = & q_A q_B + q_D q_E + q_A q_C q_E + q_B q_C q_D \\ & - q_A q_B q_D q_E - q_A q_B q_C q_E - q_A q_B q_C q_D \\ & - q_A q_C q_D q_E - q_B q_C q_D q_E - 2 q_A q_B q_C q_D q_E \quad (7) \end{aligned}$$

This small example still has 15 terms in the full expansion of equation (7). For all, but the smallest of systems, this expansion becomes prohibitively expensive in terms of computer processing time even for fast digital computers, and approximations are required.

The approximation needs to be an upper bound to take a pessimistic view of the system performance, especially when the system being analyzed is a safety system. A common approximation is the minimal cut set upper bound, Q_{MCSU} , given in equation (8). This provides a good estimation of the system failure probability and, when the minimal cut sets are independent (do not have any component failure events in common) the equation is exact.

$$Q_{\text{MCSU}} \leq 1 - \prod_{i=1}^{N_c} (1 - P(C_i)) \quad (8)$$

For the bridge network problem, the minimal cut set upper bound is:

$$\begin{aligned} Q_{\text{SYS}} \approx & 1 - (1 - q_A q_B)(1 - q_D q_E) \\ & (1 - q_A q_C q_E)(1 - q_B q_C q_D) \quad (9) \end{aligned}$$

Fault Tree Analysis

Fault tree analysis (*see Canonical Modeling; Fault Detection and Diagnosis; Expert Judgment*) is the most commonly used technique to assess the failure probability of a specific system failure mode in terms of the failure likelihood of its components. The concept was developed by Watson [5] at Bell Telephone Laboratories in 1961. The time-dependent mathematical theory, known as *kinetic tree theory*, was developed by Vesely at Idaho Nuclear Laboratories and appeared almost 10 years after the initial conception [6]. Kinetic tree theory remains the method used

today in many of the commercially available software tools that perform fault tree analysis.

The fault tree diagram is an inverted tree structure that starts with the system failure mode event, known as *the top event*. Its branches then spread downwards, systematically developing the failure logic, firstly, in terms of intermediate system events down to component failure events, known as *basic events*, which terminate the branches as illustrated in Figure 3 (developing the failure logic in sections from left to right for convenience of presentation).

The fault tree diagram features two types of symbols: events and gates. The event symbols enable the failure logic to be documented as shown. The logic expressing how the events combine to cause the events appearing at higher levels in the tree is indicated by gate symbols. Two of the fundamental gate types are the AND gate and the OR gate. The AND gate symbol (as shown for the GATE 1 event in Figure 3) represents a fully redundant structure. It has events from the lower level in the fault tree as inputs, and the higher level event (the gate output) occurs if all inputs to the gate occur. The OR gate (symbol for TOP) is used to represent a nonredundant structure, and the output to the gate occurs when at least one of the input events occurs. There are many other symbols that can be used [3], but they are mainly for the purposes of the documentation.

There are three fundamental mathematical logic operators (AND, OR, and NOT). For certain system types, the use of the NOT gate is required in the analysis [7], which means that NOT component failure events, i.e., components successfully working, contribute to the failure of the system. This is known as *noncoherence*.

Having constructed the fault tree, it can be analyzed qualitatively and quantitatively. The qualitative analysis produces minimal cut sets by performing Boolean reduction on the failure logic structure. Quantitative analysis delivers the system failure mode probability and frequency. It is also possible to extract importance measures.

As an example, consider the simple tank level control system shown in Figure 4 [8]. Initially, the system has the push button contacts open and switches 1 and 2 (SW1, SW2) contacts closed. To start the system, the push button is pressed and held. This energizes relay R1, which closes its contacts and maintains the circuit when the push button is released

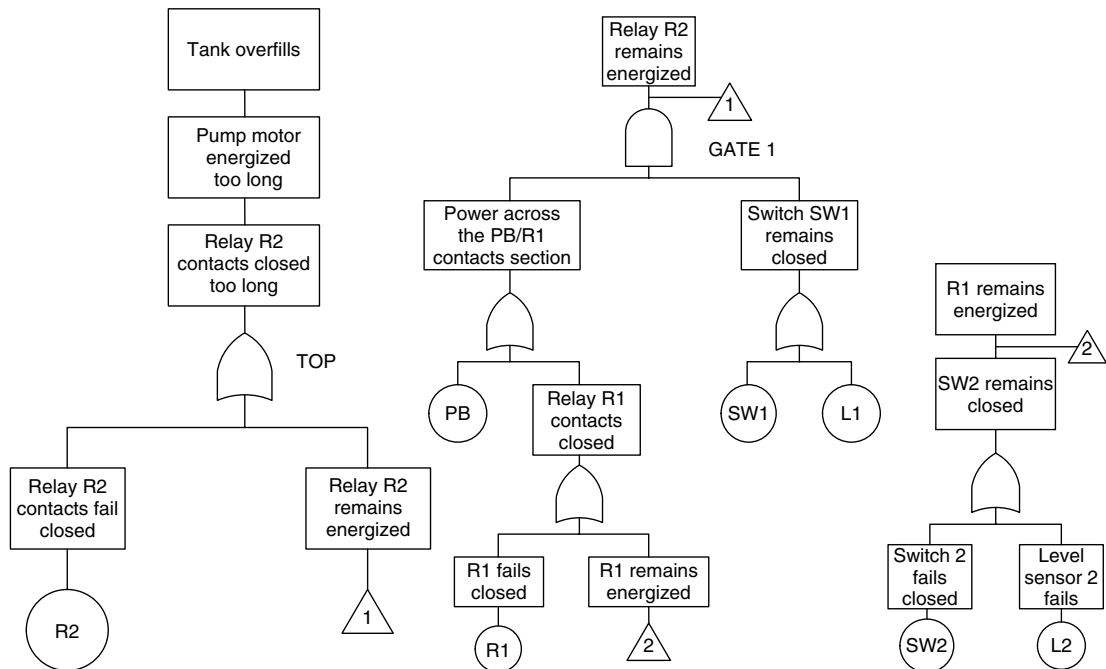


Figure 3 Fault tree for top event “tank overfills”

and opens. Relay R2 is also energized and its contacts close, starting the pump in the second circuit. The pump transfers fluid to the tank. The level of the fluid in the tank is monitored by two level sensors L1 and L2. When the fluid reaches the required level, switch SW1 opens and de-energizes relay R2, turning off the pump. When the fluid in the tank is used and the level drops, SW1 closes and pumps fluid to replace what has been used. The normal operation of the system is the switch SW1 opening and closing to turn off and turn on the pump *via* relay R2.

As a safety feature, the second level sensor, L2, is connected to switch SW2. When the fluid level is unacceptably high, SW2 opens and de-energizes relay R1. R1 contacts then open to break the control circuit. This results in R2 de-energizing; its contacts open, removing power from the pump. This would then require a manual restart.

For the system failure mode “tank overfills”, the relevant component failure modes along with the failure rate and repair time data are shown in Table 1. Some of the failure modes are revealed, such as, when relay R2 contacts get stuck and remain closed.

This component condition would mean that the pump keeps running and the problem is revealed by the tank overfilling. Its probability of failure is given by equation (1), where the repair rate is given by the reciprocal of the mean time to failure. Other components, such as relay R1 contacts failing closed are unrevealed, as this is the normal operating state for that component. All of the component failure modes associated with the safety system remain unrevealed, as for this class of events, the failure is revealed only when the component is tested/inspected or when a demand for the component to work occurs. For these component failure events, an inspection interval of 6 months (4380h) is assumed, which enables the probability of the event to be calculated as in equation (2).

The fault tree developed for the undesired top event “tank overfills” is illustrated in Figure 3. The text boxes specify exactly what each gate output event in the fault tree represents. Each branch is developed downward using only AND and OR gates until basic events (component failure events) are encountered and the failure causality development is terminated.

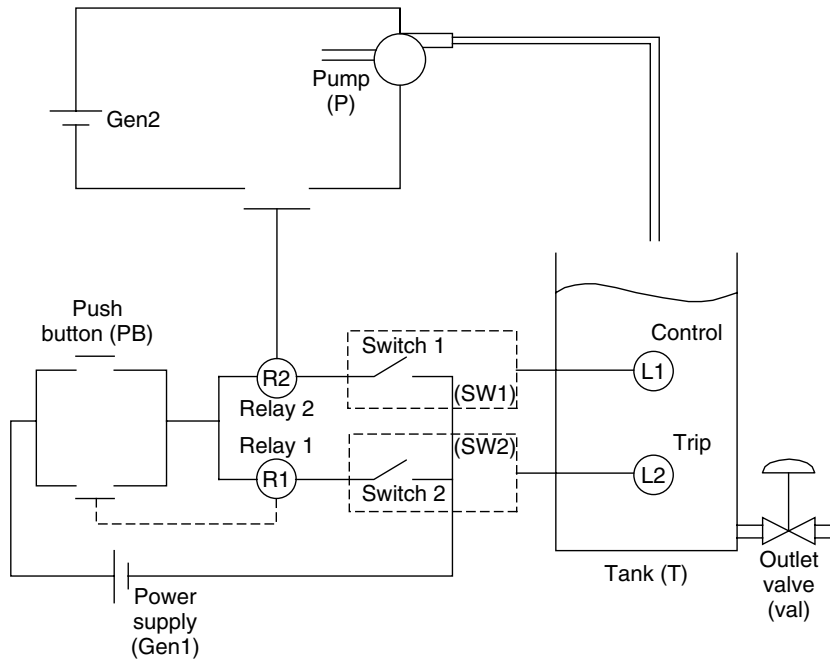


Figure 4 Simple tank level control system

Table 1 Component failure modes and data

Component	Failure mode	Code	Failure rate (h^{-1})	Mean time to repair (h)
Push button	Stuck closed	PB	5×10^{-5}	2
Relay contacts	Stuck closed	R1/R2	6×10^{-5}	10
Switch	Stuck closed	SW1/SW2	5×10^{-5}	10
Level sensors	Fail to indicate high level	L1/L2	2×10^{-6}	5

The final fault tree structure, showing how the basic events combine to cause the system level failure event, is illustrated in Figure 5. In this fault tree, the component failure events appear only once in the structure. Large fault trees, developed for real systems, commonly feature multiple occurrences of basic events in the structure.

Manipulating the Boolean equation for the top event [3] yields the minimal cut sets of the fault tree. For the tank level control system fault tree, the complete list of minimal cut sets are given in Table 2. As can be seen, there are nine failure combinations in all. One is first order (a single event causes system failure), and eight are of order two.

Using the component failure data in Table 1, the system failure parameters can be calculated as

follows:

$$\text{Top event probability} = 7.5 \times 10^{-4}$$

$$\text{Top event frequency} = 7.72 \times 10^{-5} \text{ h}^{-1}$$

If the system failure predictions indicate an unacceptable performance, the weaknesses can be identified using component importance measures. The Fussell–Vesely measure is indicated in Table 3. The Fussell–Vesely measure of importance for each component is the proportion of the system failure probability that is caused by a minimal cut set which contains the event, divided by the total system failure probability, i.e., the probability of the union of the minimal cut sets containing the event divided by

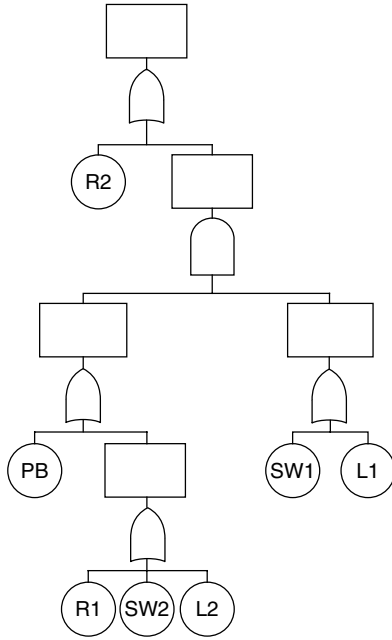


Figure 5 Tank overfill fault tree structure

Table 2 Minimal cut sets

1	R2
2	SW1 PB
3	SW1 R1
4	SW1 SW2
5	SW1 L2
6	L1 PB
7	L1 R1
8	L1 SW2
9	L1 R1

Table 3 Importance measures

Rank	Component	Fussell–Vesely
1	R2	0.781
2	SW1	0.215
3	R1	0.080
4	SW2	0.068
5	PB	0.067
6	L1	0.004
7	L2	0.003

the system failure probability. The importance analysis shows that component R2 provides the biggest contribution to system failure.

The system assessment results presented have been obtained using commercial software [9].

State Space Methods

State space models (*see* **Extreme Event Risk; Digital Governance, Hotspot Detection, and Homeland Security**) identify all possible states in which a system can reside over its lifetime and predicts the probability of its being in each of these states. The system states are usually defined by identifying which of the components are functioning or failed. By examining the component conditions and determining the state of the system, each state is then labeled as a system functioning state or a state in which the system has failed. Having performed the analysis and determined the likelihood of the system being in each state, the system performance prediction is then a matter of adding the relevant state probabilities.

System reliability modeling requires that the states identified for the system must include every possibility (exhaustive) and that the system must only be able to reside in one state at any time (mutually exclusive).

For a Markov approach to be valid two assumptions have to be appropriate:

- The system lacks memory.
- The system is homogeneous.

For the system to be memoryless, it means that the immediate future of the system is governed by its current state. The history of how the system got to reside in its current state is of no consequence. So if we consider the state X_k that the system is in at discrete-time points 1, 2, ..., k , where k is its current state, then

$$P(X_{k+1}|X_k, X_{k-1}, X_{k-2} \dots X_1) = P(X_{k+1}|X_k) \quad (10)$$

For the system to be homogeneous, it means that if it is in any state, the likelihood of moving from this state to any other does not change over time. Hence the transition probabilities or occurrence rates are constant.

Markov analysis is a state space technique commonly used for system reliability modeling. While being discrete in state space, it can be either continuous or discrete with respect to the time domain and both have applications to systems reliability.

Markov models (*see* **Reliability Optimization; Optimal Stopping and Dynamic Programming**) are of particular relevance while modeling the reliability of a system in which some aspect of its design, operation, or maintenance introduces dependencies amongst its components. Typical examples of this

are standby systems and systems in which failed components frequently have to queue for repair. Standby systems feature an operational, or primary component, and a backup, standby, component that is brought into operation to replace the primary component when it fails. For standby situations, the failure characteristics of the backup component can be dependent on the performance of the primary component. When the primary component fails, the backup component becomes operational, and its likelihood of failure increases. To provide greater insight into the method, a Markov model for the warm standby situation is discussed in the next section.

The two Markov approaches, discrete and continuous, are now considered separately.

Continuous-Time Markov Models

An example of a continuous-time Markov model (see **Multistate Models for Life Insurance Mathematics**) is shown in Figure 6. It represents the warm standby system with two identical repairable components A and B. One acts as the operational component at which point the other component acts as its standby. As can be seen, there are five nodes on the model, each representing a different state in which the system can reside. The system states are defined in terms of the states of the two components A and B, which can be operational (O), standby (S), or failed (F). Assume also that the failure rate of both components is λ when operational and $\bar{\lambda}$ when in standby. The rate at which either component is repaired when failed is ν . The five states included in the model are all mutually exclusive so that the system can only reside in one of the alternative states at any one time, and they represent all possible conditions of the system. The states are labeled 1–5 for identification purposes.

The transitions between the states occur when the conditions of the components, and therefore, the system, change due to the occurrence of a failure or repair. Assume that the initial condition of the system is that component A is operational and component B is in standby, state 1 in Figure 6. This situation can change in two ways: the operational component can fail with rate λ , which causes the backup component B to become operational and the failed component to undergo repair (state 3), or the backup component can fail with rate $\bar{\lambda}$, instigating its repair (state 4). The first of these situations is indicated on the Markov

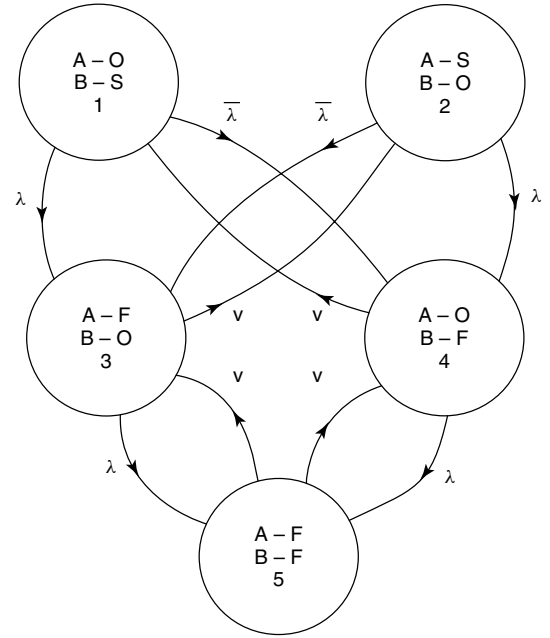


Figure 6 Continuous-time Markov model for a two-component warm standby system

state transition diagram with an arrow (directed edge) from state 1 to state 3, with associated value of λ . The second transition is indicated by an arrow from state 1 is to state 4, with associated value $\bar{\lambda}$.

Only when both components A and B are in the failed state does the system fail. This is represented by state 5 on the Markov model. To determine the availability of the system, it is required to calculate the probability that the system is in the state represented by node 5 on the diagram.

The behavior of the system is represented by a set of first-order linear differential equations which are in matrix form.

$$[\dot{Q}] = [Q] \cdot [A]$$

where $[Q] = [Q_1(t), Q_2(t), Q_3(t), Q_4(t), Q_5(t)]$

$Q_i(t)$ is the probability of being in state i at time t

and $[A] =$

$$\begin{bmatrix} -(\lambda + \bar{\lambda}) & 0 & \lambda & \bar{\lambda} & 0 \\ 0 & -(\lambda + \bar{\lambda}) & \bar{\lambda} & \lambda & 0 \\ 0 & \nu & -(\lambda + \nu) & 0 & \lambda \\ \nu & 0 & 0 & -(\lambda + \nu) & \lambda \\ 0 & 0 & \nu & \nu & -2\nu \end{bmatrix} \quad (11)$$

Note that expressing the equations in this form makes the matrix $\mathbf{A}$ very easy to construct. The off-diagonal terms in row i column j represent the transition rate from state i to state j . The diagonal terms are then calculated so that every row sums to zero.

The solution of the equations yields the time-dependent probabilities of the system residing in each of the states. The system availability is then obtained by summing the probabilities of its being in each of the states that correspond to the system working. Summing the system failure state probabilities gives the system unavailability. Depending on the method employed to solve the equations, time-dependent or steady-state solutions can be obtained.

For some systems, particularly those whose failure may result in fatalities, the requirement is to operate without failure over a period of time. For such systems, it is more appropriate to calculate their unreliability rather than their unavailability. Since system failure cannot be tolerated, the Markov model for the system is modified so that when a system failed state is entered, it cannot make a transition back to a system functioning state. To do this, the transitions that return any system failed state to a state in which the system functions (usually as a result of repair), are removed. This makes the system failed states *absorbing*. Solution of the equations resulting from this model give the system unreliability.

The Markov method is a very flexible technique for system reliability modeling. Its advantage is the ease with which it can accommodate dependencies within the model. For it to be an appropriate technique to use, it does, however, require that the Markov property holds and as a consequence, the failure and repair rates do not vary with time. As such, *wear-out* (increasing failure rate) and *burn-in* (decreasing failure rate) cannot be accommodated. It is also unusual for repair times to be governed by a random process, as implied by the exponential distribution. From a practical viewpoint, a difficulty can arise by the model size for medium to large systems, as the Markov model size increases exponentially with the number of components in the system.

Discrete-Time Markov Models

For discrete-time Markov models, the representation of the states in which the system can reside remains

the same as in its continuous-time counterpart. However, the system state is considered at the end of specified, discrete, time periods, and the values associated with the edges represent the probability of moving from one state to another in this period. An example of a discrete-time Markov model is shown in Figure 7. The number associated with each arc joining the nodes is the probability that in the time period the system moves from the state at the start of the arrow to the state that the arrow points to.

The equations that govern the probability of being in each state after n time periods $Q(n)$ are given by

$$[Q(n)] = [Q(0)][B]^n$$

$$\text{where } [Q(n)] = [Q_1(n), Q_2(n), Q_3(n)]$$

$[Q(0)]$ is the initial state probability vector

$$\text{and } [B] = \begin{bmatrix} 1 - P_1 & P_1 & 0 \\ 0 & 1 - P_2 & P_2 \\ P_4 & P_3 & 1 - P_3 - P_4 \end{bmatrix} \quad (12)$$

Like the continuous-time Markov models, the equations in this form are easily formed. For the matrix $\mathbf{B}$, each off-diagonal term in row i column j is the probability of moving from state i to state j in the discrete-time period considered. The diagonal terms are then calculated to make the rows sum to one and represent the probability that the system remains in that state over the time period.

Solution methods can yield either transient or steady-state probabilities of being in each state,

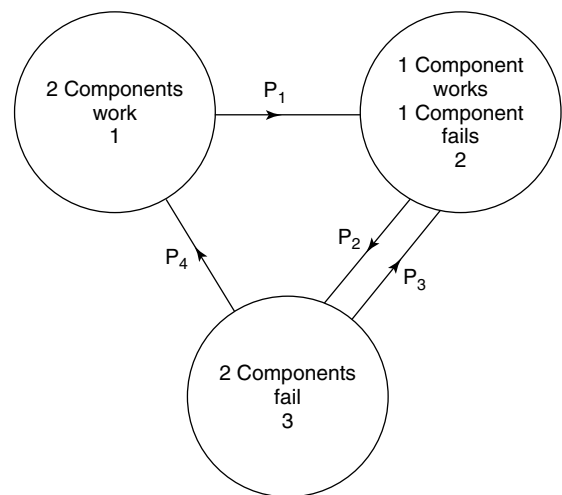


Figure 7 Discrete-time Markov model

as appropriate. For system unreliability prediction, the model is changed to make failed system states absorbing.

Simulation Methods

Monte Carlo Simulation

At times, the characteristics of a system can mean that the assumptions required for the analytical approaches such as the combinatorial or state space methods are not appropriate. There may be, for example, complex dependencies between components in the system or component failure time or repair time distributions of a form that makes the mathematics to predict failure probabilities complex or impossible to perform in a closed form. An alternative, highly flexible method, is Monte Carlo simulation (see **Global Warming; Risk Measures and Economic Capital for (Re)insurers; Simulation in Risk Management; Uncertainty Analysis and Dependence Modeling**). This is a process by which the system behavior is run as an experiment, conducted on the computer. Event times are randomly sampled from the appropriate event distributions. The system structure and operational conditions then determine when the system failure occurs. A single pass through the simulation is governed by the appropriate probability distributions and represents one possible set of events and outcome we could achieve, if we were to operate the system. We need to conduct the experiment a large number of times to get statistically significant results. This process delivers far more detail than just a point estimate of system parameters: we can get distributions of times to system

outcome events. The likelihood of system performance characteristics can be determined as a relative frequency, i.e., the number of simulations that resulted in this outcome divided by the total number of simulations. Because of the large number of simulations that are generally required for the system reliability characteristics to converge, a computer tool is essential for this method to be a practical proposition.

As an example of how simple the simulation concept is, consider a two-component parallel system whose components fail with a fixed probability. Assume the two components, labeled A and B, have failure probabilities 0.1 and 0.05, respectively.

The simulation of the system requires the state of each component to be established. This is achieved using random numbers in the range $[0,1]$. If a number within this range is generated randomly, then all numbers are equally likely. Consider the unit lines for each of the two components shown in Figure 8.

The first component has a 0.1 chance of failure. The chance that a random number generated lies in the range $[0,0.1]$ is 0.1. Therefore, if the random number generated lies in this range the component A is considered to be failed for this particular simulation. A random number in the range $[0.1,1.0]$ would indicate component functionality. For the simulation illustrated, the random numbers have resulted in the failure of component A and the functioning of component B. Since the system is a parallel structure, it continues to be operational with component A failed. This provides one simulation. Repetition of the experiment many times would be required to provide an adequate indication of the expected system performance.

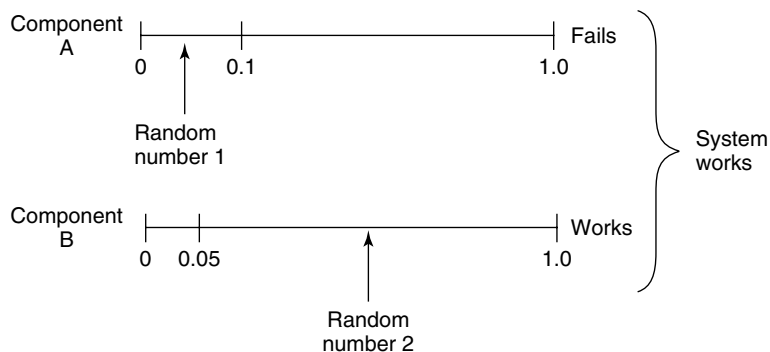


Figure 8 Simulation example – fixed probabilities

More complex system simulations require random samples to be taken from probability distributions representing the time at which events occur. The method used to do this depends on the form of the probability distribution [10]. A process, which works for some distributions, is the inverse transformation method. In this case, a random number is generated that lies in the range $[0,1]$. This variable has the same properties as the cumulative failure probability distribution $F(t)$. Therefore, by generating the random number and equating it to the cumulative probability, the time, t , at which the event occurred can be generated (see Figure 9). Times, generated in this way, conform to the distribution. Where the closed form of the distribution (such as for the exponential and Weibull) can be inverted, this method can be used.

Consider the exponential distribution with failure rate parameter λ given by

$$F(t) = 1 - e^{-\lambda t} \quad (13)$$

Equating $F(t)$ to a random number X in the range $[0,1]$ and rearranging (noting that $1 - X$ is also uniform on $[0,1]$), gives a random sample from this distribution.

$$t = -\frac{1}{\lambda} \ln X \quad (14)$$

Consider a two-component parallel system with identical components, A and B, both with failure rate $1 \times 10^{-4} \text{ h}^{-1}$. Using two random numbers, 0.456 and 0.778, to generate the failure times for A and B,

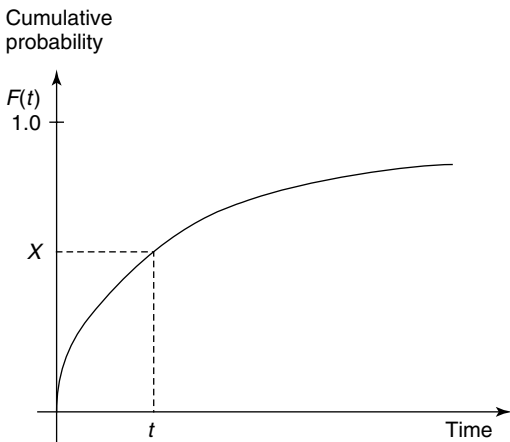


Figure 9 Inverse transformation sampling method

respectively, gives

$$\begin{aligned} t_A &= -\frac{1}{1 \times 10^{-4}} \ln 0.456 = 7853 \text{ h} \\ t_B &= -\frac{1}{1 \times 10^{-4}} \ln 0.778 = 2510 \text{ h} \end{aligned} \quad (15)$$

A parallel, fully redundant, system of nonreparable components works until the last component fails, i.e. for a two-component system, the system failure time $t_{\text{sys}} = \max(t_A, t_B)$. The last component failure time for this particular simulation is 7853 h.

For a series system made of nonreparable components, the failure time for the system is the minimum of the failure times for the components.

Monte Carlo simulation is a very flexible means of modeling the system behavior. No assumptions are required by the analyst, and the difficulty of its implementation is in developing an efficient housekeeping routine to keep track of the events and the system status as the simulation progresses. The disadvantage of the method is the large amounts of simulations that are required to get meaningful statistics on the system performance. For example, consider a system of n components. If the system is such that its actual failure probability was 10^{-4} , then it would need $n \times 10^4$ random samples from probability distributions to be generated before we could expect to produce one system failure event.

Petri Nets

Petri nets (PNs) [11] (see **Human Reliability Assessment, Markov Modeling in Reliability**) are an adaptable and versatile graphical modeling tool used for dynamic system representation. A PN is a bipartite directed graph with two types of node: *places*, which are circular, and *transitions*, shown as bars or rectangles (Figure 10). Places link only to transitions, and *vice versa*, using directed arcs. It is possible for a

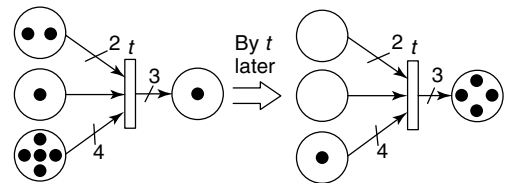


Figure 10 Transition enabling and switching

place to have several arcs to or from the same transition, which is condensed down into a single arc with a *weight* or *multiplicity*, denoted by a slash through the arc with a number next to it. If there is no slash, the multiplicity is assumed to be 1. The dynamic aspect of the PN is formed by *tokens* or *marks*, which abide within places, and are passed between them by the *switching* of transitions.

Figure 10 shows a transition that has three places as inputs. The first place has an arc weight of 2, the middle one has a multiplicity of 1, while the third has a weight of 4. Because the number of tokens in all of the input places to the transition contains at least the weight-number of tokens, the transition can be said to be *enabled*. Transitions can also be associated with a delay, which forces the transition to postpone switching for a period after being enabled. This delay can be zero or randomly sampled from a given distribution. In Figure 10, there is a time delay of t applied once the transition is enabled. Once the time period has passed and the transition remains enabled, the switching takes place. This process removes the number of tokens in each input place corresponding to the multiplicity of the relevant arc, and “creates” the weight-number of tokens in each output place. This is shown in Figure 10 where the switching removes two, one, and four tokens from each of the input places, and deposits three tokens in the output place. The transition is then disabled, as the input places do not have the correct number of tokens.

It is possible to prevent a transition switching by using an *inhibitor arc*. This special arc, shown by a line with a small circle on the end instead of an arrow, connects only an input place to a transition (see Figure 11). It acts such that if the number of tokens within the place is at least that of the arc weighting, the transition cannot switch, regardless of whether it is enabled or not. In Figure 11, the

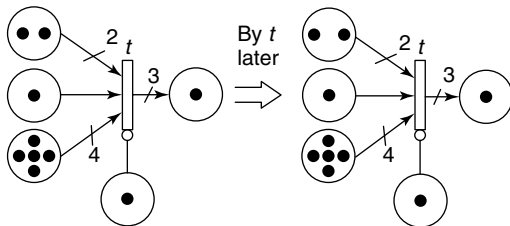


Figure 11 Inhibitor arc preventing switching

otherwise enabled transition can wait for time t to expire, but cannot switch – no tokens are moved by that transition, while the inhibiting place contains the relevant number of tokens.

It is the switching of the transitions that represents the dynamic behavior of the PN model – the ability to transport tokens around the net, thereby changing the *marking* with each switch. The *net marking* is a term given to the distribution of the tokens throughout the whole PN, and each of its forms represents a different system state. This is what is of interest to the analyst.

As a simple illustration of how the PN can be used to model system states, consider the PN shown in Figure 12. This represents a two-component parallel system in which the components A and B are both repairable. Initially, the system has both A and B working, and these states are marked with tokens (as shown). The failure and repair times of the components are F_A , R_A and F_B , R_B . With the input places to transitions 1 and 3 marked, then after F_A and F_B , respectively, these transitions will fire. When transition 1 fires, it removes the token from place 1 and puts tokens in the output places 2 and 5. Should component A then be repaired, by the firing of transition 2, this removes the token from places 2 and 5 and puts a token in place 1. The failure and repair of component B is modeled in the same way. It is only when both components have failed that there are two tokens in place 5, which then fires immediately (zero delay) and places a token in place 6, which represents the system failed state.

Event Tree Analysis

Event tree analysis is a method which forms the cornerstone of many risk assessments. It differs from techniques such as fault tree analysis in that it is an inductive rather than deductive approach and takes into account all possible consequences that can result following the hazard.

An event tree has its branches spreading from left to right across the diagram, an example of which is illustrated in Figure 13.

The event on the far left is known as the initiating event and usually represents the occurrence of a hazard. On the rest of the diagram, each column represents an event which influences the outcome that results following the occurrence of the hazard. Commonly, each column represents a system that is

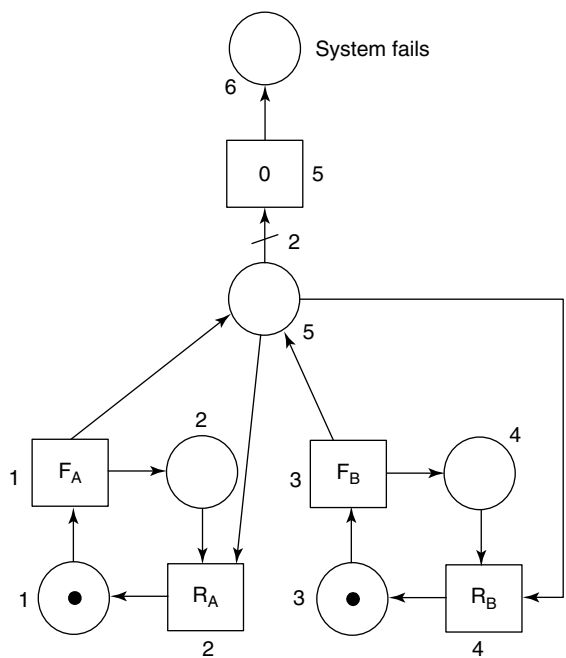


Figure 12 Petri net for a two-component parallel system

designed to respond to the hazard. Within that column, the branches correspond to the success or failure of the system; or a null branch, which means that the system success or failure, given the events which have happened before it, is irrelevant in determining the consequence. For instance, in the small example illustrated in Figure 13, the initiating event is a gas release. In these circumstances, there are three systems that are designed to respond to this hazard. A gas detection system that would establish that the release has occurred, and two systems triggered by the detection system. The first is an isolation system that isolates and shuts down the process, the second is a blowdown system to depressurize the gas line. Both of these measures are designed to limit the quantity of gas that can be released. The gas release event is the event represented on the first single left hand branch. The three systems that are designed to respond to this event are then considered in the order that they operate in the next three columns. The detection system is considered first in the second column. The two branches passing through this column represent the success (top

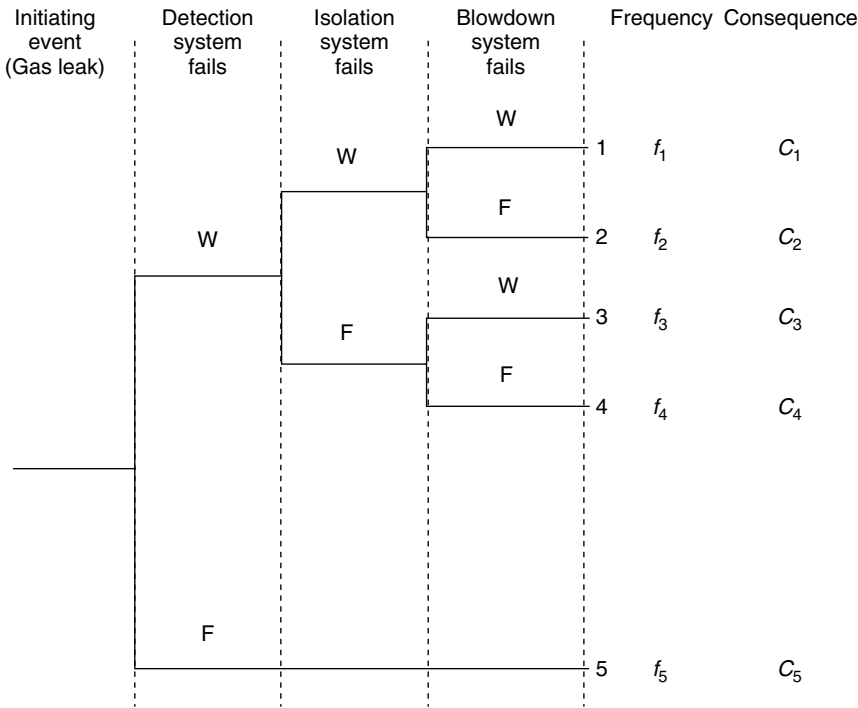


Figure 13 Event tree example

branch) and failure (bottom branch) for this system. In the next two columns, the functionality of the isolation and blowdown systems are considered. All combinations of success and failure of these systems are represented. Following the failure of the detection system, the isolation and blowdown systems are not activated, and so the branch that passes through the columns representing the functionality of these systems is a null branch, since the functionality of these systems is irrelevant in determining the final outcome.

Having constructed the event tree diagram, all possible responses to the hazard have been identified. Each response is represented by the end points of the branches on the far right of the diagram. For example, in Figure 13, the sequence of events labeled 1 at the top of the outcomes represents the situation in which all three systems have responded, as intended, to the hazard. The next two outcomes represent the correct functioning of the leak detection system, with only one of the isolation or blowdown systems responding. Sequence, outcome 4, has the detected leak being neither isolated or blowdown. The final outcome has the detection system failed, and so there is no response to the leak.

All outcome events have an associated consequence, C_i , indicated in a column at the end of the event tree diagram. The consequence depends on the nature of the hazard and can be expressed in terms of quantities such as: financial loss, injuries, fatalities, environmental damage, or the magnitude of a nuclear release.

Quantification of the event tree diagram produces the frequency of each of the outcome events, f_i . The data required to determine this is the frequency of the initiating event, and the probability of progressing along each of the branch points represents the failure or functionality of each of the systems included on the diagram. The failure of each system can be determined by any of the other methods discussed in this article, depending on the characteristics of the system concerned. Fault trees are, however, the most commonly used technique for this purpose. The event tree shows all possible responses to the initiating event and so, when quantified, the total of the frequencies for each of the outcome events will sum to the frequency of the initiating event.

The results from an event tree analysis can be represented in many forms. The total risk can be

determined from

$$Risk = \sum_{i=1}^n f_i C_i \quad (16)$$

where f_i and C_i are the frequency and consequence, respectively, of each of the n branch outcomes.

The results can also be represented graphically as a plot of consequence (severity) against frequency. It is common to use log scales on the axes of such a plot. Also, the most common form to express the results is as a cumulative frequency versus severity plot, where the cumulative frequency is the frequency by which the associated severity level is exceeded.

Summary of When to Use Each Method

With the choice of possible methods available to conduct an assessment for any given system, it has to be decided which one to select. The selection depends on the characteristics of the system and the performance parameter which is to be predicted. The list of possible methods is of those whose implicit assumptions are appropriate to model the system. From this list, the method selected would be the one that provides the most efficient analysis. Figure 14 shows the relationship between complexity and efficiency of the systems reliability techniques discussed. The most efficient techniques are placed on the left of the diagram. As the line moves from left to right, this indicates increasing complexity of the techniques associated with decreased efficiency. As such, the preferred method would be the method located as far to the left as possible of the diagram, which has the capability to model the system parameters of interest and whose assumptions are appropriate to model the system structure, operation, and maintenance.

If the requirement is to model the system unavailability or failure rate, then all methods can be used. The combinatorial methods on the left extreme of the diagram would be the preferred option. For these methods to be appropriate, the component failures must be independent of each other, and the component failure and repair rates must be constant (an implementation limitation of most of the commercial codes). Under these circumstances, fault tree analysis or reliability networks can be used. Fault tree analysis is the preferred choice between these two options, since the text boxes in the diagram enable the analysis to be documented. This is a very useful

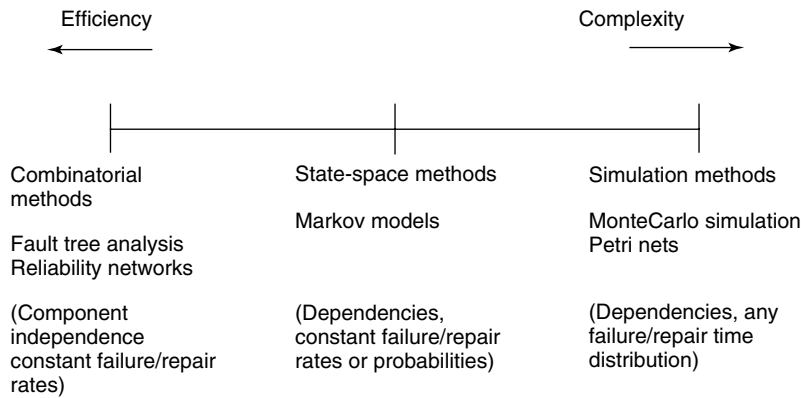


Figure 14 Characteristics of system reliability quantification methods

feature to explain the thinking behind the failure logic development to peers, managers, designers, and most importantly, regulators. When the assumption of independence cannot be made, then, if constant component failure and repair rates are appropriate, the Markov models can be used. In the event that the size of the system containing dependencies leads to a state space explosion, then either of the two simulation based approaches should be used.

If the system reliability is to be predicted, then the choices lie between the state space and simulation methods. The choice is again governed by the factors described in the previous paragraph.

In the situation in which the component failure or repair time distributions are not exponential, then the selection would be one of the simulation methods.

The above methods predict the probability that the system fails in terms of the likelihood of its component failures and the structure of the system. These predictions can then be incorporated into an event tree analysis to perform a risk analysis of the system.

References

- [1] The Royal Society (1992). *Risk Analysis Management and Perception*, Royal Society Publishing.
- [2] Andrews, J.D. & Moss, T.R. (2002). *Reliability and Risk Assessment*, Professional Engineering Publishing.
- [3] Henley, E.J. & Kumamoto, H. (1981). *Reliability Engineering and Risk Assessment*, Prentice Hall.
- [4] Billinton, R. & Allan, R. (1983). *Reliability Evaluation of Engineering Systems*, Pitman.
- [5] Watson, H.A. (1961). *Launch Control Safety Study, Section VII*, Bell Labs, Murray Hill, Vol. 1.
- [6] Vesely, W.E. (1970). A time-dependent methodology for fault tree evaluation, *Nuclear Engineering and Design* **13**, 337–360.
- [7] Andrews, J.D. (2001). The use of NOT logic in fault tree analysis, *Quality and Reliability Engineering International* **17**, 143–150.
- [8] Andrews, J.D. (2006). Fault tree analysis using binary decision diagrams, Tutorial Notes, *RAMS*, Newport Beach.
- [9] FaultTree+, Software supplied by Isograph Limited, Warrington.
- [10] Marseguerra, M. & Zio, E. (2002). *Basics of the Monte Carlo Method with Application to System Reliability*, LiLoLe Publishing Company.
- [11] Schneeweiss, W.G. (1999). *Petri Nets for Reliability Modelling*, LiLoLe Publishing Company.

JOHN D. ANDREWS

Threshold Models

What is a Threshold Model?

Dose–response models (*see Dose–Response Analysis*) link the dose of an agent, d , to the effect observed in the organism it is administered to. The shape of dose–response at low doses is of considerable interest because exposure of humans to carcinogens is usually orders of magnitude below the dose level used in bioassays. A *threshold model* focuses on the low-dose behavior of dose–response models and assumes that doses below a threshold, d^* , do not have an effect on response, i.e., doses below d^* lead to the same response as background. Specifically in cancer risk assessment, for example, response is often considered as lifetime tumor probability. Mathematically, a general threshold model for the effect probability $P(d, t)$ in observational period t is given by Edler and Kopp-Schneider [1]

$$P(d, t) = \begin{cases} p_0(t) & \text{if } d \leq d^* \\ p_0(t) + (1 - p_0(t))F(d, d^*, t) & \text{if } d > d^* \end{cases} \quad (1)$$

where $p_0(t)$ denotes the background response and $F(d, d^*, t)$ represents a continuous – usually monotonically increasing – function with values between 0 and 1 and $F(d^*, d^*, t) = 0$. Usually t is considered as fixed. Schematically, a threshold model is visualized in Figure 1.

Dependence of $P(d, t)$ on the observation period is often neglected. This may be justified in typical bioassays with e.g., a 2-year duration. However, variations in observation period will lead to alterations in dose–response relationship and obviously also in potential threshold doses. Equation (1) and Figure 1 show a strict threshold. Less stringent definitions of threshold models have been discussed, requiring that the derivative with respect to dose, $\partial P(d, t)/\partial d$, be zero at $d = 0$. This definition requires that dose–response starts horizontally in dose 0, leading to a curve with a shape similar to Figure 1. All polynomials of order >1 show this property. However, no clear-cut threshold dose can be defined whenever the derivative is positive for positive dose.

Why Were Threshold Models Introduced?

The motivation for introducing the threshold concept is as follows. If threshold doses existed and could be estimated, safe exposure levels could be derived, leading to tolerable small exposures. In this case, exposure below the threshold dose would be acceptable and no further endeavor is necessary to reduce exposure. On the other hand, efforts to reduce exposure in the environment to extremely small values are justified if dose–response models do not exhibit a threshold. Hence, the existence or nonexistence of thresholds has tremendous consequences for practical risk assessment. As a result, the existence of thresholds for cancer risk models has been a matter of fierce controversy.

The observation that organisms in a bioassay or epidemiologic study receiving a positive dose show no statistically significant difference in response compared to background may be taken as argument for the existence of an apparent threshold. However, the absence of a statistically significant effect depends on bioassay sample size and therefore, study power. Lack of evidence for a significant effect may not be interpreted as evidence for the lack of the effect. Apparent thresholds may also be observed depending on the choice of the scale when plotting dose–response curves: linear dose–response curves may show a thresholdlike behavior as depicted in Figure 1 if dose is plotted on a logarithmic scale [2].

Do Threshold Doses Really Exist?

Theoretical arguments have been collected to support the existence of thresholds. Possible threshold mechanisms include metabolic inactivation of the compound, DNA repair, cell cycle arrest, apoptosis of damaged cells, and control of the immune system [3]. The discussion about the existence of thresholds is closely related to the discussed distinction between genotoxic and nongenotoxic carcinogenic compounds. The existence of thresholds for nongenotoxic agents (e.g., tumor promoters) is more or less accepted in the scientific community, whereas genotoxic compounds have usually been associated with a nonthreshold behavior. In this context, it has recently been suggested that two types of threshold should be distinguished. Dose–response

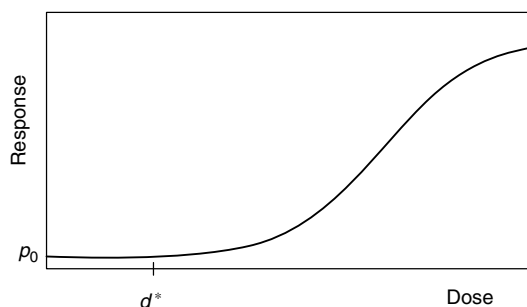


Figure 1 Schematic dose–response relationship with threshold dose d^* for fixed observational period t

of nongenotoxic carcinogens as well as some genotoxic agents acting as e.g., mitotic spindle poisons, may show a “perfect” threshold. Definition of a “practical” threshold is based on the idea that the chemical should cause no genotoxic effect at very low exposure levels owing to e.g., rapid degradation or other factors that limit target exposure. Using this definition of thresholds, it has been suggested that carcinogens can be classified in the following way [4, 5]: (a) nonthreshold genotoxic agents; (b) genotoxic compounds for which the existence of a threshold is possible, but is not yet sufficiently supported; (c) genotoxic carcinogens for which a “practical” threshold is supported by studies on mechanism and/or toxicokinetics; (d) genotoxic carcinogens for which a “perfect” threshold is associated with a no-observed effect level; and (e) nongenotoxic carcinogens for which a “perfect” threshold is associated with a no-observed effect level.

However, even if a threshold exists, determination of its actual value, d^* , may be hampered by practical problems. Estimation from study data involves the

issues of choice of doses, sample size, and study power. It should be noted in this context that the no-observed adverse effect level (NOAEL) has been taken as surrogate for threshold dose.

References

- [1] Edler, L. & Kopp-Schneider, A. (1998). Statistical models for low dose exposure, *Mutation Research* **405**, 227–236.
- [2] Edler, L., Portier, C.J. & Kopp-Schneider, A. (1994). Zur existenz von schwellenwerten: wissenschaftliche methode oder statistisches artefakt in der risikoabschätzung, *Zentralblatt für Arbeitsmedizin* **44**, 16–21.
- [3] Hengstler, J.G., Bogdanffy, M.S., Bolt, H.M. & Oesch, F. (2003). Challenging dogma: thresholds for genotoxic carcinogens? The case of vinyl acetate, *Annual Review of Pharmacology and Toxicology* **43**, 485–520.
- [4] Bolt, H.M., Foth, H., Hengstler, J.G. & Degen, G.H. (2004). Carcinogenicity categorization of chemicals – new aspects to be considered in a European perspective, *Toxicology Letters* **15**, 29–41.
- [5] Fukushima, S., Kinoshita, A., Puatanachokchai, R., Kushida, M., Wanibuchi, H. & Morimura, K. (2005). Hormesis and dose-response-mediated mechanisms in carcinogenesis: evidence for a threshold in carcinogenicity of non-genotoxic carcinogens, *Carcinogenesis* **26**, 1835–1845.

Related Articles

Environmental Carcinogenesis Risk

Gene–Environment Interaction

Low-Dose Extrapolation

ANNETTE KOPP-SCHNEIDER

Toxic Torts: Implications for Risk Management

This article introduces aspects of the law of torts that relate exposures to noxious agents (possibly due to either willful or negligent conduct), to their possible adverse effects. Tort law is a form of civil (not criminal) law that intervenes to redress wrongs from willful or negligible behavior and compensate (generally with money awards) those legally proved injured by those actions. In those cases, law and science intersect at the key issue of causation: scientific evidence and proof of causation must fit within concepts of legal causation. All of these use elements of scientific evidence and causal reasoning that are part of quantitative risk assessment methods. In particular, concepts such as increased risk, dose–response, statistical significance and inference, and probabilistic reasoning are essential. Examples from the United States, United Kingdom, and Australia illustrate differences among jurisdictions and modern views of the interactions between science and law.

Toxic tort law concerns past and, in some limited instances, feared future injuries from exposure to toxicants and carcinogens generally produced, emitted, or disposed in ways that (are alleged to) contravene certain social norms that constitute the current law of torts. (For brevity, this article does not address injunctive relief and intentional torts such as assault and battery.) In tort law, the scientific evidence concerning human health generally consists of epidemiological studies, although animal studies (such as those conducted with rodents – see **Cancer Risk Evaluation from Animal Studies**) and tests on lower organisms (e.g., bacteria) may also be considered. Establishing causation based on these heterogeneous data, and often competing etiological theories, requires abstractions and reasoning with probabilities, instead of relying on deterministic certainty. Cranor [1] provides the detail of some of the methods used to develop evidence *via* a “weight-of-evidence” system to show causation in regulatory law.

To deal with these complications, the law uses seemingly objective legal tests of causation, asking whether the defendant’s acts were a “substantial factor” producing injury, or whether, “but for” the defendant’s actions, the plaintiff would not have

suffered injury. The former compares the extent and importance of the defendant’s tortious activity in causing harm with other known and unknown causes, but it is too subjective to produce predictable results in most toxic exposure cases. The latter deterministically limits evaluation of causal factors to one – it seeks to answer whether “but for the act of the defendant, the plaintiff would not have been injured”. However, often such deterministic tests are too limiting in toxic tort and environmental law. For instance, in *Allen et al. v. U.S.* (588 F.Supp. 247 (1984), rev’d on other grounds, 816 F. 2d 1417 (10th Cir. 1988), cert. den’d, 484 US 1004 (1988)), several plaintiffs brought action against the United States under the Federal Torts Claim Act, attempting to recover for leukemia alleged to be the result of exposure to atomic bomb fallout after those weapons were tested in Nevada. The trial court discussed tests such as “natural and probable” consequences, “substantial factor”, and “more likely than not”, finally opting for the “substantial factor” test, rather than “but for”. The *Allen* court held that each plaintiff should meet at the “preponderance of the evidence” level the burden of proof. The probability of exposure to fallout resulting in doses in significant excess of “background” was predicated on the following basis:

1. injury consistent with those known to be caused by ionizing radiation; and
2. residence near the Nevada Test Site for at least some, if not all, of the years when the tests were conducted.

Reading tort law cases produces other terms of art used to assess what is known about the adverse effects of toxic exposures. For example, in the case *Smith v. Humboldt Dye Works, Inc.* (34 App. Div. 2d 1041 (1970)), the court held that 25 years of exposure to “known” carcinogens that were causally linked with bladder tumor were a “substantial evidence” of causality, even though the scientific evidence was in conflict and the statistical results unclear. Also in the United States, the Restatement Second of Torts (Restatement (II) Section 431, comment (a)) uses “legal cause”, rather than “proximate cause”, predicated on determining whether a tortfeasor’s act constitutes a “substantial factor” (*In Re Bendectin Litigation*, 857 F2d. 290 (6th Cir. 1988)), in bringing about a result, considering, *inter alia*, the “number of factors” in the causal structure, the conduct of the defendant in creating or setting

in motion forces leading to injury, those factors that go beyond the responsibility of the defendant, the time lag, and foreseeability between cause and effect [2].

Regarding the duty of care (the legal duty from the defendant to the plaintiff, which is an essential element to be proved in a tort law case) to warn consumer of potentially hazardous exposures (i.e., voluntarily taken vaccinations), in *Davis v. Wyeth* (399 F. 2d 121 (9th Cir. 1986)), the court held that a duty to warn (for the manufacture defendant) exists even though the probability of contracting poliomyelitis from being vaccinated was one in a million (but it apparently did happen to the plaintiff who contracted polio and thus sued the manufacturer).

In reading this article, key questions to keep in mind include the following: (a) what is the probability that the defendant's act caused harm to the plaintiff; (b) what is the probability that the plaintiff is among those harmed; and (c), given the answers to (a) and (b), what should be done to compensate the plaintiff properly?

Causation in Tort Law

A Snippet of History: How to Establish Legal Causation?

In the United Kingdom, cause has been assessed under different terms of art, from "real effective cause" to "direct" cause to the cause that was "natural and probable" to "proximate cause". These causal descriptions culminated with *Overseas Tankship (UK) Ltd v. Morts Dock and Engineering Co. Ltd. (The Wagon Mound)* ([1961] A. C. 388), in which "direct consequence" was disapproved, and was replaced by the "reasonable foreseeability" of harm in *Wagon Mound No. 2 (Overseas Tankship (UK) Ltd v. The Miller Steamship Co Pty* [1967] 1 A. C. 617). The dissent by Andrews J., in *Palsgraf v. Long Island RR* (162 N.E. 99 (N.Y. 1928)) defines "proximate cause" as follows: "What we do mean by the word 'proximate' is, that because of convenience, of public policy, of a rough sense of justice, the law arbitrarily declines to trace a series of events beyond a certain point. It is all a question of expediency". Moreover, "[t]he foresight of which the courts speaks assumes prevision".

A more recent case, *Lee v. Heyden (Pamela Lee B. v. Heyden*, 94 C.D.O.S. 4265 (1994)) summarizes causation as follows:

[t]he causation element is satisfied when the plaintiff establishes (1) that the defendant's breach of duty (...) was a substantial factor in bringing about the plaintiff's harm and (2) there is no rule of law relieving the defendant's of liability.

In England, Lord Reid, in the case *McGhee v. National Coal Board* ([1973] 1 WLR 1, at 5) stated as follows:

... it has often been said that the legal concept of causation is not based on logic or philosophy. It is based on the practical way in which the ordinary man's mind works in the everyday affairs of life.

In Australia, in *Bennett v Minister of Community Welfare* ((1992) 107 ALR 617), the Australian High Court followed the majority in the earlier case *March v. E & MH Stramare Pty Ltd* ((1991) 171 CLR 506), holding as follows:

In ... negligence, causation is essentially a question of fact, to be resolved as a matter of common sense. In resolving that question, the "but for" test, applied as a negative criterion of causation, has an important role to play but it is not a comprehensive and exclusive test of causation; value judgments and policy considerations necessarily intrude.

In light of these admonitions, consider an environmental or occupational case in which the empirical evidence is epidemiological. The plaintiff demonstrates "general causation" by showing a statistically significant increase in the relative risk. In *Usury v. Turner Elkhorn Mining Co.*, 428 U.S. 1 (1976), the US Supreme Court held that uncertain causation does not bar *statutory* recovery. And, a negative finding in a radiological examination does not suffice to deny a worker's compensation claim. *Usury* holds that Congress can develop *probabilistic* health protection standards and provide just compensation in the factual instances of uncertain etiological factors associated with dreaded, fatal, and causally complex diseases.

More recently, in *James Mitchell et al. v. Jose L. Gonzales et al.* (54 Cal. 3d 1041; 819 P.2d 872; 199; 1 Cal. Rptr. 2d 913), the Supreme Court of California disapproved the "'but for' test of cause in fact" ruling in favor of the "substantial factor" test, further illustrating the pathology of causation

in legal reasoning. The issue before the Supreme Court of California concerned the instructions to the fact finder about *factual causation*. According to this court, *proximate cause* contains *cause in fact* or actual cause; this is the scientific aspect of legal causation. Yet, using the phrase *proximate cause* to satisfy both factual and legal causation has and continues to generate much confusion. Thus, in *Mitchell*, the Supreme Court of California cited Prosser's (1950) statement "that 'proximate cause' . . . is a complex term of highly uncertain meaning under which other rules, doctrines and reasons lie buried . . .". For jury instruction, according to Charrow [3] "the term *proximate cause* was misunderstood by 23% of the subjects . . . They interpreted it as 'approximate cause,' 'estimated cause,' or some fabrication". The "substantial factor" test applies when there are concurrent (not necessarily jointly necessary and sufficient) causal events such that applying the "but for" test would result in both defendants escaping justice. Prosser [4] declared the *substantial factor* test "sufficiently intelligible to any layman to furnish an adequate guide to the jury, and it is neither possible nor desirable to reduce it to lower terms". However, in the *Mitchell* case, the court never stated what it required, quantitatively, for a *factor* to be "substantial".

Legal Aspects of Causation: Foreseeability, Duty of Care, Proximity, and Relationship. The normal is typically the foreseeable. As the Supreme Court of California put it, "[o]n a clear day, some courts can see forever, at least when they are looking for negligence". In an earlier case, it had held that "[e]xperience has shown that . . . foresight alone [does not] provides a socially and judicially acceptable limit on recovery of damages for injury". Foreseeability limits tort liability by limiting the causal inquiry about the factual chain of events alleged by the plaintiff. In some cases, legal duty is a way out of the difficulties of setting legal bounds on causally complex facts. For example, in Australia, in *Chapman v. Hearse* ((1962), 106 C. L. R. 112), the Australian High Court held that "[u]nless all events which happened were reasonably foreseeable", Chapman was not guilty of a breach of his legal duty. The phrase "reasonably foreseeable" was not a direct test of causation; rather, "it marks the limits beyond which a tortfeasor will not be held responsible. . ." as the theoretical proposition that foreseeability limits legal duty.

If there is a duty, it must be related to physical cause so that the legal limit can be set with confidence. There is a one-to-one correspondence between foreseeability and this limit on the factual, but *ex post*, reconstructed view of factual causation. In the United Kingdom, *Caparo Plc v. Dickman* ([1990] 2 A. C. 605 (H. L.)), at 617–618) developed a legal test articulated in terms of the following:

Foreseeability of damage (which itself has a substantial history in case law),

Legal proximity or neighborhood (which also has a rich history of interpretation), and the

Fair, just and reasonable imposition of a legal duty.

The plaintiff does not need to show with accuracy *all* of the elements required by the "reasonably foreseeable" test. It is sufficient to show "injury to a class of persons of which the plaintiff was one" and that such injury "might reasonably be foreseen as a consequence" of the defendant's tortious act.

Justification of risk management policy choices requires developing the formal nexus between factual causation and the limits of "foreseeability" further. Formal probabilistic analysis may help in defining what should be considered "foreseeable" on theoretical and empirical grounds.

Balancing Adversarial Evidence at Trial: "The Balance of the Probabilities" and Similar Tests

In tort cases, the *reasonable* mind is given the task of deciding causation, often according to innate common sense, even in cases involving scientific evidence of the greatest complexity. In traditional tort law cases, a trier of fact who is "51%" convinced of causation allows recovery, and one who is "49%" convinced does not. The question thus arises: how can a combination of subjective notions of uncertainty and common sense causation meet the burden of proof?

The "More Likely than Not" Test

Plaintiffs in the United States, the United Kingdom, Canada, and Australia must prove their facts using the "more probable than not" standard, thus maximizing the expected number of correct civil decisions. The

prevalent American tort law balancing standard is alternatively stated as “the balance of probabilities”, “preponderance of the evidence”, “more likely than not”, and “51%”. These phrases are supposed to impose the same burden of proof on the parties; once one party’s evidence is greater than that of the other, that party prevails according to the all or nothing rule of traditional tort liability. The essence of the test (*In Re Agent Orange Product Liability Litigation*, 597 F.Supp. 740, at 835) is that even “if there near certainty as to general causation, if there were significant uncertainty as to individual causation, traditional tort principles would dictate that causation be determined on a case-by-case basis. . .”. Some courts have distinguished between degrees of scientific certainty and “legal sufficiency”, and some have held that a lesser standard should apply to scientific facts. Other courts shift the burden to the defendant and keep the “more likely than not” test. On occasions, a higher standard of proof than the “more likely than not” applies in civil litigation. The standard of proof may be altered for events characterized by severe consequences – such as cancer – even if they have low probability. The idea that the standard of proof should be adjusted so that “high probability” is commensurate with “what is at stake” (United Kingdom, *Khawaja v. Secretary of State for the Home Department* [1984] AC 74) has sometimes been rejected because of the difficulties confronting a plaintiff in meeting those standards, and a perception that any balancing more stringent than the “more likely than not” can be, *a priori*, unfair to the toxic tort plaintiff (Australia High Court, *Neat Holdings Pty Ltd v. Karajan Holdings Pty Ltd*. ([1992] 67 ALJR 170 at 171)).

The interpretation of the “more likely than not” test as a probability number has led to various strains of this legal standard. The “strong” version would not accept statistical measures of association between events without additional evidence of a “. . . direct and actual knowledge of the causal relationship between the defendant’s tortious conduct and the plaintiff’s injury” [5]. The *Agent Orange* (597 F.Supp. 740, at 835) court has understood this to mean as follows:

... ‘proof that can provide direct and actual knowledge of the causal relationship between the defendant’s tortious conduct and the plaintiff’s injury’ is needed.

The “weak” version, on the other hand, suggests that statistical evidence alone can suffice to make a decision. Once the probabilistic balancing has taken place, the fact finder has a simple decision to make: either the balance of evidence has swung against the defendant, or it has not. Judge Weinstein echoed this view in stating that “particularistic evidence . . . is no less probabilistic than . . . statistical evidence.” His court favored the use of the “strong” version of the standard for individual plaintiffs, but appeared to support the “weak” version for class actions lawsuits (597 F.Supp. 740). For example, he noted that if 1000 cancers are expected and 1100 are found (and the result is statistically significant), then individual plaintiffs would never recover because $100/1100 = 9\%$ is below the conventional 51%. To avoid this patently unfair result for class action cases, Judge Weinstein proposed that results falling outside two standard deviations from the expected incidence of the disease, computed to be 31.6 excess cancers relative to the 1000 expected, should suffice to allow the more likely than not test to be met. The 100 excess cancers meet this new version of the more likely than not test. This test includes a reasonable measure of variability of the data, namely, the standard deviation, assuming – a strong assumption – that all plaintiffs are homogeneous with respect to exposure response.

An American tort law test that balances probability of harm against cost of mitigating that harm was enunciated by Judge Learned Hand, in *United States v. Carroll Towing Co.* (159 F.2d 169 (2d Cir. 1947)). This utilitarian formulation focuses on economic efficiency, assuming that the societal objective is to increase expenditures on safety measures until their marginal social cost equals the marginal social benefit. As Judge Learned Hand stated,

... the owner’s duty ... to provide against resulting injuries is a function of three variables: (1) the probability that she (the ship that caused initiated the sequence leading to damage) will break away; (2) the gravity of the resulting injury, if she does; (3) the burden of adequate precautions.

He then developed the formula for balancing these three variables as follows:

Possibly it serves to bring this notion into relief to state in algebraic terms: if the probability be called P, the injury, L; and the burden, B; liability depends upon whether B is less than L multiplied by P ...

In this formula, the expected (or average) value of the injury or economic loss controls. The cost can be either the “marginal” cost or the “average” cost.

Similarly, English law allows the balancing the magnitude of consequences against cost of reducing risk (*Edwards v. National Coal Board* ([1949] 1 KB 704)), while Australia defines reasonableness as the balancing of “magnitude of the risk and the degree of probability of occurrence along with the expense and difficulty and inconvenience of taking alleviating action and other conflicting responsibilities the defendant may have” (*Wyong Shire v. Shirt* (1980) 146 CLR 40).

What is, Legally, Good Science Regarding a Finding of Scientific Causation

Pretrial Admissibility of Scientific Evidence: Frye v. Daubert—Joiner—Kumho

The famous *Frye* standard for admissibility of scientific evidence based on what the court construes as being “generally accepted” in the relevant scientific community clashes with rapid scientific advances and the lag between new science and general acceptance. The *Frye* tests guard against false positives (especially in criminal cases, where proof must be “beyond a reasonable doubt” rather than merely by “preponderance of evidence”), but not against false negatives. Today, the increasing pace of information development and their contingency on newer information often makes *Frye* inappropriate for torts, although it persists in several states. The net effect of the rule is to exclude scientific evidence that could, nevertheless, be appropriate for the fact finders to consider. A version used in California (*Kelly–Frye*) seemingly reduces scientific opinions and findings by experts to a counting of the number of scientific peers proclaiming “yays” and “nays” and then deciding what correct causation is. However, Californian courts appear to limit *Kelly–Frye* to novel scientific methods and devices, such as DNA testing and voiceprint.

In Australia, in *R. v. Gilmore* [1977] 2 NSWLR 935 at 938–939), the following view was advanced (following a US case):

Absolute certainty . . . or unanimity of scientific opinion is not required for admissibility. ‘Every useful new development must have its first day in court . . .’

... Unless an exaggerated popular opinion . . . makes its use prejudicial or likely to mislead the jury, it is better to admit relevant scientific evidence . . . and allow its weight to be attacked by cross-examination and refutation

This “exaggerated popular opinion” is one of the two aspects of expert testimony that can lead to unfair results. The second aspect concerns ill-constructed expert testimony about causation. Expert opinion is generally given to assist the trier of fact (the jury, most of the time, or the judge in some situations), given all the other relevant evidences and various procedural constraints. This function argues for a formal and coherent balancing of the advantages and disadvantages of admission for both parties.

In the US federal system *Frye*’s hegemony ended in 1993 with the US Supreme Court case of *Daubert v. Merrell Dow Pharmaceuticals Inc* (509 U.S. 579 (1993)). The factors enumerated by the Supreme Court in *Daubert* include the following:

1. reliability (the empirical testing of scientific hypotheses or opinions that the expert intends to introduce as evidence);
2. scientific validity of the methods and reasoning and scientific peer review.

Daubert requires four factors to establish the reliability of *expert knowledge*: (a) a “theory or techniques . . . can be (and has been) tested”; (b) its “peer review and publication”; (c) its “known or potential rate of error” and “standards controlling the technique’s operation”; and (d) its “general acceptance in the relevant scientific community”. The context of the case determines the appropriate number of factors, reflecting the flexibility of the federal rules of evidence (FRE). Notably, *Daubert* places trial judges in the role of gatekeepers of scientific evidence:

The objective of this requirement is to ensure the reliability and relevancy of expert testimony. It is to make certain that an expert, whether basing testimony upon professional studies or personal experience, employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field.

The reach of *Daubert* was extended from scientific evidence to all expert evidence in *Kumho Tire Co., Ltd. et al. v. Carmichael, etc., et al.* (526 U.S. 137 (1999)). *Kumho* holds that *Daubert* applies to all contexts to which FRE 702 applies, regardless of whether scientific, technical, or other specialized knowledge

is introduced. The US Supreme Court confirmed in *Kumho* “judicial broad latitude” allowing gatekeeping by trial judges. Under *Kumho*, trial courts now determine the admissibility of expert testimony by analysis of the “reasonableness” of inference, and the expert’s methods for analysis and *whether those reasons lead to the conclusion derived*.

Standard of Review on Appeal: *General Electric Co. v. Joiner* (522 U.S. 136 (1997)). In *Joiner*, the US Supreme Court reconfirmed the trial judges’ considerable latitude in admitting scientific evidence of causation. The plaintiff’s experts advanced a theory of lung cancer causation based on cancer-promoting activity of PCBs in mice at high doses. One expert testified that it was “more likely than not that Mr. Joiner’s lung cancer was causally linked to cigarette smoking and PCB exposure”, and another testified that “lung cancer was caused or contributed to in a significant degree by the material with which Joiner had worked”. The plaintiff’s experts, however, would not testify that there was a “link between increases in lung cancer and PCB exposure among the workers they examined” on the basis of their own studies. They had used a “weight of the evidence” method to evaluate the causal relationship between Joiner’s exposure and his lung cancer. The evidence they considered included four epidemiological studies and animal bioassay results, as well as the plaintiff’s medical records. The PCB exposures in those studies were confounded by other agents, or the studies did not support a causal association, or PCBs were not mentioned in the exposures. After scrutinizing *some* of the epidemiological studies *individually*, as opposed to *aggregating* the evidence to get a better sense of overall results, the trial court (the federal district court) excluded that evidence. The court of appeal upheld this approach, finding that the studies did not provide sufficient evidence of a causal link to overturn the trial court findings of inadmissibility of the scientific opinions. The Supreme Court agreed with the trial court’s logic; it refused to accept expert *assertions* as a sufficient basis for admitting evidence.

Quantitative Risk Assessment, QRA, in Legal Causation

There is no satisfactory alternative to methods of QRA for developing sound causal arguments in toxic

tort law. Legal tests of causation require substantive quantitative and qualitative analyses to bound potential causation by accounting for likelihoods and expert opinion (perhaps encoded as prior probabilities or distributions, or as conditional independence relations in *influence diagrams*), to yield sound formulations of causation that combine both. Although qualitative aspects can be met by expert opinions, as *Daubert’s* “intellectual rigor” standard enjoins, methods of decision theory and QRA are also essential to understand and quantify potential outcomes of actions and to assess actions according to specific criteria of optimality for risk taking. QRA enables a correct, appropriately caveated understanding (expressly conditioned on available information) of the strength of a causal relation, and thus support and enhance legal causal analysis, regardless of the type of legal test used by a court.

Conclusions

Scientific causation can be crucial in determining legal causation using standards (such as *but for* or *more likely than not*) by which plaintiffs must legally demonstrate that the actions of the defendants (probably) caused their injuries. Courts in multiple countries use scientific causation and evidence in a much broader context by examining *reasonable* conduct, existence of *legal duty* and breach of such duty, *foreseeability of injury*, the resulting *damages* to the plaintiffs, and other factors to reach a verdict. Thus, potential causation, in the sense of hazard, adverse effect, and risk are necessary but not sufficient aspects of tort law cases. To prevail, plaintiffs must meet all of these aspects, and not just establish a strong causal set of scientific facts.

Toxic torts and tort law in general have been subjected to many attempts to reform their legal structure, in part because of the magnitude of punitive damages (awarded if the breach of duty is deemed to be beyond negligence on the part of the defendants). Although the magnitude of awards by juries (particularly in the United States) falls outside the scope of health risk assessment, it is relevant to risk management and should be considered in prudent risk management. The risk manager may use probabilistic simulations to include ranges of options and their overall consequences using, for example, decision analysis to account for varying amounts of

damages, contingent on the outcome from potential law suits (and include the cost of settlements or having to go to trial). Game-theoretic models of pre-trial arbitration and settlement [6] have also been extensively developed, and QRA methods and models can provide crucial information to those approaches by helping the parties potentially to agree on a settlement based on the quantitative probabilistic relation between defendant (and, perhaps, plaintiff actions) and resulting harm to the plaintiff, rather than proceeding to a trial that can be costly as well as result in an unpredictable outcome for the parties.

References

- [1] Cranor, C.F. (1996). Judicial boundary-drawing and the need for context-sensitive science in toxic torts after *Daubert v. Merrell-Dow Pharmaceutical*, *Virginia Environmental Law Journal* **16**, 1–77.
- [2] Sanders, J. (1998). *Bendectin on Trial: A Study of Mass Tort Litigation*, University of Michigan Press, Ann Harbor.
- [3] Charrow, R.P. (1979). Making legal language understandable: a psycholinguistic study of jury instructions, *Columbia Law Review* **79**, 1306, 1353.
- [4] Prosser, W.L. (1950). Proximate cause in California, *California Law Review* **38**, 369, 375.
- [5] Rosenberg, D. (1984). The causal connection in mass exposure cases: a ‘public law’ vision of the tort system, *Harvard Law Review* **97**, 857 at p.870.
- [6] Daughety, A.F. & Reinganum, J.F. (1994). Settlement negotiations with two-sided asymmetric information: model duality, information distribution, and efficiency, *International Review of Law and Economics* **14**(3), 283–298.

Related Articles

Causality/Causation

Environmental Hazard

Environmental Health Risk

Expert Judgment

Hazardous Waste Site(s)

Occupational Risks

Subjective Probability

Toxicity (Adverse Events)

What are Hazardous Materials?

PAOLO F. RICCI

Toxicity (Adverse Events)

An adverse event (AE), or toxicity, in a clinical trial setting (*see* **Randomized Controlled Trials**), is any unfavorable change in structure (signs), function (symptoms), or chemistry (laboratory findings) in a subject or patient that is temporally associated with participation in the trial, whether or not this change is related to the trial intervention [1]. An AE is treatment emergent if it is not present prior to the start of the clinical trial, is not associated with a prior medical condition, and develops or worsens (if present at the start of the trial) during the trial [1]. Treatment-emergent AEs are classified on the basis of their relatedness to the therapeutic intervention: unrelated (clearly not related); unlikely related (doubtfully related); possibly related (may be related); probably related (likely related); or definitely related (clearly related) [2]. We note that, in randomized controlled trials, only increases in AEs in the intervention group compared with the control group are attributable to the intervention. An AE that is unrelated to the therapeutic intervention could be attributable to a concomitant medication (i.e., a medication prescribed for a coexisting disease).

A serious adverse event (SAE) is an AE that is at least possibly related to the intervention and that results in death; life-threatening experience; inpatient hospitalization or prolongation of existing hospitalization; persistent and significant disability or incapacity; jeopardy to patient and requires intervention; or a congenital anomaly or birth defect. Other important medical events may be considered SAEs if the patient or subject is jeopardized and requires medical or surgical intervention to prevent one of the outcomes of an SAE [2].

An unexpected adverse event is as any AE that is unanticipated in its nature or severity that may be related to the intervention [2].

In general, SAEs and unexpected AEs are reported to local Institutional Review Boards (IRBs), to study sponsors, and to appropriate regulatory authorities within 24 h [2].

We classify AEs using various standardized coding dictionaries. These dictionaries vary by regulatory group, study phase, disease, and type of therapy. For instance, the United States Federal Drug Administration (US FDA) used the Coding Symbols for Thesaurus of an Adverse Reaction Term (COSTART)

[3] until 1997; the FDA then instituted the use of the Medical Dictionary for Regulatory Activities (MedDRA [4]) developed by the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) [5]. The United States National Cancer Institute (US NCI) and the European Organization of Research and Treatment of Cancers (EORTC) use, for cancer studies, the Common Toxicity Criteria – version 2 (CTC v2.0 [6]) and the Common Terminology Criteria for Adverse Events – version 3 (CTCAE v3.0 [7]). Some research groups, for instance, the Radiation Therapy Oncology Group (RTOG) use the CTC v2.0 for acute (within 90 days from the start of therapy) and late (beyond 90 days from the start of therapy) radiation toxicities; others use the late effects on normal tissues (LENTs) scoring system and the subjective, objective, management and analytic (SOMA) scales for late radiation toxicities [8–10]. Studies run by the Eastern Cooperative Oncology Group (ECOG) use their own dictionary [11].

Most coding dictionaries classify AEs into five severity grades. For example, the grade definitions used in the CTC v2.0 [2] are as follows:

- 1 = mild adverse event
- 2 = moderate adverse event
- 3 = severe and undesirable adverse event
- 4 = life-threatening or disabling adverse event
- 5 = fatal adverse event.

Several coding dictionaries classify AEs in broad categories based on anatomy and/or pathophysiology (changes in body functionality caused by disease), and then classify them by a specific term within each category. In the CTCAE v3.0, for example, “autoimmune reaction” is a specific AE within the category of “allergy/immunology” [7].

In clinical trials, we perform routine laboratory tests before study enrollment or randomization (in controlled trials) to screen patients for eligibility and during the trials in order to ensure patient **Safety**. We consider a laboratory test result that is outside the normal range (as defined for a specific institution or group) as an abnormality. Such abnormal findings are unbiased indicators of changes in the chemistry of the body and are considered AEs.

Adverse experiences and serious adverse experiences are monitored throughout a clinical trial and are summarized to provide assessments of the **Safety** of the treatment under study.

References

- [1] Northington, B. (1996). A review of issues in the collection and reporting of adverse events, *Biopharmaceutical Report of the American Statistical Association* **4**, 1–5.
- [2] Cancer Therapy Evaluation Program (CTEP) (2005). *CTEP, NCI Guidelines: Adverse Event Reporting Requirements*, at http://ctep.cancer.gov/reporting/new-adverse_2006.pdf.
- [3] Center for Drug Evaluation and Research (CDER) (1995). *Coding Symbols for Thesaurus of Adverse Reaction Term (COSTART)*, 5th Edition, United States Food and Drug Administration, Rockville.
- [4] *Medical Dictionary for Regulatory Activities (MedDRA) [online]*, at <http://www.meddrasso.com/MSSOWeb/index.htm>.
- [5] Brown, E.-G., Wood, L. & Wood, S. (1999). The Medical Dictionary for Regulatory Activities (MedDRA), *Drug Safety* **2**, 109–117.
- [6] Cancer Center Evaluation Program (CTEP) (1999). *Common Toxicity Criteria v2.0 (CTC v2.0)*, at http://ctep.cancer.gov/forms/CTCv20_4-30-992.pdf.
- [7] Cancer Center Evaluation Program (CTEP) (2006). *Common Terminology Criteria for Adverse Events v3.0 (CTCAE)*, at <http://ctep.cancer.gov/forms/CTCAEv3.pdf>.
- [8] LENT SOMA tables (1995). *Radiotherapy and Oncology* **35**, 17–60.
- [9] Rubin, P., Constine, L.S., Fajardo, L.F., Phillips, T.L. & Wasserman, T.H. (1995). Overview of late effects normal tissues (LENT) scoring system, *Radiotherapy and Oncology* **35**, 9–10.
- [10] Pavy, J.-J., Denekamp, J., Letschert, J., Littbrand, B., Mornex, F., Bernier, J., Gonzales-Gonzales, D., Horiot, J.-C., Bolla, M. & Bartelink, H. (1995). Late effects toxicity scoring: the SOMA scale, *Radiotherapy and Oncology* **35**, 11–15.
- [11] Oken, M.M., Creech, R.H., Tormey, D.C., Horton, J., Davis, T.E., McFadden, E.T. & Carbone, P.P. (1982). Toxicity and response criteria of the Eastern Cooperative Oncology Group, *American Journal of Clinical Oncology* **5**, 649–655.

Related Articles

Safety

BENJAMIN LEVINSON AND JUDITH D.
GOLDBERG

Truncation

Types of Truncation

Truncation is one of the two major features that distinguish time-to-event data or failure-time data from other data and often occurs in risk analysis or other types of data [1]. The other feature is censoring [1, 2], which is discussed in a different article (see **Censoring**). Truncation represents or is usually defined as a condition in the study that produces the time-to-event data, and censoring usually refers to a type of incompleteness of observed data. For truncated data, only subjects who experience the events defining the condition and the times to the events of interest from those subjects are observed in the study. Subjects have to satisfy some conditions (e.g., exposure to a disease) or experience an intermediate event prior to the event of interest to be included in study. To give an example, suppose that one is interested in a follow-up study of subjects who were exposed to certain radiation to estimate their survival times after the exposure among other objectives. For this, one common design is to find and follow up on some subjects who were known to be exposed to the radiation. It is clear that a person who was exposed but died before the study was started would not be included in the study. For each subject in the study, his or her survival time T will be greater than the time Y , which is the period from the occurrence of the exposure to the radiation to the enrollment in the study. As defined below, this means that T is left truncated by Y .

Several types of truncations exist in practice: left truncation, right truncation, and general truncation. Let T denote the time to an event of interest, and let Y be the time to the truncation event. In left truncation, only subjects with $T \geq Y$ are observed or included in the study. A common example of left-truncated data is that subjects enter a study at a random time rather than at the origin of the event of interest, and are followed up from their delayed entry time to the occurrence of the event or to the time of being right censored. The reason behind this could be that subjects are simply not available for the study or study investigators. Another example of left truncation could be in a study of estimating cancer survival times based on a particular population. Some patients with cancer may not move

into the population and thus may not be included in the study until some time after the cancer is diagnosed. Also, subjects may have to experience a different intermediate event to be recognized as being at risk of experiencing the event of interest. In this case, the time to the occurrence of the intermediate event serves as the left-truncation time.

Another type of truncation that often occurs in risk and other studies is right truncation. Let T and Y be defined as before. In right truncation, only subjects with $T \leq Y$ are observed or included in the study. Right truncation occurs when only subjects who have experienced the event of interest are included in the study sample, and any subject who has not yet experienced the event is not observed. A common reason for this is the limitation of available information about the event or the study design. An example of right-truncated data arises from a mortality study based on death records. Of course, like censoring, a study could involve both left and right truncation [3].

Although there are many similarities between censoring and truncation, there are differences. One difference is that truncation often occurs with censoring but much less frequently. In consequence, there exists much less literature for the former than for the latter. Another difference between the two is that censored observations provide partial information about the response variable of interest, T , while the subjects who suffer truncation (or are truncated out) provide no information. To see this, consider again, the radiation example discussed above and suppose that the goal is to follow up and study a particular subpopulation who suffered certain radiation and for which the population size is unknown. Also suppose that the study started a while after the exposure occurred and there was no record of the subjects who had died from the exposure before the study commenced. In this case, as discussed above, the survival time of each subject in the study is left truncated and, in particular, it is unknown how many people were left out of the study owing to the left truncation caused by death. In contrast, if the study started right after the exposure, then one would not have to face the truncation and only censoring may exist. Specifically, the sample or subpopulation size is known.

The third type of truncation is general truncation, which occurs when, for example, one measures concentration of exposure to contamination (e.g.,

contaminated drinking water or radon exposure in homes [3, 4]). This is limited by the instrument's precision. When characterizing risk, risk factors are assumed to have some distribution (e.g., log normal distribution). Truncation makes the resulting distribution of the risk, which is assumed to be either a multiplicative or a general risk function, very complex. We discuss this issue in another articles (see **Risk Characterization and Uncertainty and Variability Characterization and Measures in Risk Assessment**). Here we focus on truncation in the context of time-to-event data.

Analysis of Truncated Data

As mentioned above, truncation often arises with censoring. Thus, most existing statistical approaches for truncated data were developed for both censored and truncated data. Let T and Y be defined as above, and assume that one observes right-censored and left-truncated data during the time to an event of interest. In this situation, the observed data have the form $\{X_i = \min(T_i, C_i), \delta_i = I(X_i = T_i), Y_i; i = 1, \dots, n\}$, where T_i and Y_i denote the same variables as defined above, but with subject i ; C_i denotes the right-censoring time associated with subject i , and δ_i is the censor indicator. Assume that both Y_i and C_i are independent of T_i as usual and all variables are continuous. Then the likelihood function can be written as

$$L = \prod_{i=1}^n \frac{f^{\delta_i}(x_i) [1 - F(x_i)]^{1-\delta_i}}{1 - F(y_i)} \quad (1)$$

where f and F denote the probability density function (pdf) and cumulative distribution function (cdf) of the T_i 's, respectively. The inference about F or any unknown parameters involved in F can be made on the basis of the maximum-likelihood approach.

It is interesting to note that if there is no truncation, that is, all $Y_i = 0$, the likelihood function L becomes

$$L = \prod_{i=1}^n f^{\delta_i}(x_i) [1 - F(x_i)]^{1-\delta_i} \quad (2)$$

On the other hand, without censoring, that is, all $C_i = \infty$, we have

$$L = \prod_{i=1}^n \frac{f(x_i)}{1 - F(y_i)} \quad (3)$$

since in this case, all $\delta_i = 1$. It is easy to see that, in terms of nonparametric estimation of F , the maximization of L in equation (2) gives an estimate of F over the whole support region, while the maximization of L in equation (3) only yields an estimate of F over $t \geq \max(y_i; i = 1, \dots, n)$.

In the case of right truncation or more general truncation with more general censoring, one can write the corresponding likelihood function similarly; this approach has been used by many authors. Among others, one of the early papers discussing truncation was given by Turnbull [3], who proposed a self-consistency algorithm for the nonparametric maximum-likelihood estimate of a distribution based on interval-censored [4] and truncated data. The algorithm applies to both right- and left-truncated data. More recently, Pan and Chappell [5] and Sun [6], among others, considered the one-sample estimation problem when left truncation and general truncation are involved. Lim *et al.* [7] also discussed the one-sample estimation problem with general truncation and the existence of a change point. For regression analysis of censored and truncated data, one can read Alioum and Commenges [8] and Pan and Chappell [9], who considered the problem by using the proportional hazards model [2].

Acknowledgment

This research was supported in part by the Grant CA21 765 from the National Institute of Health and by the American Lebanese Syrian Associated Charities (ALSAC). The authors thank Dr. Angela McArthur for the critical scientific editing. The authors are also grateful to the Editor and a referee for providing many helpful suggestions, which significantly improved the manuscript.

References

- [1] Klein, J.P. & Moeschberger, M.L. (2003). *Survival Analysis*, Springer-Verlag, New York.
- [2] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Edition, John Wiley & Sons, New York.
- [3] Turnbull, B.W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data, *Journal of the Royal Statistical Society, Series C* **38**, 290–295.
- [4] Rai, S.N., Bartlett, S., Krewski, D. & Peterson, J. (2002). Probabilistic risk assessment and its applications to drinking water guidelines, *Human and Ecological Risk Assessment* **8**, 493–509.

- [5] Pan, W. & Chappell, R. (1998). Estimating survival curves with left-truncated and interval-censored data under monotone hazards, *Biometrics* **54**, 1053–1060.
- [6] Sun, J. (1997). Self-consistency estimation of distributions based on truncated and doubly censored data with applications to AIDS cohort studies, *Lifetime Data Analysis* **3**, 305–313.
- [7] Lim, H.J., Sun, J. & Matthews, D.E. (2002). Maximum likelihood estimation of a survival function with a change-point for truncated and interval-censored data, *Statistics in Medicine* **21**, 743–752.
- [8] Alioum, A. & Commenges, D. (1996). A proportional hazards model for arbitrarily censored and truncated data, *Biometrics* **52**, 512–524.
- [9] Pan, W. & Chappell, R. (2002). Estimation in the Cox proportional hazards model with left truncated and interval censored data, *Biometrics* **58**, 64–70.

Further Reading

- Krewski, D., Rai, S.N., Bartlett, S., Zielinski, J. & Hopke, P.K. (1999). Characterization of uncertainty and variability characterization in residential radon cancer risks, *The Annals of the New York Academy of Sciences* **895**, 245–272.
- Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer, New York.

Related Articles

Hazard and Hazard Ratio

SHESH N. RAI AND JIANGUO SUN

Ultrahigh Reliability

Ultrareliability is a broad approach to achieving high levels of system reliability by taking a system of systems perspective (*see Multistate Systems*). It involves analyzing each individual system component, interactions between the components, and the emergent effects when multiple components are combined. In the end, the objective is to increase reliability by a full order of magnitude compared to the traditional reliability process. It is best viewed as a process, with the benefits accruing continuously, as more of these interactions are addressed.

Table 1 lists the typical direct and indirect benefits of ultrareliability.

Types of Failures

In order to achieve these benefits, the scope of the analysis must include the physical failures on which reliability analysis often focuses, as well as behavioral and procedural failures. Behavioral failures are those caused by a human behavior that is on the root cause path of the failure. It could include failures caused by outright human negligent behavior, where behavior is the root cause, or it could be a training or communication failure that leads to an incorrect human behavior. Procedural failures are those that emerge when a process or procedure is inappropriate for a given situation, even when correctly applied. This can occur when system context changes without corresponding updates to the process or if the process was not developed correctly in the first place.

Analysis of Failures

As failures are identified, they are analyzed to understand the root cause, the path from the root cause to the failure, and the effects of the failure on the system. In general, there are six aspects of each failure that should be modeled: the root cause, the root cause path, contextual factors that increase or decrease the likelihood of failure, the overall probability of failure in each use context, the severity of the failure in each context, and possible methods to control or prevent the failure.

Stages of Reliability Evolution

As organizations begin to apply the concepts of ultrareliability to a system, they tend to evolve through a series of five stages. Pursuing this gradual approach tends to be more successful than attempting to jump directly to ultrareliability because many aspects of an ultrareliable system, such as management culture, take time to develop. The stages are as follows:

1. Reactive: failures are addressed and repaired as they occur.
2. Preventive: maintenance is conducted on a fixed schedule based on historical performance to prevent most failures.
3. Predictive: maintenance is conducted based on real-time data acquisition of system status to eliminate failures just in time.
4. Proactive: systems are tracked and modeled to understand why they fail and this insight is integrated into the design and specification processes.
5. Ultra: reliability is integrated at every stage of the system life cycle and interactions among system components are modeled.

System Life Cycle

The system life cycle can be divided into eight stages with three cycles of iteration (*see Figure 1*). While this chapter cannot go into significant details about every stage in detail, an overview can provide insight into how ultrareliability can be integrated into each stage.

Requirements Specification

Requirements specification is where the measures that define system success are determined (*see Subjective Expected Utility*). This should include reliability metrics as well as performance metrics. Key issues to achieve ultrareliability involve identifying the key stakeholders and the metrics that are critical to each one. Balanced scorecards [1] can be used to balance the needs of each group. Methods to solicit requirements are being adapted from diverse fields such as ethnography and marketing [2]. Integrating requirements analysis into early stages of the design process is critical to ensure that they are effectively supported. Iterative testing cycles can be used to verify and validate their delivery. Achieving ultrareliability also

Table 1 Typical direct and indirect benefits achieved through the application of ultrareliability to a system

Direct benefits	Reduced need for corrective and preventive maintenance, leading to lower maintenance costs and less disruption of normal operations
	Reduced inventory requirements for spare parts
	Reduced risk of catastrophic failure, with an associated reduction in safety risk for human operators and others in the proximity of the system when it fails
	Increased operating life of the system
Indirect benefits	Streamlined design and development process
	Fewer project delays and rework
	Better customer satisfaction and loyalty
	Greater return on investment
	Better understanding of why things fail

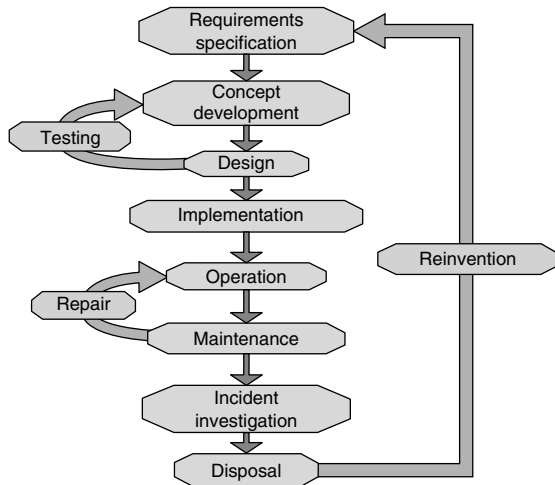


Figure 1 System life cycle for ultrareliability

requires that the landscape is continuously surveyed to track the evolution of stakeholder requirements and that these are integrated into the system even after it is launched.

Concept Development and Design

One of the key steps to maximize reliability is to use a systems perspective. This entails evaluating each component of the system, including architectural components, hardware, software, and user interface objects. Once the reliability of each component has been evaluated individually, the connections and modules are analyzed next. Then the full subsystems and the system as a whole are considered.

There are many modeling techniques that can be used individually or in combination to evaluate the reliability of components and subsystems. At the very least, a combination that considers both short-term and long-term failure profiles should be used. Adding safety factors, fault tolerances, and/or redundancy can increase the reliability of the system design. At a higher level, design strategies can pursue multiple concepts and test them against each other (see **Optimal Stopping and Dynamic Programming**).

Iterative Testing

Iterative cycles of testing are among the most effective strategies to maximize system effectiveness and reliability. As the design evolves and expands, additional cycles of testing should be executed. This can be accomplished cost-effectively and in a timely manner by using a variety of testing methods, such as simulation testing, low fidelity prototyping, and non-destructive testing. At later stages of the design process, integration testing and ecological field testing should be used. After the system has been launched, monitoring and reporting systems can be used to identify failures that were not predicted as well as validate the failure profiles that were not eliminated from the system design (see **R&D Planning and Risk Management**).

Operation, Maintenance, and Repair

After the system has been implemented, the next cycle should include operation, maintenance, and repair. While all design processes include these to some extent, there are techniques that can be

used to enhance reliability. At the operation stage, consideration should be given to the workflow, production schedules, workforce schedules, knowledge management, the use of automation, and housekeeping practices.

Workflows should be evaluated for their physical difficulty and cognitive complexity and how these are matched to the capabilities of the employee population. Production schedules should be reasonable, given the operational capabilities of machines and equipment. Workforce schedules should consider fatigue and error profiles of workers, given the challenges of vigilance [3] and fatigue [4]. Knowledge management requires a comprehensive strategy to facilitate participation [5]. Compliance and reliance-related system failures should be modeled for any automation that is used [6]. Finally, housekeeping practices should be evaluated for any possible links to degraded performance.

All systems require maintenance and repair operations at some point during their operation. These practices should be carefully designed to achieve the necessary system reliability. Maintenance and repair has a similar evolution to reliability in general. At the basic level, corrective maintenance involves using diagnostics to evaluate failures as they occur and repair them. Rapid intervention involves designing inspection practices to discover failures immediately as they occur to minimize the consequences of the failure. New developments in software, hardware, and material designs have created self-repairing systems that detect their own failures and take steps to repair them. The advanced phase of maintenance evolution is to take steps to prevent failures before they occur. This can be accomplished through preventive maintenance on fixed schedules based on historical failure rates, diagnostics that are monitored to identify states that may lead to failure for early intervention, and at the highest level, predictive modeling that predicts system failure based on individual system characteristics and context.

Incident Investigation

When unexpected/unplanned failures occur, a systematic incident investigation process is necessary to ensure that these failures are integrated into the operation/maintenance/repair cycle in the future. Note that this process is not about ‘accident’ investigation. The term accident connotes that the random aspect of the failure cannot be controlled. This

approach is unfortunate because an effective analysis can generally provide some control of every failure. Another common error in incident investigation is to attribute the failure to a proximate cause rather than continue the investigation until the root cause is identified and the root cause path is modeled.

There are two basic approaches to effective incident investigation. Root cause analysis involves combining theoretical and interview techniques to track the series of events that led to the failure. Stop rules are determined *a priori* to set a threshold for how far back to investigate.

The second approach is to establish an incident database that can be used to empirically analyze failures. As failures occur, the criteria and characteristics are stored in a database. The database is analyzed on a regular schedule to identify trends in failures and recurring contextual issues that contribute to failures.

Disposal

At the end of system life, either because of obsolescence or performance degradation, systems are disposed of. There are costs and consequences of this disposal that should be considered as part of the system design process. Some systems include hazardous wastes that can have environmental significance if not disposed of properly. Other systems may have warranties that can be managed to minimize the cost of replacement. Still other systems may have components that can be reused even as the rest of the system is scrapped. Typical strategies to plan for disposal include graceful degradation, obsolescence planning, end-of-life modeling, and disassembly planning. Environmental effects should also be modeled for all components and materials.

Interactions

One of the greatest challenges of ultrareliability is anticipating and modeling interactions among system components. The variability of failure profiles of each component are magnified when interactions are considered. There are several aspects of system performance that can be targeted to identify possible sources of failure that may be missed in a typical reliability analysis. These include management, training, communication, and human factors.

Management

Management is a common source of variability that can lead to system failure. At the broadest level, corporate cultures can develop that prioritize performance measures that are not aligned with customer demands. Similarly, supervisory and human resource practices can focus on unaligned measures. These can manifest in explicit organizational policies or in tacit practices that become accepted and permeate the organization.

Training

Even when system design and management practices are aligned, ineffective training can lead to system failure. Training design needs to consider the selection of media and content, the level of fidelity at which the training should be delivered, and the scheduling of training delivery. Employee performance in training and application must be measured and managed. Records must be maintained systematically, either through the use of certification or informally.

Communication

Ineffective communication is another contributor to system failure. As organizations become more geographically and culturally diverse, communication problems will be magnified [7]. The design and management of effective communication requires the consideration of communication mode, content, timing, tone, technology, and user characteristics. Modeling and controlling system failures requires consideration of the interactive effects of communication with design specifications.

Human Factors

Human factors involve describing the capabilities and limitations of all human components of the system and modeling on how they interact with system failure profiles. Categories of human capability include physical, motor, perceptual, cognitive, behavioral, and emotional. System requirements that exceed human capabilities clearly increase the likelihood and severity of failure. In addition, human capabilities change with context and these changes can also impact system failure profiles.

External Context

All systems are affected by their external context and this often has an impact on system reliability. For example, systems exist within an external culture depending on the region, the state of the economy, and the characteristics of internal and external customers. As globalization and international sourcing becomes more common, the ability of systems to succeed in a variety of societal cultures will increase in importance. The integration of value chains from basic resources through processing, manufacture, distribution, retail, delivery, and recently consumer participation creates new reliability challenges. Including these factors in the system design process and reliability analysis can enhance reliability significantly and reduce the impact or emergence of additional failures.

Case Study

On account of space limitations, this case study must be brief, but it will attempt to illustrate the contributions to ultrareliability of some of the key stages and interactions of the system life cycle and context. Consider the capacity expansion of an oil company in a remote region such as a deepwater ocean rig or a pipeline in Alaska or a desert. Clearly, this is an expensive endeavor, relying on efficient and effective processes, and maximizing reliability is critical. Recent examples such as BP [8] glaringly illustrate the importance of ultrareliability.

Ultrareliability in the System Life Cycle

The system life cycle shown in Figure 2 has many stages. An ultrareliability approach at each stage can impact the system's ultimate performance. At the user requirements stage, the wide variety of stakeholders for this kind of project must be balanced. Financial requirements must be met or the project should not be initiated in the first place. Design requirements must be specified to ensure that the system has the capacity to meet necessary flow rates and the durability to last as long as the expected lifetime. The requirements of the workforce stationed in a remote location must be considered to achieve the necessary productivity and reliability. Contextual requirements such as withstanding extreme weather conditions or accommodating the needs of local residents should be

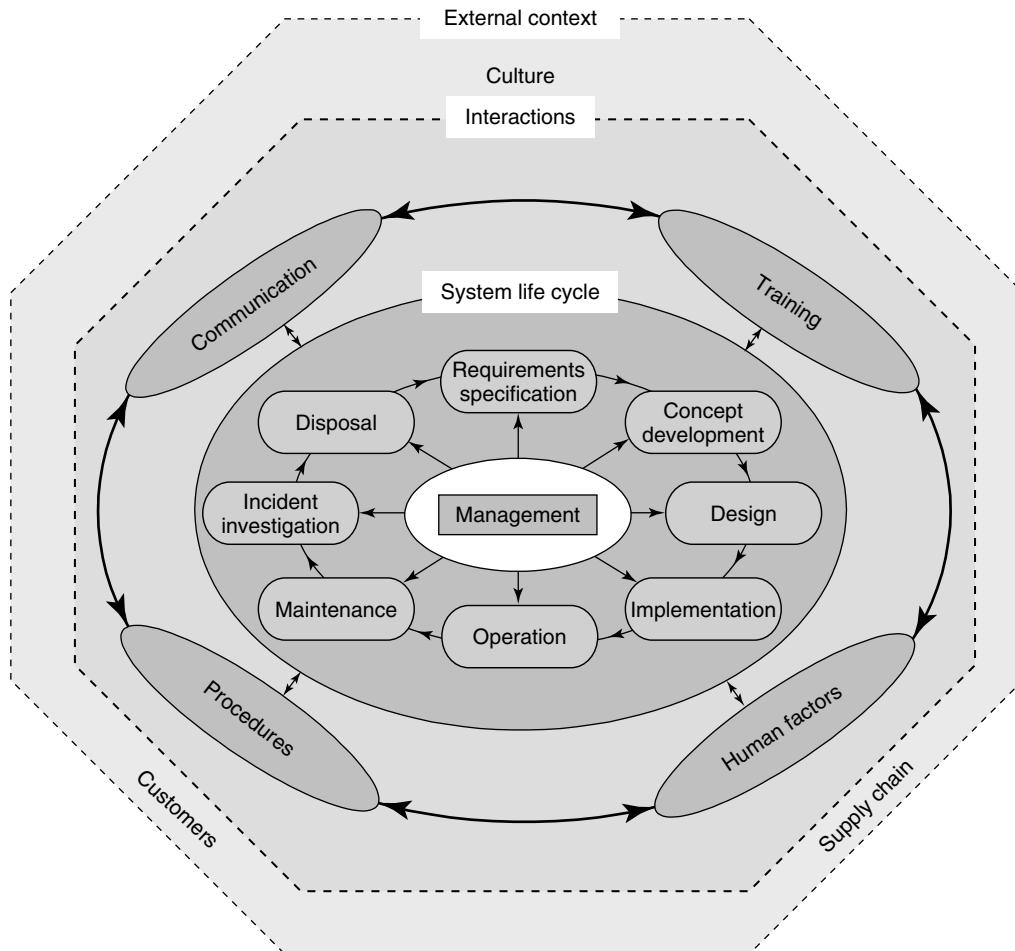


Figure 2 System life cycle, interactions, and external context for ultrareliability

included. Environmental and political requirements may also be critical for the system's success. Ignoring any of these aspects will reduce the eventual reliability of the system.

At the testing stage, using an ultrareliability approach is evident. Naturalistic testing must be used to evaluate the system performance in the harsh climate. Failures in remote areas or environmentally sensitive regions can be extremely expensive. Accelerated life testing should be used because of the extended duration for which the system is intended to last.

Advanced maintenance approaches, such as predictive maintenance and self-healing materials, are critical because bringing in maintenance personnel to

remote areas on an emergency basis can be difficult and costly. Scheduled noninvasive fatigue testing and similar measures can be used to verify that prediction models accurately reflect the degradation of the system over time.

Disposal is also an essential consideration because of the environmentally sensitive nature of the location and the use of hazardous materials, but in addition to having the appropriate procedures in place, measures must be taken to ensure that workers comply with these procedures and do not take shortcuts [9].

Reinvention is a stage that is often discounted, but is very important for long-term projects such as this case. New technologies, governmental policies,

environmental requirements, or economic realities can significantly change the requirements of the system over time. Constant monitoring of the landscape allows the company to be proactive in making changes and preempting legal or public relations challenges.

Interactions

Many of the interactions illustrated in Figure 2 significantly impact this case, particularly because of the remote location. Communication may not be always possible, so detailed procedures must be balanced with worker empowerment, to prevent on-site personnel from modifying important processes but allowing them to react to changing conditions. Training thus becomes essential to ensure that they are qualified, capable, and motivated to implement this balanced implementation. Management must be seen as committed to this approach or workers will discount the necessary activities to execute effectively. Each of these factors interacts, complicating the design of the system context and requiring an ultrareliability approach at all levels.

External Context

Last, but not least, the external context must be considered. The supply chain is critical because of the variety of technologies that must be combined to make a large infrastructure project of this kind of work. Compatibility issues will be the key. Any suppliers of outsourcing services must be also reviewed to verify that their personnel policies are also compatible. When workers from contractors work on the same facility as internal employees, different safety rules, incentive plans, and other management initiatives can play havoc with their ability to work together. System reliability can suffer as a result.

Conclusions

In order to elevate system reliability to the highest level, each of these aspects must be analyzed at the individual level, subsystem level, and integrated level. This perspective is the only way to achieve the goals of the ultrareliability process, but achieving these levels of reliability is not easy or simple, either from design or management perspectives. Ultrareliability can be addressed in two ways, depending on the commitment of management and the resources

available. A systematic, quantitative process to minimize all failure modes throughout the system can be attempted at the earliest stages of system design. This requires significant up-front investment of time and resources. Alternatively, a continuous improvement process can be used to reduce exposure to failure modes over time. This requires continuous attention, consistency of management support, and ongoing resource allocation to the ultrareliability effort, but the gradual nature of this approach is often easier to achieve.

References

- [1] Maltz, A.C., Shenhar, A.J. & Reilly, R.R. (2003). Beyond the balanced scorecard: refining the search for organizational success measures, *Long Range Planning* **36**, 187–204.
- [2] Sperssneider, W. & Bagger, K. (2003). Ethnographic fieldwork under industrial constraints: toward design-in-context, *International Journal of Human-Computer Interaction* **15**(1), 41–50.
- [3] Davies, D. & Parasuraman, R. (1982). *The Psychology of Vigilance*, Academic Press, London.
- [4] Circadian Technologies (2006). The secret cost of fatigue, *Managing 24/7 Electronic Newsletter* Received on 06/14/06.
- [5] Resnick, M.L. (2002). Knowledge management in the virtual organization, in *Proceedings of the 11th Annual Conference on Management of Technology*, International Association for Management of Technology, Miami.
- [6] Wickens, C.D., Lee, J.D., Liu, Y. & Gordon Becker, S.E. (2004). *An Introduction to Human Factors Engineering*, 2nd Edition, Pearson, New Jersey.
- [7] Goldbrunner, T., Doz, Y. & Wilson, K. (2006). The well-designed global R&D network, *Strategy + Business Summer*, Retrieved on 6/8/06 at <http://www.strategy-business.com/press/article/06217?pg=all&tid=230>.
- [8] Reed, S. (2007). BP feels the heat, *BusinessWeek*.
- [9] Resnick, M.L. (2007). The effects of organizational culture on system reliability: a cross-industry analysis, in *Proceedings of the Industrial Engineering Research Conference*, Institute of Industrial Engineers, Norcross.

Related Articles

Decision Trees

Multivariate Reliability Models and Methods

Stress Screening

Structural Reliability

Uncertainty Analysis and Dependence Modeling

Wags and Bogsats^a

“...whether true or not [it] is at least probable; and he who tells nothing exceeding the bounds of probability has a right to demand that they should believe him who cannot contradict him”. Samuel Johnson, author of the first English dictionary, wrote this in 1735. He was referring to the Jesuit priest Jeronimo Lobos’ account of the unicorns he saw during his visit to Abyssinia in the seventeenth century [1, p. 200].

Johnson could have been the apologist for much of what passed as decision support in the period after World War II, when think tanks, forecasters, and expert judgment burst upon the scientific stage. Most salient in this genre is the book *The Year 2000* [2], in which the authors published 25 “even money bets” predicting features of the year 2000, including interplanetary engineering and conversion of humans to fluid breathers. Essentially, these are statements without pedigree or warrant, whose credibility rests on shifting the burden of proof. Their cavalier attitude toward uncertainty in quantitative decision support is representative of the period. Readers interested in knowing how many of these even money bets the authors have won, and in other examples from this period are referred to [3, Chapter 1].

Quantitative models pervade all aspects of decision making, for example, failure probabilities of unlaunched rockets, risks of nuclear reactors, effects of pollutants on health and the environment, or consequences of economic policies. Such quantitative models generally require values for parameters that cannot be measured or assessed with certainty. Engineers and scientists sometimes cover their modesty with churlish acronyms designating the source of ungrounded assessments. Wild-ass guess “(Wags)” and bunch of guys sitting around a table “(bogsats)” are two examples found in published documentation.

Decision makers, especially those in the public arena, increasingly recognize that input to quantitative models is uncertain, and demand that this uncertainty be quantified and propagated through the models (*see Decision Modeling*).

Initially, the modelers were the ones who provided assessments of uncertainty and did the propagating. Not surprisingly, this activity was considered secondary to the main activity of computing “nominal values” or “best estimates” to be used for forecasting and planning, and received cursory attention.

Figure 1 shows the result of such in-house uncertainty analysis performed by the National Radiological Protection Board (NRPB) and The Kernforschungszentrum Karlsruhe (KFK) in the late 1980s [4, 5]. The models in question predict the dispersion in the atmosphere of radioactive material following an accident with a nuclear reactor. The figure shows predicted lateral dispersion under stable conditions, and also shows wider and narrower plumes, which the modelers are 90% certain will enclose an actual plume under the stated conditions.

It soon became evident that if things were uncertain, then experts might disagree, and using one expert-modeler’s estimates of uncertainty might not be sufficient. Structured expert judgment (*see Expert Judgment*) has since become an accepted method for quantifying models with uncertain input. “Structured” means that the experts are identifiable, the assessments are traceable, and the computations are transparent. To appreciate the difference between structured and unstructured expert judgment, Figure 2 shows the results of a structured expert judgment quantification of the same uncertainty pictured in Figure 1 [6]. Evidently, the picture of uncertainty emerging from these two figures are quite different.

One of the reasons for the difference between these figures is the following: the lateral spread of a plume as a function of down wind distance x is modeled, per stability class, as

$$\sigma(x) = Ax^B \quad (1)$$

Both the constants A and B are uncertain, as attested by spreads in published values of these coefficients. However, these uncertainties can *not* be independent. Obviously, if A takes a large value, then B will tend to take smaller values. Recognizing the implausibility of assigning A and B as *independent* uncertainty distributions, and the difficulty of assessing a joint distribution on A and B , the modelers elected to consider B as a constant; that is, as known with certainty.^b

The differences between these two figures reflect a change in perception regarding the goal of quantitative modeling. With the first picture the main effort

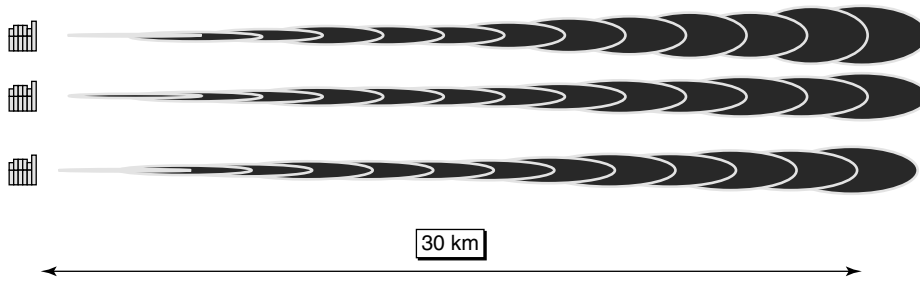


Figure 1 Plume widths (stability D) of 5, 50, and 95% computed by NRPB and KFK

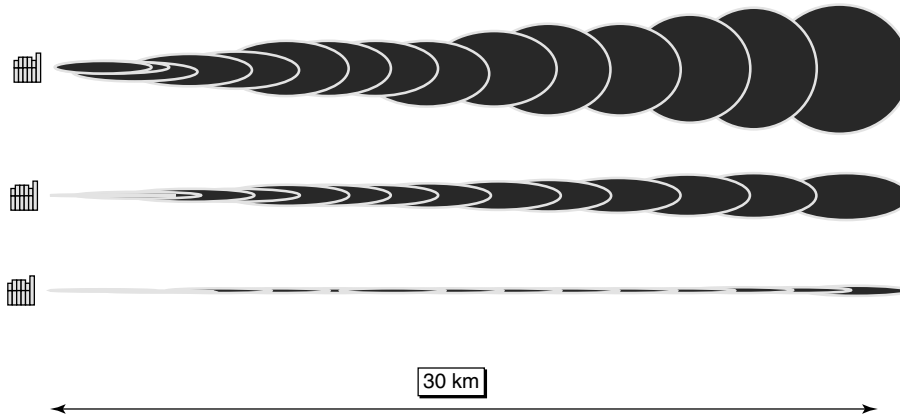


Figure 2 Plume widths (stability D) of 5, 50, and 95% computed by the EU-USNRC Uncertainty Analysis of accident consequence codes

has gone into constructing a quantitative deterministic model, to which uncertainty quantification and propagation are added on; however, in the second picture, the model is essentially about capturing uncertainty. Quantitative models are useful insofar as they help us resolve and reduce uncertainty. Three major differences in the practice of quantitative decision support follow from this shift in perception.

- First of all, the representation of uncertainty *via* expert judgment, or some other method is seen as a scientific activity subject to methodological rules every bit as rigorous as those governing the use of measurement or experimental data.
- Second, it is recognized that an essential part of uncertainty analysis is the analysis of dependence. Indeed, if all uncertainties are independent, then their propagation is mathematically trivial (though perhaps computationally challenging). Sampling and propagating independent

uncertainties can easily be trusted to the modelers themselves. However, when uncertainties are dependent, things become much more subtle, and we enter a domain for which modelers' training has not prepared them.

- Finally the domains of communication with the problem owner, model evaluation, etc. undergo significant transformations, once we recognize that the main purpose of models is to capture uncertainty.

Uncertainty Analysis and Decision Support: a Recent Example

A recent example serves to illustrate many of the issues that arise in quantifying uncertainty for decision support. The example concerns transport of *Campylobacter* infection in chicken processing lines.

The intention here is not to understand *Campylobacter* infection, but to introduce topics later in this entry. For details on *Campylobacter*, see [7–9].

Campylobacter contamination of chicken meat may be responsible for up to 40% of *Campylobacter* associated gastroenteritis and for a similar proportion of deaths. A recent effort to rank various control options for *Campylobacter* contamination has led to the development of a mathematical model of a processing line for chicken meat (these chickens are termed *broilers*).

A typical broiler processing line involves a number of phases as shown in Figure 3. Each phase is characterized by transfers of *Campylobacter* colony forming units from the chicken surface to the environment, from the environment back to the surface, from the feces to the surface (until evisceration), and the destruction of the colonies. The general model, applicable with variations in each processing phase, is shown in Figure 4.

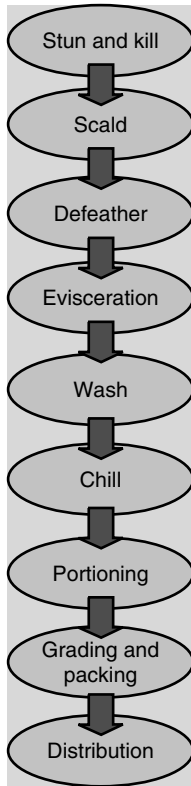


Figure 3 Broiler chicken processing line

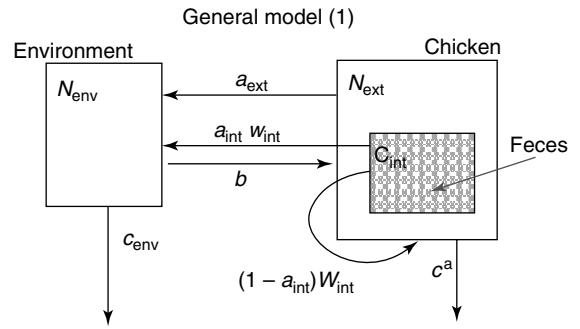


Figure 4 Transfer coefficients in a typical phase of a broiler chicken processing line

Given the number of campylobacter on and in the chickens at the inception of processing, and given the number initially in the environment, one can run the model with values for the transfer coefficients and compute the number of campylobacter colonies on the skin of a broiler and in the environment at the end of each phase. Ideally, we would like to have field measurements or experiments to determine values for the coefficients in Figure 4. Unfortunately, these are not feasible. Failing that, we must quantify the *uncertainty* in the transfer coefficients, and propagate this uncertainty through the model to obtain uncertainty distributions on the model output.

This model has been quantified in an expert judgment study involving 12 experts [9]. Methods for applying expert judgments are reviewed in [3, 10]. We may note here that expert uncertainty assessments are regarded as statistical hypotheses which may be tested against data and combined with a view to optimize performance of the resulting “decision maker”.

The experts have detailed knowledge of processing lines; but, owing to the scarcity of measurements, they have no direct knowledge of the transfer mechanisms defined by the model. Indeed, use of environmental transport models is rather new in this area, and unfamiliar. Uncertainty about the transfer mechanisms can be large; and, as in the dispersion example discussed above, it is unlikely that these uncertainties could be independent. Combining possible values for transfer and removal mechanism independently would not generally yield a plausible picture. Hence, uncertainty in one transfer mechanism cannot be addressed independently of the rest of the model.

Our quantification problem has the following features:

- There are no experiments or measurements for determining values.
- There is relevant expert knowledge, but it is not directly applicable.
- The uncertainties may be large and may not be presumed to be independent, and hence dependence must be quantified.

These obstacles will be readily recognized by anyone engaged in mathematical modeling for decision support beyond the perimeter of direct experimentation and measurement. As the need for quantitative decision support rapidly outstrips the resources of experimentation, these obstacles must be confronted and overcome. The alternative is regression to wages and bogsats.

Although experts cannot provide useful quantification for the transfer coefficients, they are able to quantify their uncertainty regarding the number of campylobacter colonies on a broiler in the situation described below taken from the elicitation protocol:

At the beginning of a new slaughtering day a thinned-flock is slaughtered in a 'typical large broiler chicken slaughterhouse'. ... We suppose every chicken to be externally infected with 10^5 campylobacters per carcass and internally with 10^8 campylobacters per gram of caecal content at the beginning of each slaughtering stage.... Question A1: All chickens of the particular flock are passing successively each slaughtering stage. How many campylobacters (per carcass) will be found after each of the mentioned stages of the slaughtering process, each time on the first chicken of the flock?

Experts respond to questions of this form, for different infection levels, by stating the 5, 50, and 95% quantiles, or percentiles of their uncertainty distributions. If distributions on the transfer coefficients in Figure 4 are given, then distributions, per processing phase, for the number of campylobacter per carcass (the quantity assessed by the experts) can be computed by Monte Carlo simulation: we sample a vector of values for the transfer coefficients, compute a vector of campylobacter per carcass, and repeat this until suitable distributions are constructed. We would like the distributions over the assessed quantities computed in this way to agree with the quantiles given by the combined expert assessments. Of course, we could guess an initial distribution over

the transfer coefficients, perform this Monte Carlo computation, and see if the resulting distributions over the assessed quantities happen to agree with the experts' assessments. In general, they will not, and this trial-and-error method is quite unlikely to produce agreement. Instead, we start with a diffuse distribution over the transfer coefficients, and adapt this distribution to fit the requirements in a procedure called *probabilistic inversion*.

More precisely, let X and Y be n - and m -dimensional random vectors, respectively, and let G be a function from $\mathbb{R}^n$ to $\mathbb{R}^m$. We call $x \in \mathbb{R}^n$ an inverse of $y \in \mathbb{R}^m$ under G if $G(x) = y$. Similarly, we call X a probabilistic inverse of Y under G if $G(X) \sim Y$, where $\sim$ means 'has the same distribution as'. If $\{Y|Y \in C\}$ is the set of random vectors satisfying constraints C , then we say that X is an element of the probabilistic inverse of $\{Y|Y \in C\}$ under G if $G(X) \in C$. Equivalently, and more conveniently, if the distribution of Y is partially specified, then we say that X is a probabilistic inverse of Y under G if $G(X)$ satisfies the partial specification of Y . In the current context, the transfer coefficients in Figure 4 play the role of X , and the assessed quantities play the role of Y .

In our campylobacter example, the probabilistic inversion problem may now be expressed as follows: find a joint distribution over the transfer coefficients, such that the quantiles of the assessed quantities agree with the experts' quantiles. If more than one such joint distribution exists, pick the least informative of these. If no such joint distribution exists, pick a "best fitting" distribution, and assess its goodness of fit.

In fact, the best fit produced with the model in Figure 4 was not very good. It was not possible to find a distribution over the transfer coefficients which, when pushed through the model, yielded distributions matching those of the experts. Reviewing the experts' reasoning, it was found that the "best" expert in fact recognized two types of transfer from the chicken skin to the environment. A rapid transfer applied to campylobacter on the feathers, and a slow transfer applied to campylobacter in the pores of the skin. When the model was extended to accommodate this feature, a satisfactory fit was found. The second model, developed after the first probabilistic inversion, is shown in Figure 5.

Distributions resulting from probabilistic inversion typically have dependencies. In fact, this is one of the

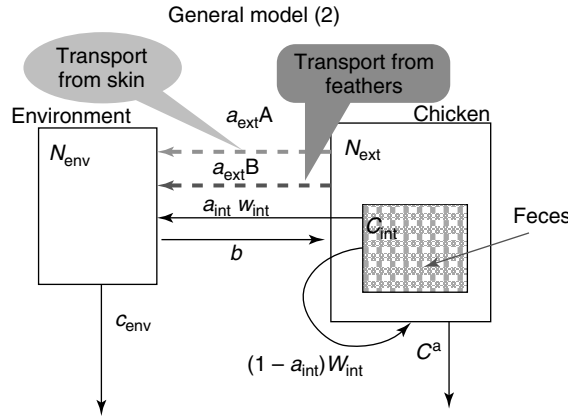


Figure 5 Processing phase model after probabilistic inversion

ways in which dependence arises in uncertainty analysis. We require tools for studying such dependencies. One simple method is simply to compute rank correlations. Rank correlation is the correlation of the quantile transforms of the random variables, where the quantile transform simply replaces the value of a random variable by its quantile. There are many reasons for preferring the rank correlation to the more familiar product moment correlation in uncertainty analysis.^c For now, it will suffice simply to display in Table 1 the rank correlation matrix for the transfer coefficients in Figure 5, for the scalding phase.

Table 1 shows a pronounced negative correlation between the rapid transfer from the skin (a_{extA}) and evacuation from the chicken (ca), but other correlations are rather small. Correlations do not tell the whole story; they are averages after all. To obtain a detailed picture of interactions, other tools must be applied. One such tool is the cobweb plot. In a cobweb plot, variables are represented as vertical lines. Each sample realizes one value

of each variable. Connecting these values by line segments, one sample is represented as a jagged line intersecting all the vertical lines. Figure 6 shows 2000 such jagged lines and gives a picture of the joint distribution. In this case we have plotted the quantiles, or percentiles, or ranks of the variables rather than the values themselves. Contracting the name of a_{extA} to axa , the negative rank correlation between axa and ca is readily visible if the picture is viewed in color: the lines hitting low values of a_{extA} are red, and the lines hitting high values of ca are also red.

We see that the rank dependence structure is quite complex. Thus, we see that low values of the variable ce (c_{env} , the names have been shortened for this graph) are strongly associated with high values of b , but high values of ce may occur equally with high and low values of b . Correlation (rank or otherwise) is an average association over all sample values and may not reveal complex interactions that are critical for decision making. One simple illustration highlights their use in this example.

Suppose we have a choice of accelerating the removal from the environment ce or from the chicken ca ; which would be more effective in reducing campylobacter transmission? To answer this, we add two output variables: $a1$ (corresponding to the elicitation question given above) is the amount on the *first* chicken of the flock as it leaves the processing phase, and $a2$ is the amount on the *last* chicken of the flock as it leaves the processing phase. In Figure 7, we have conditionalized the joint distribution by

Table 1 Rank correlation matrix of transfer coefficients, scalding phase

Variable	a_{extA}	a_{extB}	ca	b	ce	$aint$
a_{extA}	1.00	0.17	-0.60	-0.04	0.03	0.00
a_{extB}	0.17	1.00	-0.19	-0.10	-0.06	0.00
ca	-0.60	-0.19	1.00	0.01	0.02	0.00
b	-0.04	-0.10	0.01	1.00	0.02	0.00
ce	0.03	-0.06	0.02	0.02	1.00	0.00
$aint$	0.00	0.00	0.00	0.00	0.00	0.00

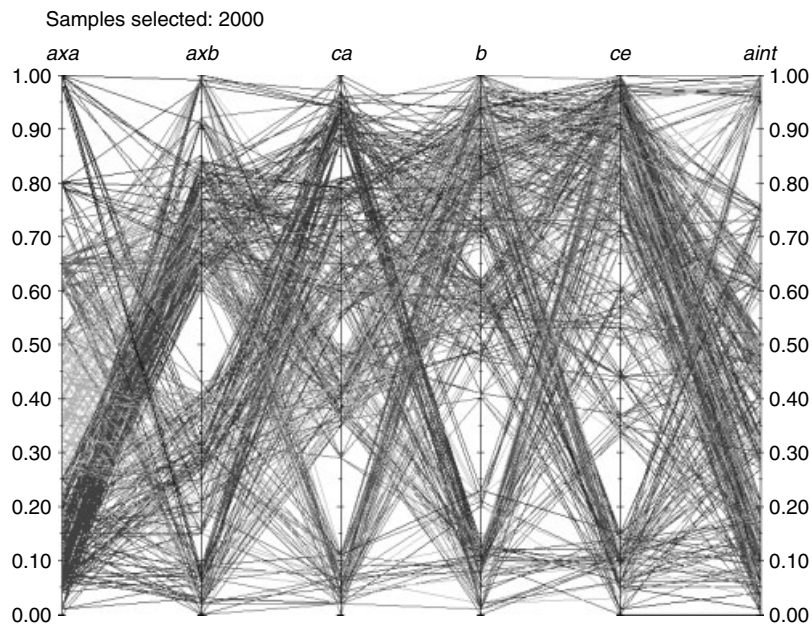


Figure 6 Cobweb plot for the transfer coefficients in the extended model

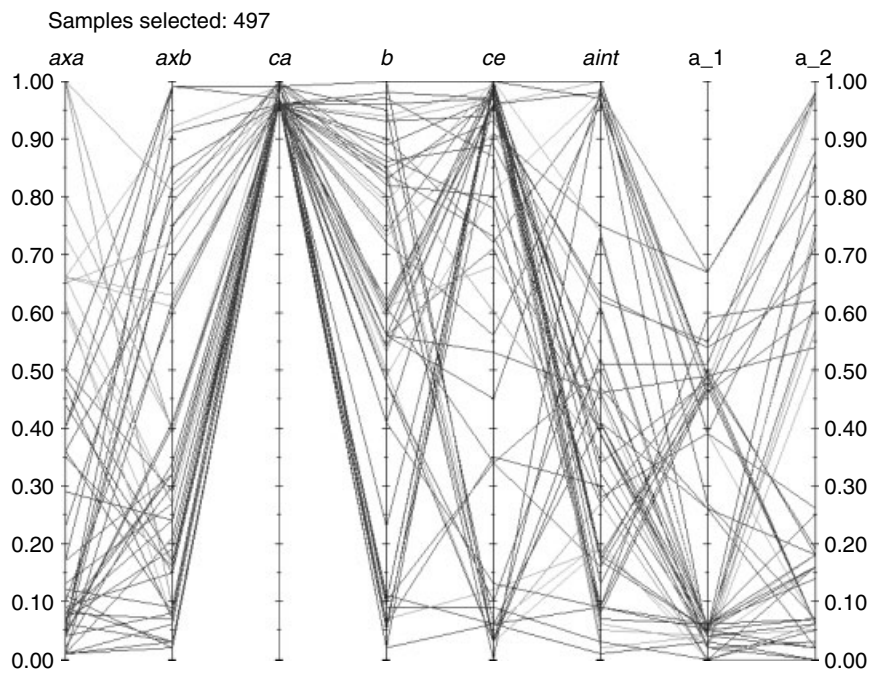


Figure 7 Cobweb plot conditional on high *ca*

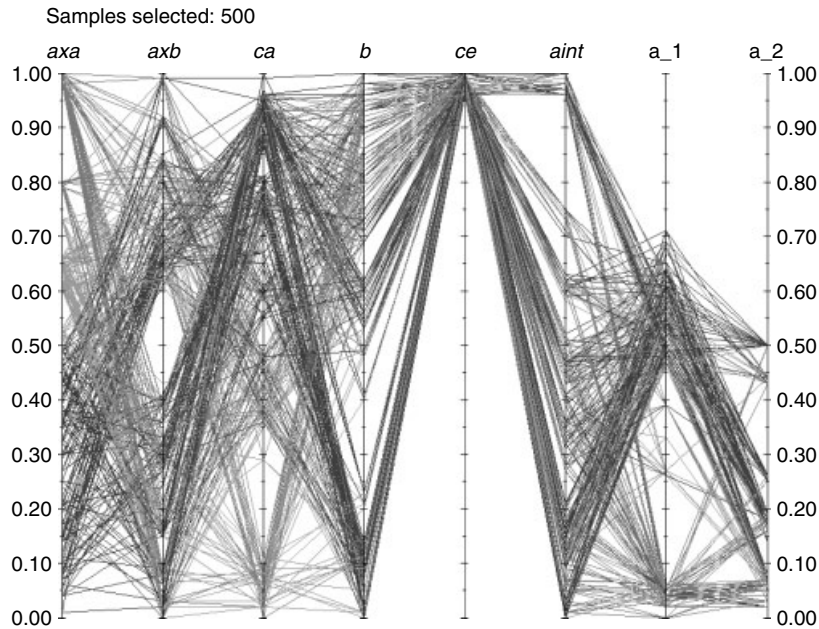


Figure 8 Cobweb plot conditional on high *ce*

selecting the upper 5% of the distribution for *ca*; in Figure 8, we do the same for *ce*.

Let us assume that the intervention simply conditionalizes our uncertainty, without changing other causal relations in the processing line (otherwise, we should have to rebuild our model). On this assumption, it is readily apparent that *ce* is more effective than that on *ca*, especially for the last chicken.

This example illustrates a feature that pervades quantitative decision support, namely, that input parameters of the mathematical models cannot be known with certainty. In such situations, mathematical models should be used to capture and propagate uncertainty. They should not be used to help a bunch of guys sitting around a table make statements that should be believed if they cannot be contradicted. In particular, it shows that

- Expert knowledge can be brought to bear in situations where direct experiment or measurement is not possible, namely, by quantifying expert uncertainty on variables that the models should predict.
- Utilizing techniques like probabilistic inversion in such situations, models become vehicles for capturing and propagating uncertainty.

- Configured in this way, expert input can play an effective role in evaluating and improving models.
- Models quantified with uncertainty, rather than wags and bogsats, can provide meaningful decision support.

Uncertainty Analysis – the State of the Art

The following remarks focus on techniques for uncertainty analysis that are generally applicable. This means, uncertainty distributions may *not* be assumed to conform to any parametric form and techniques for specifying, sampling, and analyzing high dimensional distributions should therefore be nonparametric.

Some techniques, in particular, those associated with bivariate dependence modeling, are becoming familiar to a wide range of users. Copulae, or multivariate distributions with uniform margins are used to capture rank correlation structure which is independent of marginal distributions (*see Copulas and Other Measures of Dependency*). While two-dimensional copulae are becoming familiar, multidimensional copulae remain extremely limited in their ability to represent dependence. Good books are available for bivariate dependence, such as [11–14].

High dimensional dependence models, sampling methods, postprocessing analysis, and probabilistic inversion are “breaking news” to nonspecialists, both mathematicians and modelers.

With regard to dependence in higher dimensions, much is not known. For example, we do not know whether an arbitrary correlation matrix is also a rank correlation matrix.^d We do know that characterizing dependence in higher dimensions *via* product moment correlation matrices is *not* the way to go. Product moment correlations impose unwelcome constraints on the one-dimensional distributions. Further, correlation matrices must be positive definite, and must be completely specified. In practice, data errors, rounding errors, or simply vacant cells lead to intractable problems with regard to positive definiteness. We must design other, friendlier ways to let the world tell us, and to let us tell computers which high dimensional distribution to calculate. We take the position that graphical models are the weapon of choice. These may be Markov trees, vines, independence graphs or Bayesian belief nets. For constructing sampling routines capable of realizing richly complex dependence structures, we advocate regular vines. They also allow us to move beyond discrete Bayesian belief nets without defaulting to the joint normal distribution. Much of this material is new and only very recently available in the literature [15–20].

Problems in measuring, inferring, and modeling high dimensional dependencies are mirrored at the end of the analysis by problems in communicating this information to problem owners and decision makers. This is sometimes called *sensitivity analysis* [21].

Whereas, communicating uncertainty has received some attention, much less attention has gone into *utilizing* uncertainty. It is safe to say that our ability to quantify and propagate uncertainty far outstrips our ability to *use* this quantification to our advantage in structured decision problems.

End notes

^a. This chapter is based on the first Chapter of [1] to which the reader may refer for definitions and details.

^b. This is certainly not the only reason for the differences between Figures 1 and 2. There was also ambivalence with regard to what the uncertainty should capture. Should it capture the plume uncertainty in a single accidental release, or the

uncertainty in the average plume spread in a large number of accidents? Risk analysts clearly required the former, but meteorologists are more inclined to think in terms of the latter.

^c. Among them the rank correlation always exists, is independent of the marginal distributions, and is invariant under monotonic increasing transformations of the original variables.

^d. We have recently received a manuscript from H. Joe that purports to answer this question in the negative for dimension greater than four.

References

- [1] Shepard, O. (1930). *The Lore of the Unicorn*, Avenel Books, New York.
- [2] Kahn, H. & Wiener, A.J. (1967). *The Year 2000: A Framework for Speculation*, Macmillan, New York.
- [3] Cooke, R.M. (1991). *Experts in Uncertainty*, Oxford University Press, New York.
- [4] Fischer, F., Ehrhardt, J. & Hasemann, I. (1990). Uncertainty and sensitivity analyses of the complete program system ufomod and selected submodels, Technical Report 4627, Kernforschungszentrum Karlsruhe.
- [5] Crick, J.J., Hofer, E., Johnes, J.A. & Haywood, S.M. (1988). Uncertainty analysis of the foodchain and atmospheric dispersion modules of MARC, Technical report NRB-P184, National Radiological Protection Board, Chilton, Didcot, Oxon.
- [6] Cooke, R.M. (1997). Uncertainty modeling: examples and issues, *Safety Science* **26**(1/2) 49–60.
- [7] Nauta, M.J., van der Fels-Klerx, H.J. & Havelaar, A.H. (2004). A poultry processing model for quantitative microbiological risk assessment, *Risk Analysis* (submitted).
- [8] Cooke, R.M., Nauta, M., Havelaar, A.H. & van der Fels, I. Probabilistic inversion for chicken processing lines, *Reliability Engineering and System Safety* (in press).
- [9] van der Fels-Klerx, H.J., Cooke, R.M., Nauta, M.J., Goossens, L.H.J. & Havelaar, A.H. (2005). A structured expert judgment study for a model of campylobacter contamination during broiler chicken processing, *Risk Analysis* **25**, 109–124.
- [10] Cooke, R.M. & Goossens, L.H.J. (2000). Procedures guide for structured expert judgement, Technical Report EUR 18820 EN, European Commission, Directorate-General for Research, Nuclear Science and Technology, Brussels.
- [11] Doruet Mari, D. & Kotz, S. (2001). *Correlation and Dependence*, Imperial College Press, London.
- [12] Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.
- [13] Nelsen, R.B. (1999). *An Introduction to Copulas*, Springer, New York.

-
- [14] Dall’Aglia, G., Kotz, S. & Salinetti, G. (1991). *Probability Distributions with Given Marginals; Beyond the Copulas*, Kulwer Academic Publishers.
 - [15] Bedford, T.J. & Cooke, R.M. (2001). *Probabilistic Risk Analysis; Foundations and Methods*, Cambridge University Press, Cambridge.
 - [16] Bedford, T.J. & Cooke, R.M. (2002). Vines – a new graphical model for dependent random variables, *Annals of Statistics* **30**(4), 1031–1068.
 - [17] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons, Chichester.
 - [18] Cowell, R.G., Dawid, A.P., Lauritzen, S.L. & Spiegelhalter, D.J. (1999). Probabilistic networks and expert systems, in *Statistics for Engineering and Information Sciences*, Springer-Verlag, New York.
 - [19] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufman Publishers, San Mateo.
 - [20] Kurowicka, D. & Cooke, R.M. (2006). *Uncertainty Analysis with High Dimensional Dependence Modeling*, John Wiley & Sons, New York.
 - [21] Saltelli, A., Chan, K. & Schott, E. (2000). *Sensitivity Analysis*, John Wiley & Sons, New York.

Related Articles

Bayesian Statistics in Quantitative Risk Assessment

Decision Analysis

Group Decision

Societal Decision Making

ROGER M. COOKE AND DOROTA KUROWICKA

Uncertainty and Variability Characterization and Measures in Risk Assessment

The risks experienced by individuals within a given population can vary appreciably owing to differences in susceptibility and exposure patterns [1]. Susceptibility can vary depending on genetic predisposition to certain diseases such as cancer or pharmacokinetic factors that control the rates of metabolism and elimination of xenobiotics from the body. Susceptibility can also vary as a function of age, e.g., developing tissues in the pediatric population are sometimes more susceptible than adult tissues. Exposure to environmental toxicants can markedly vary among individuals; food consumption patterns influencing dietary intake of naturally occurring or synthetic dietary components vary both among individuals and within an individual over time (*see Managing Foodborne Risk; Environmental Hazard*).

In the past, conservative assumptions have often been invoked in the face of uncertainty and variability. In the absence of evidence to the contrary, it is often assumed that humans are as sensitive as the most sensitive animal species tested. When exposure varies greatly, consideration is given to the risks faced by individuals at or toward the upper end of the exposure range. These conservative assumptions can compound and lead to much higher estimates of risk than warranted [2]. Although such conservative strategies are appropriate in the interests of public health protection, excessive conservatism can be cost ineffective. Moreover, by focusing on single-summary estimates of risk obtained by the application of conservative inference guidelines, one can lose sight of the fact that such estimates are more properly viewed as upper limits rather than as best estimates.

Distinguishing clearly between uncertainty and variability is very important in quantitative risk assessment [3–5]. Uncertainty represents the degree of ignorance about the precise value of a particular parameter (risk factor) such as the body weight of a given individual. In this case, uncertainty may be

caused by a systematic or random error associated with the instrument (e.g., a simple scale or more sophisticated mass balance) used to measure that parameter. Variability represents inherent variation in the value within the population of interest. In addition to being uncertain, body weight also varies among individuals. A parameter such as body weight, which can be determined with a high degree of accuracy and precision and may be subject to little uncertainty, can be highly variable. Other parameters may be subject to little variability but substantial uncertainty, or they may be both highly uncertain and highly variable.

For the purposes of this article, uncertainty and variability are defined as follows. When the input variable is fixed (deterministic in a broad sense) but unknown, the distribution of values obtained from repeated observations represents uncertainty. The repeated observations can be used to construct a confidence interval for which there is a given percent chance of bounding the true value. Note that large amounts of uncertainty indicate that the information base for decision making may be improved by additional research in areas where information is most doubtful. When the input variable is stochastic and represented by a distribution with known parameters, variability is present in the input variable. Note that, other things being equal, large amounts of real variability indicate that societal resources may be more efficiently targeted to reduce risk where it is most intense. An input variable can have both uncertainty and variability. When the input variable is stochastic and represented by a distribution with unknown parameters, variability is present in the input variable, and uncertainty is present in the parameters of the distribution of the input variable.

In the remainder of this article, we consider a general framework for the analysis of uncertainties in a general risk model in which the input variables are stochastic. In the section titled “General Risk Model Incorporating Uncertainty and Variability”, we describe a general risk model in which the input variables are subject to uncertainty, variability, or both. The section also describes a graphical as well as numerical measure of uncertainties. In the section titled “Discussion”, we present the discussion on uncertainty and variability in a general risk model. The article concludes with some observations about

the assumptions that underlie the various measures that we have proposed.

General Risk Model Incorporating Uncertainty and Variability

In this section, we describe briefly a general risk model that incorporates uncertainty and variability in risk factors. Rai *et al.* [6] have developed a general framework for characterizing and analyzing uncertainty and variability for arbitrary risk models. Suppose that the risk R depends on p risk factors $X_1, \dots, X_p$ according to the function

$$R = H(X_1, X_2, \dots, X_p) \quad (1)$$

Each risk factor X_i may vary within the population of interest according to some distribution with probability density function $f_i(X_i|\theta_i)$, which is conditional upon the parameter θ_i . Uncertainty in X_i is characterized by the distribution $g_i(\theta_i|\theta_i^0)$ for θ_i , where θ_i^0 is a known constant. The total uncertainty/variability in X_i (marginal distribution of X_i) is then described by the density function

$$c_i(X_i|\theta_i^0) = \int f_i(X_i|\theta_i) g_i(\theta_i|\theta_i^0) d\theta_i \quad (2)$$

If θ_i is a vector valued, g_i is a multivariate distribution. Here, it is assumed that the forms of the distributions f and g are known.

For any correlated pair (i, j) , let X_i and X_j have a joint distribution f_{ij} , conditional upon the parameter θ_{ij} ; here, f_{ij} represents the joint distribution of variability in X_i and X_j . The parameter vector θ_{ij} , in turn, has a (possibly multivariate) distribution g_{ij} with known parameter θ_{ij}^0 ; here, g_{ij} represents the joint distribution of uncertainty in X_i and X_j . As described

below, the first two moments of these bivariate distributions can be used to describe uncertainty and variability in risk and characterization of risk.

Relative Uncertainty/Variability

Rai *et al.* [6] define uncertainty and variability measures in a Taylor series expansion of equation (1) around the means of the risk factors. However, in the risk models, many risk factors are assumed to be lognormally distributed. Also many risk models are multiplicative [7]. Therefore, developing uncertainty measures and distributions of the risk based on $R^* = \log R$ may be robust and appropriate.

For any $i, i = 1, 2, \dots, p$, let $X_i^* = \log X_i$ and $R^* = \log R$. For uncertainty analysis, we need only two moments, mean and variance (or standard deviation). We describe the tree diagram in Figures 1 and 2 to identify the terms needed for the uncertainty analysis.

Note that

$$E(X_i^*) = E_{g_i} E_{f_i}(X_i^*) = E_{g_i}(m_i) = \mu_i^0 \quad (3)$$

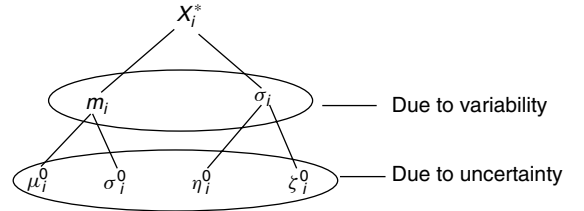


Figure 1 Summary of uncertainty and variability in a risk factor. The factors that have $\sigma_i = 0$ are subject to only uncertainty and not variability, and those that have $\sigma_i^0 = 0$ and $\xi_i^0 = 0$ are subject to only variability and not uncertainty

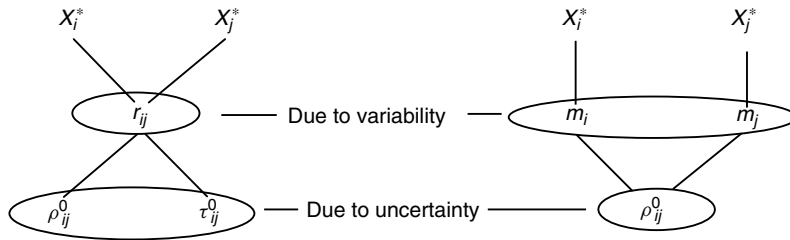


Figure 2 Correlation between a pair of risk factors. The correlated factors that have $\tau_{ij}^0 = 0$ are subject to only variability but not uncertainty

$$\begin{aligned}\text{Var}(X_i^*) &= \text{Var}_{g_i}(m_i) + E_{g_i}(\sigma_i^2) \\ &= \sigma_i^{0^2} + (\eta_i^{0^2} + \xi_i^{0^2})\end{aligned}\quad (4)$$

and

$$\begin{aligned}\text{Cov}(X_i^*, X_j^*) &= \text{Cov}(m_i, m_j) + E(\sigma_{ij}) \\ &\cong \tau_{ij}^0 \sigma_i^0 \sigma_j^0 + \tau_{ij}^0 \eta_i^0 \eta_j^0\end{aligned}\quad (5)$$

By expanding the risk R^* , i.e., risk measured on a logarithmic scale, around the point $(X_1^* = \mu_1^0, X_2^* = \mu_2^0, \dots, X_p^* = \mu_p^0)$ by using Taylor series expansion and ignoring all but the first-order terms, we have

$$\begin{aligned}R^* &\approx \tilde{R}^* = \log H(\mu_1^0, \mu_2^0, \dots, \mu_p^0) \\ &\quad + \sum_{i=1}^p h_i (X_i^* - \mu_i^0)\end{aligned}\quad (6)$$

where

$$\begin{aligned}h_i &= \frac{\mu_i}{H(\mu_1, \dots, \mu_p)} \\ &\times \left. \frac{\partial H(X_1, X_2, \dots, X_p)}{\partial X_i} \right|_{(\mu_1^0, \mu_2^0, \dots, \mu_p^0)}\end{aligned}\quad (7)$$

is the partial derivative of H with respect to X_i . This approximation reduces the log-scale risk R^* to a simple linear combination of the p risk factors measured on the log scale X_i^* . When the risk factor is multiplicative, $h_i = 1$. Rai and Krewski [8] defined uncertainty measures for the multiplicative model.

Partitioning Uncertainty (U) and Variability (V)

Following Rai *et al.* [6] and Rai and Krewski [8], we now derive an expression for U , which is the measure of uncertainty in R due to uncertainty in all of the risk factors combined, and V , which is the measure of variability in R that encompasses the variability in all of the risk factors. Without loss of generality, we can assume that the variables $X_1, \dots, X_{p_1}$ in equation (2) are subject only to uncertainty, that $X_{p_1+1}, \dots, X_{p_1+p_2}$ exhibit only variability, and that the remaining $p - (p_1 + p_2)$ variables are subject to both variability and uncertainty. Thus,

$$\begin{aligned}\tilde{R}^* &= \log H(\mu_1^0, \mu_2^0, \dots, \mu_p^0) + \sum_{i=1}^{p_1} h_i (X_i^* - \mu_i^0) \\ &\quad + \sum_{i=p_1+1}^{p_1+p_2} h_i (X_i^* - \mu_i^0) + \sum_{i=p_1+p_2+1}^p h_i (X_i^* - \mu_i^0)\end{aligned}\quad (8)$$

Following Rai and Krewski [8], we note that equation (8) leads to

$$\begin{aligned}U &= \sum_{i=1}^{p_1} h_i^2 \text{Var}(X_i^*) + \sum_{p_1+p_2+1}^p h_i^2 \text{Var}(m_i) \\ &\quad + \sum_{i=1}^{p_1} \sum_{j(\neq i)=1}^{p_1} h_i h_j \text{Cov}(m_i, m_j) = \sum_{i=1}^{p_1} h_i^2 \sigma_i^{0^2} \\ &\quad + \sum_{p_1+p_2+1}^p h_i^2 \sigma_i^{0^2} + \sum_{i=1}^{p_1} \sum_{j(\neq i)=1}^{p_1} h_i h_j \rho_{ij}^0 \sigma_i^0 \sigma_j^0\end{aligned}\quad (9)$$

and

$$\begin{aligned}V &= \sum_{i=p_1+1}^{p_1+p_2} h_i^2 \text{Var}(X_i^*) + \sum_{p_1+p_2+1}^p h_i^2 E(\sigma_i^2) \\ &\quad + \sum_{i=1}^{p_1} \sum_{j(\neq i)=1}^{p_1} h_i h_j E(\sigma_{ij}) = \sum_{i=p_1+1}^{p_1+p_2} h_i^2 \eta_i^{0^2} \\ &\quad + \sum_{p_1+p_2+1}^p h_i^2 (\eta_i^{0^2} + \xi_i^{0^2}) + \sum_{i=1}^{p_1} \sum_{j(\neq i)=1}^{p_1} h_i h_j \rho_{ij}^0 \eta_i^0 \eta_j^0\end{aligned}\quad (10)$$

The total uncertainty/variability in R due to all factors is given as

$$W = U + V\quad (11)$$

The total uncertainty/variability in R due to X_i is given as

$$W_i = U_i + V_i\quad (12)$$

where

$$\begin{aligned}U_i &= h_i^2 \text{Var}(m_i) + \sum_{j(\neq i)=1}^p h_i h_j \text{Cov}(m_i, m_j) \text{ and} \\ V_i &= h_i^2 E(\sigma_i^2) + \sum_{j(\neq i)=1}^p h_i h_j E(\sigma_{ij})\end{aligned}\quad (13)$$

Note that when the risk factors $X_1, \dots, X_p$ are independent, the covariance terms vanish. In this case, the evaluation of individual uncertainty and variability may be easy. Also note that some other alternatives for uncertainty and variability in a correlated risk factor defined in equation (13) are possible. These partitioned variances can be used to find the influential risk factors, contributions to risk made by uncertainty, variability, or both.

Following Rai *et al.* [9], these measures of uncertainty and variability are used to define

useful relative indicators showing the percentage of uncertainty and/or variability due to different sources. Specifically, the percentage of overall uncertainty and variability W in R due to uncertainty and variability W_i in X_i is denoted by $RW_i = (W_i/W) \times 100$. Similarly, $RU = (U/W) \times 100$ and $RV = (V/W) \times 100$ indicate the percentages of overall uncertainty and variability due to uncertainty U and variability V , respectively. Also, $RU_i = (U_i/W) \times 100$ and $RV_i = (V_i/W) \times 100$ are the percentages of overall uncertainty and variability W due to uncertainty U_i due to X_i and variability V_i in X_i , respectively.

Furthermore, $RU_i^{(a)}$ and $RV_i^{(a)}$ denote the percentages of uncertainty and variability W_i in X_i due to uncertainty U_i in X_i and variability V_i in X_i , respectively. The measures $RU_i^{(b)}$ and $RV_i^{(b)}$ denote the percentage of overall uncertainty U due to uncertainty U_i in X_i and the percentage of overall variability V due to variability V_i in X_i , respectively.

Risk Distributions

To provide the fullest possible characterization of risk, consider the distribution of R^* . Allowing for both uncertainty and variability in risk, the distribution of R^* is given by

$$C^*(R^* \leq r^*) = \int_{\log H(x_1, \dots, x_p) \leq r^*} \times \int c(X_1, \dots, X_p | \theta^0) dX_p, \dots, dX_1 \quad (14)$$

where $c(\cdot)$ is a joint-density function of all the risk factors. The parameter of this joint distribution is $\theta^0 = \{\theta_i^0, \theta_{ij}^0; i, j = 1, \dots, p\}$. The joint-density function, $c(\cdot | \theta^0)$, can be partitioned into two parts: $f(\cdot | \theta)$ represents the joint density due to the variability in the risk factors and $g(\cdot | \theta^0)$ represents the joint-density function due to uncertainty in the risk factors. The probability function (11) can then be further expanded as follows:

$$C^*(R^* \leq r^*) = \int_{\log H(x_1, \dots, x_p) \leq r^*} \times \int \{f(X_1, \dots, X_p | \theta) dX_p, \dots, dX_1\} \times \{g(\theta | \theta^0) d\theta\} \quad (15)$$

The mean and variance of the distribution of R^* can be approximated by $\log H(\mu_1^0, \mu_2^0, \dots, \mu_p^0)$ and W .

If all of the risk factors in equation (1) are lognormally distributed and the risk model is of the multiplicative form, then the distribution of R^* is normal with a mean $\log H(\mu_1^0, \mu_2^0, \dots, \mu_p^0)$ and variance W . Risk factors are typically assumed to be beta, triangular, or log-triangular distributed, if not lognormally distributed. In this case, they can be approximated very well by a lognormal distribution. Thus, if all of the risk factors are approximately lognormally distributed, then the distribution of R^* can be approximated by a normal distribution. The distribution of R^* can also be approximated by Monte Carlo simulation. Although straightforward, Monte Carlo simulation can become computationally intensive with a moderate number of risk factors.

In addition to examining the distribution of R^* and taking into account both uncertainty and variability in the $\{X_i\}$, it is of interest to examine the distribution of R^* considering only uncertainty in the $\{X_i\}$ or only the variability in the $\{X_i\}$. By comparing the distributions of R^* and allowing uncertainty and variability, only variability, or only uncertainty, one can gauge the relative contribution of uncertainty and variability to the overall uncertainty/variability in risk.

The density of R^* based on only uncertainty in the $\{X_i\}$ is the density of $\log H(X_1, X_2, \dots, X_p)$, but the risk factors $\{X_i^*\}$ are replaced by m_i , where

$$m_i = E_{f_i}(X_i^*) = \int_{-\infty}^{\infty} X_i^* f_i(X_i | \theta_i) dX_i \quad (16)$$

Note that this density depends on the distribution of the means of the log-scaled risk factors. For example, if the risk factor X_i has a lognormal distribution, then equation (16) reduces to a simple expression. Under some mild restrictions, this density can be approximated by a normal distribution with a mean $\log H(\mu_1^0, \mu_2^0, \dots, \mu_p^0)$ and variance U .

The density of R^* based on only variability in the $\{X_i\}$ is the density involving $\{\tilde{X}_i^*\}$ in the general risk model, where $\tilde{X}_i^*$ has distribution f_i with the following known parameter:

$$E_{g_i}(\theta_i) = \int_{-\infty}^{\infty} \theta_i g_i(\theta_i | \theta_i^0) d\theta_i \quad (17)$$

This density can also be approximated by a normal distribution with mean $\log H(\mu_1^0, \mu_2^0, \dots, \mu_p^0)$ and variance U .

The density of R^* based on both variability and uncertainty, only uncertainty, or only variability in the risk factor X_i can be approximated by a normal distribution with a common mean $\log H(\mu_1^0, \mu_2^0, \dots, \mu_p^0)$ and variances W_i , U_i , and V_i , respectively.

Risk due to Exposure to Trihalomethanes in Drinking Water

For demonstration purposes, we briefly describe the example discussed in Rai *et al.* [9]. Trihalomethanes (THMs) form in drinking water, primarily as the result of chlorination of organic matter that is naturally present in untreated water supplies. THMs are classified as probably carcinogenic in humans. Although the available epidemiological evidence was considered incomplete, preliminary data were consistent with the hypothesis that chlorinated water, if not THMs, is causally related to human bladder and colon cancer [10]. Using the model given in Krewski *et al.* [11], we can calculate the cancer risk as 1.69×10^{-3} for a lifetime exposure to 1 mg/kg-body weight/day of chloroform. This is comparable to the unit risk of THMs given in the US Environmental Protection Agency's Integrated Risks Information System [12].

We can apply our probabilistic risk assessment techniques to evaluate the potential risk associated with the interim maximum acceptable concentration

(IMAC) of THMs. The risk R is computed as follows:

$$R = \frac{UR \times IMAC \times C}{BW \times 10^3} = 3.6 \times 10^{-6} \quad (18)$$

Here, UR denotes the unit risk associated with lifetime exposure to 1 mg/kg-body weight/day (1.69×10^{-3}) with the $IMAC$ set at $100 \mu\text{g l}^{-1}$. The factor BW denotes the average body weight of an adult (70 kg), and C is the average daily consumption of drinking water for an adult (1.51 day^{-1}). Descriptions of the uncertainty and variability distributions are given in Table 1.

The results of uncertainty and variability analyses are given in Table 2. The uncertainty in the unit risk estimate contributes 67.8% to the overall uncertainty and variability; thus, it is the most influential factor. Of the remaining factors, the rate of consumption of drinking water (86.1%) exhibits substantially more variability than does body weight (13.9%).

The distributions of uncertainty, variability, and both uncertainty and variability are displayed in Figure 3. The distributions of R due to uncertainty in all of the input variables, due to variability in all of the input variables, and due to both uncertainty and variability in all of the input variables are lognormal with the same geometric mean (0.75×10^{-6}) but with different geometric standard deviations (1.21, 1.50, and 1.82, respectively). Considering both uncertainty and variability (for a random individual) in the input variables, the lifetime risk is about 3.6×10^{-6} for more than 90% of the population. Thus, about 10%

Table 1 Distributions of uncertainty and variability in the probabilistic risk assessment of chloroform

Factor	Transformed factor	Variability in X_i^*	Uncertainty in X_i^*
$X_1 = IMAC^{(a)}$	$X_1^* = \log_{10} X_1$	Constant = m_1	$\mu_1^0 = 2, \sigma_1^0 = 0$
$X_2 = UR^{(b)}$	$X_2^* = \log_{10} X_2$	Constant = m_2	$N^{(c)} (\mu_2^0 = -3.3925, \sigma_1^0 = 0.4161)$
$X_3 = BW^{(d)}$	$X_3^* = \log_{10} X_3$	Bivariate normal ($\mu_3, \sigma_3; \mu_4, \sigma_4; r_{34}$)	Multivariate normal
$X_4 = C^{(e)}$	$X_4^* = \log_{10} X_4$		$\mu_3^0 = 1.8451, \sigma_3^0 = 0.04139;$ $\eta_3^0 = 0.08636, \xi_3^0 = 0.04139;$ $\mu_4^0 = 0.1139, \sigma_4^0 = 0.04139;$ $\eta_4^0 = 0.23044, \xi_4^0 = 0.04139;$ $r_{34} \sim \text{Uniform}(0.3, 0.6)$

(a) $IMAC$: interim maximum acceptable concentration

(b) UR : unit risk

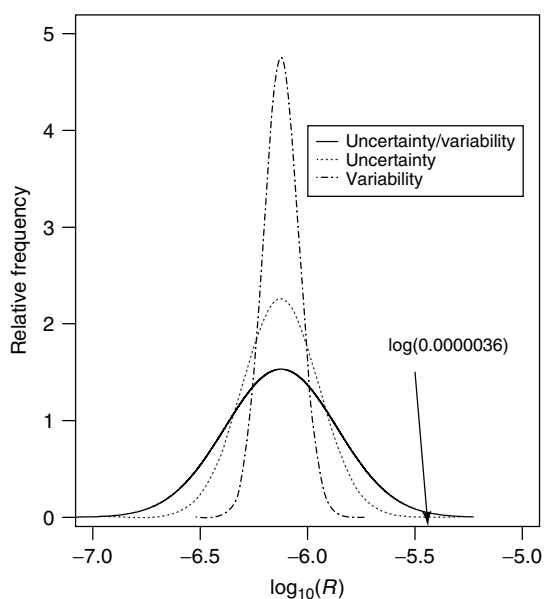
(c) N : normal distribution

(d) BW : average body weight

(e) C : daily consumption

Table 2 Uncertainty and variability in chloroform risks

Factor	Uncertainty			Variability			Both
	$RU_i = \frac{U_i}{W} \times 100$	$RU_i^{(a)} = \frac{U_i}{W_i} \times 100$	$RU_i^{(b)} = \frac{U_i}{U} \times 100$	$RV_i = \frac{V_i}{W} \times 100$	$RV_i^{(a)} = \frac{V_i}{W_i} \times 100$	$RV_i^{(b)} = \frac{V_i}{V} \times 100$	$RW_i = \frac{W_i}{W} \times 100$
$X_1 = IMAC^{(a)}$	0	0	0	0	0	0	0
$X_2 = UR^{(b)}$	66.5	98.1	100	0	0	0	66.5
$X_3 = BW^{(c)}$	0.7	1.0	2.3	27.7	86.1	97.7	28.4
$X_4 = C^{(d)}$	0.7	1.0	12.8	4.5	13.9	87.2	5.1
Total	67.8	100	NA ^(e)	32.2	100	NA ^(e)	100

^(a) IMAC: interim maximum acceptable concentration^(b) UR: unit risk^(c) BW: average body weight^(d) C: daily consumption^(e) NA: not applicable**Figure 3** Distributions of uncertainty and/or variability obtained using lognormal approximation in a multiplicative model for cancer risk (R) due to exposure to trihalomethanes in drinking water

of the population experiences higher risk than that set by the agency. To lower this percentage, the agency may want to adjust the *IMAC* level for this chemical.

Discussion

In this article, we have summarized a framework for addressing uncertainty and variability in a general

risk model for quantitative risk assessment. We have demonstrated an example of potential risks associated with the presence of THMs in drinking water. This method takes into account uncertainty and variability in each of the risk factors or input variables affecting risk to determine the overall uncertainty and/or variability in risk. Our methods also facilitate the identification of those factors that contribute the most to uncertainty and variability in risk. The objective of the example was not to revisit the established guidelines for acceptable THM levels in drinking water, but rather to illustrate how probabilistic risk assessment can provide a more complete characterization of risk than traditional best estimates or confidence limits of risk.

Note that our measures of uncertainty under a general (multiplicative or nonmultiplicative) model for risk utilize the full characterization of joint distribution of risk factors. Our measures take into account the resulting correlation structure due to any two risk factors.

One may fear that these results convey a false sense of reality about the nature of the data. However, on the contrary, we believe that such methodology helps to put the existing data in better context that could be used to better communicate the role of variability and uncertainty in determining risk and exposures.

References

- [1] National Research Council (1994). *Science and Judgment in Risk Assessment*, National Academy Press, Washington, DC.

- [2] Bogen, K.T. (1995). Methods to approximate joint uncertainty and variability in risk, *Risk Analysis* **15**, 411–419.
- [3] Morgan, M., Henrion, M. & Small, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, New York.
- [4] Hattis, D. & Burmaster, D.E. (1994). Assessment of variability and uncertainty distributions for practical risk analysis, *Risk Analysis* **14**, 713–730.
- [5] Hoffman, F.O. & Hammonds, J.S. (1994). Propagation of uncertainty in risk assessments: the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability, *Risk Analysis* **14**, 707–712.
- [6] Rai, S.N., Krewski, D. & Bartlett, S. (1996). A general framework for the analysis of uncertainty and variability in risk assessment, *Human and Ecological Risk Assessment* **2**, 972–989.
- [7] Brattin, W. (1994). *Quantitative uncertainty analysis of radon risks attributable to drinking water*, *Proceedings of the Toxicology Forum*, The Given Institute of Pathology, Aspen, pp. 338–350.
- [8] Rai, S.N. & Krewski, D. (1998). Uncertainty and variability analysis in multiplicative risk models, *Risk Analysis* **18**, 37–45.
- [9] Rai, S.N., Bartlett, S., Krewski, D. & Peterson, J. (2002). The use of probabilistic risk assessment in establishing drinking water quality objectives, *Human and Ecological Risk Assessment* **8**, 493–509.
- [10] Health Canada (1993). *Guidelines for Canadian Drinking Water Quality*, Supporting Documentation, Part II, Trihalomethanes, Environmental Health Directorate, Ottawa.
- [11] Krewski, D., Gaylor, D. & Szyszkowicz, M. (1991). A model-free approach to low-dose extrapolation, *Environmental Health Perspectives* **90**, 279–285.
- [12] USEPA (U.S. Environmental Protection Agency) (1985). *Drinking Water Criteria Document for Trihalomethanes*, Office of Drinking Water, Washington, DC.

Related Articles

Environmental Health Risk

Risk Characterization

Water Pollution Risk

SHESH N. RAI

Understanding Large-Scale Structure in Massive Data Sets

Massive data sets are ubiquitous in this day and age. New computer technologies have made and will continue to make data collection ever easier. At the same time increasing computer capabilities also enable development of new data analysis methods, but to what extent have the latter been brought to bear on the former? Or, do we require new thinking altogether to meet this challenge? (In this article we distinguish between massive data sets and massive data streams on the basis of whether or not they are stored. Massive data sets are recorded and available for follow-up analysis, while data streams are not. Analysis of massive data streams permit just one look at each datum as the stream flows.)

In 1995 the National Academy of Sciences Committee on Applied and Theoretical Statistics held a workshop on massive data sets. It included more than 50 participants from academia, industry, and government, who spoke about the nature of and challenges posed by massive data sets in such diverse areas as Earth and space science, information retrieval, social science, medicine, biology, business, manufacturing, telecommunications, high-energy physics, astronomy, text mining, national security, and government statistics [1]. At that time massive data sets were measured in gigabytes. Today they are measured in petabytes, and in the future they will no doubt be measured in exabytes and beyond. Nonetheless, many of the ideas presented at the workshop have borne fruit over the last decade. For example, there was extensive discussion at the workshop on the role of models in analyzing massive data sets. Later in this article we discuss the use of “local” models in understanding global structure.

We proceed as follows. First, we discuss various definitions of massive data sets. Then we review several analysis approaches that specifically make use of massiveness rather than simply scaling up existing techniques. This is followed by a discussion of how these methods can be used to understand the structure of massive data sets, and finally by a discussion of the challenges posed by massive data sets for understanding and managing risk.

What Are Massive Data Sets?

There are many ways to answer this question. Some have defined massive data sets on the basis of their sizes in bytes. For example, The Huber–Wegman taxonomy of massive data sets [2] is often cited, but now somewhat out of date in its absolute numbers. Others have concentrated on the idea of complexity both in information-theoretic and less formal senses [3]. In fact Wegman relates data set sizes to their computational, algorithmic, storage, transmission, and even visualization complexities [2], recognizing that it is difficult to give meaning to the idea of massiveness without some notion of how the data will be used.

Going even further, one can argue that massiveness depends not only on usage, but also on available resources: massiveness is in the eye of the beholder. An analyst with only a modest desktop computer may find a multiterabyte data set to be “massive” given his or her objectives while the same data set would not be considered massive by an analyst with access to a supercomputer.

A massive data set could be defined as one too large to be held in the memory of the user’s computer. This definition is criticized as overly broad, so instead we propose the following: a massive data set is a sample that is so large it can’t be analyzed like a sample. An analyst faced with such a data set might naturally choose to take and analyze a subsample from it, but this begs the question “what good are the original data if they are never used”?

Massive data sets occupy a middle-ground between populations and samples. They are like populations in the sense that making statements about them requires inference and sampling. They are like samples because they are collections of observations from even larger finite or infinite populations. If an analyst samples from and makes inferences about a massive data set, those inferences can be the basis for a second stage of inference back to the true population.

Figures 1 and 2 contrast the familiar statistical paradigm with a modified version that accommodates the role of massive data sets. Figure 1 shows schematically how a sample drawn from a population with distribution F , depending on unknown parameter(s) θ , is used to make inferences about θ . One computes a point estimate based on the sample, and it is the

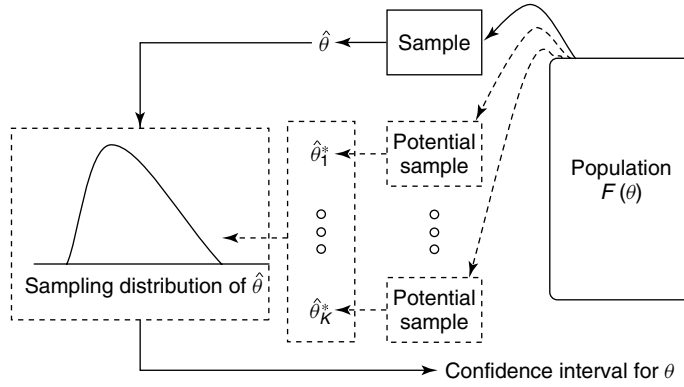


Figure 1 The familiar statistical paradigm. The population is assumed to follow distribution F , which depends on parameter(s) θ . Interest is in estimating θ . A single sample is drawn from F , and the statistic $\hat{\theta}$ is computed. The sampling distribution of $\hat{\theta}$ is derived either mathematically or by simulation, and in either case represents the distribution of $\hat{\theta}$ that would be obtained by drawing many samples and computing $\hat{\theta}$ for each one. The position of $\hat{\theta}$ in the sampling distribution determines a confidence interval for θ

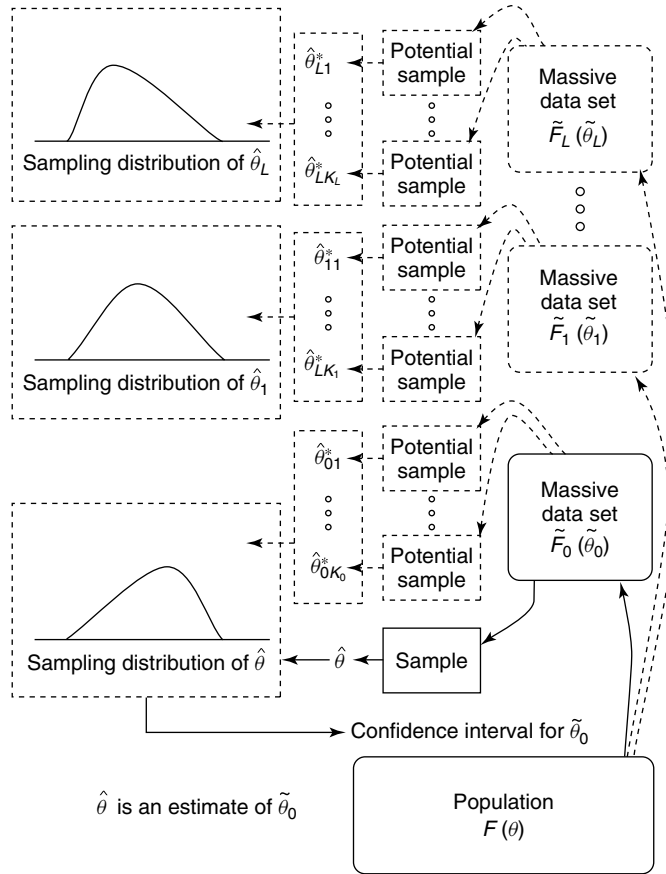


Figure 2 The statistical paradigm modified to incorporate the role of massive data sets

sampling distribution of that estimate, $\hat{\theta}$, that determines its uncertainty. The sampling distribution is represented in Figure 1 by the notion of other potential samples (designated by dashed lines) that may have been realized, and the values of $\hat{\theta}$ that would have been calculated from them.

In Figure 2, we depict our massive data set, $\tilde{F}_0$, as both a sample from the true population, F , and the data source for a subsequent sampling step. The latter is required because even though $\tilde{F}_0$ is itself a sample, it is too large to be treated like one according to our definition above. Just as there are unrealized potential samples that could have been drawn from F in Figure 1, there are unrealized potential massive data sets and unrealized potential samples from each of those in Figure 2.

To be more concrete, suppose as in Figure 1 that F is the underlying population distribution function, and that it depends on population parameter θ , e.g., the quantiles of F . F may be discrete or continuous. The massive data set collected by sampling from F is denoted by $\tilde{F}_0$, and is by definition a discrete, empirical distribution function. It depends on parameter $\tilde{\theta}_0$, which would be the quantiles of $\tilde{F}_0$ in this example. $\tilde{\theta}_0$ is unknown, and since $\tilde{F}_0$ is massive, it is not possible to calculate $\tilde{\theta}_0$ directly owing to the data set's size. Other massive data sets besides $\tilde{F}_0$ that could have been realized are represented by $\tilde{F}_1, \dots, \tilde{F}_L$, and their parameters by $\tilde{\theta}_1, \dots, \tilde{\theta}_L$. Each massive data set, including the one we actually have ($\tilde{F}_0$) gives rise to a set of potential samples, and implies a sampling distribution used to form an estimate of $\tilde{\theta}_i$. However, only the actual sample and the single estimate $\hat{\theta}$ are realized. If we are to draw conclusions about θ (rather than $\tilde{\theta}_0$) using only $\hat{\theta}$, we must account for the fact that there are now two sources of variation: one due to randomness in drawing our actual sample from the massive data set we actually have ($\tilde{F}_0$), and one due to randomness in drawing the massive data set we have from the population, F .

In Figure 1 we require knowledge of how our actual sample was obtained from the population in order to specify the sampling distribution of $\hat{\theta}$. In Figure 2 we require knowledge of how our actual sample was obtained from our actual massive data set, and also how our actual massive data set was obtained from the population. Notice that Figure 1 is actually embedded in Figure 2; in the terminology of the latter, the former gives a confidence interval for $\tilde{\theta}_0$,

not θ . In fact, the appropriate sampling distribution of $\hat{\theta}$ it as an estimate of θ is a mixture of all the distributions on the left side of Figure 2. Thus, it is crucial to understand and properly model the sampling procedures that take us from population to massive data set, and from massive data set to sample.

When faced with a massive data set, an analyst must first determine whether it is “massive” in the sense that it needs to be treated according to the model in Figure 2. The question really is one of model specification: is the goal to make an inference about the massive data set itself, or about the population from which it came? There is merit in both, and the answer depends critically upon the substantive problem at hand.

Massive data sets tend to be collected opportunistically from their parent populations, rather than according to a sampling design. For instance, NASA's Earth Observing System (EOS) remote sensing satellites are in polar orbit, and always observe the Earth within about 20 min of the same time of day, local time. A host of physical requirements determine orbit characteristics, instrument design, and data collection strategies before statistical representativeness is considered. If the quantity of interest is, say, column integrated atmospheric water vapor at 10:30 a.m. (and the overpass time is nominally 10:30 a.m.) then all is reasonably well because the massive data set $\tilde{F}_0$ turns out to be representative of the population, F . If, however, the quantity of interest is column integrated atmospheric water vapor over all times of day, $\tilde{F}_0$ is not representative of F . The need to account for the relationship between F and $\tilde{F}_0$ is generally not as well understood in massive data set analysis as is the need to account for the relationship between $\tilde{F}_0$ and samples drawn from it.

In many cases massive data sets result from (a) automated collection of vast amounts of information long before specific scientific hypotheses are formulated and (b) situations where it would be impossible to go back and collect the data over again if they were later found to be wanting. We call this the *kitchen sink* strategy: get everything you can (including the kitchen sink) now because you don't know if you'll need it eventually or not. For example, in manufacturing semiconductors, structures are built directly into chips that report on aspects of chip health for quality control purposes [1]. No formal statistical hypothesis underlies the decision to collect these data – just a general sense that the information will be useful.

Once the chips leave the manufacturing facility they can't be brought back, retooled, and reevaluated.

Modern technology makes it relatively easy to collect these observational data sets when the proper situations present themselves, but those situations are not generally repeatable or under experimental control. The tendency is to compensate with massiveness when opportunities do appear. Moreover, one does not generally know, at the time of collection, all the uses to which the data will be put, so the representativeness of $\tilde{F}_0$ for specific subsequent purposes is not clear.

Massive Data Set Analysis

Most existing work on massive data set analysis focuses on drawing conclusions about them rather than making inferences about their underlying populations. The broad area known as *data mining* is largely concerned with this problem, and in fact one definition of data mining is automated descriptive analysis of massive data sets. If massive data sets are thought to be representative of their underlying populations, then conclusions carry over, although uncertainty estimates are rarely, if ever, provided. When they are, uncertainties presented are usually determined by within-sample cross-validation, not by comparisons (see **Global Warming; Volatility Smile**) to the sampling distribution relevant to the underlying population (see Figure 2).

Nonetheless, a vast body of statistical, data mining (see **Early Warning Systems (EWSs) for Predicting Financial Crisis; Risk in Credit Granting and Lending Decisions: Credit Scoring**), and algorithmic work has been done to develop methods for understanding and quantifying structure in massive data sets. The lines separating these research areas are not crisp, but the distinctions are useful for discussion purposes. In this article we focus on approaches that are purely statistical in the sense that they have lineage reaching back to the fundamental statistical concept of data reduction.

The most direct link to data reduction arises from using probability models as descriptive summaries of data sets. For purposes of characterizing and ultimately understanding data structure, the model is a descriptive device. Although all data sets are discrete, models may be continuous if viewed as idealized descriptors, and so density estimation methods provide an important set of tools.

There is a large literature on density estimation (see **Enterprise Risk Management (ERM)**), but less on the problem of density estimation with massive data sets. Scott and Masahikow [4] provide a good discussion of the problem, and offer a potential solution called the *polynomial histogram*. Another line of work (Cheng *et al.* [5] for example) concentrates on scaling up with a “quick kernel density estimator” (see **Bayesian Statistics in Quantitative Risk Assessment**) that leverages a computational simplification obtained by constructing a pseudo-data set from the original data. Bayesian analysis methods that simulate posterior distributions such as Markov chain Monte Carlo (MCMC) (see **Non-life Loss Reserving; Reliability Demonstration**) are computationally expensive, so Ridgeway and Madigan [6] use a subset of the data for the computationally demanding portion. The rest of the data are used *via* importance sampling to reweight the initial MCMC draws. Similarly, Braverman *et al.* [7] use samples from a massive data set to construct adaptive histogram bins, which are then used to summarize all the raw data.

These last three examples all use some subset, subsample, or transformation of the original information to provide an initial solution, which is then sharpened by a computationally simpler step that incorporates the rest of the data. They all aim to approximate the distribution of the original data, either parametrically or nonparametrically. Distribution estimates can then be used to answer a wide range of questions about data structure.

Cheng *et al.* and Ridgeway and Madigan provide functional distribution estimates, nonparametric and parametric respectively. Braverman *et al.* produce a set of representative points, associated weights, and measures of variability that can be viewed as a coarsened representation of the original empirical distribution, or as a smaller compressed version of the original data that approximately retains their distributional characteristics. This is not unlike the method of data squashing [8, 9] which provides a representative set of points and weights such that characteristics of the likelihood function are preserved, and inferences based on squashed data are the same as inferences based on the original data. Pan *et al.* [10] define “the representative subset” as a “small subset of the original dataset that captures most of the original information compared to other subsets of the same size and has a low redundancy”.

They use an algorithm based on information-theoretic measures to determine such a subset.

There is a strong connection between these data reduction approaches and quantization in signal processing. Quantization, which is a form of source coding, contemplates a stream of random variables from an unknown probability distribution called an *information source*. Each realization, or signal, is to be transmitted over a channel with finite capacity. This means that signals can't be transmitted with perfect precision because their exact binary representations have too many bits. Thus, they must be quantized: each distinct potential realization is assigned to a group represented by a single value called a *codeword*. For each signal, only its group index (in binary form) is transmitted, and at the receiver the index is replaced by the codeword for the group as an estimate of the original signal. Codewords are normally the centroids of their groups. The number of groups and the assignment of potential signal realizations to them determine the average number of bits needed for transmission, and this must be less than the capacity of the channel.

The average number of bits needed for transmission is measured by the entropy (see **Expert Judgment**) (H) of the distribution of codewords. The error of the reproduced signal stream is measured by distortion (Δ), the average squared distance between the original signals and their codewords. The entire system is characterized by the entropy-distortion pair, and for every information source there is a theoretical curve called the *rate-distortion function* (see **Insurance Pricing/Nonlife; Risk Measures and Economic Capital for (Re)insurers**) which gives a lower bound on entropy as a function of distortion [11].

If we identify the signal stream with random draws from a data set, then we can say that the rate-distortion function is a descriptive characteristic of the data set. Further, we define data reduction as reduction in entropy from that of the original data. If every raw signal is unique (to infinite precision), then data reduction is $\log_2(N) - H$, where N is the number of data points. Thus there is a one-to-one relationship between data reduction and entropy.

The rate-distortion function provides a gold standard against which to judge any data reduction scheme that can be framed as a form of quantization. This includes, for example, clustering algorithms, and in fact some of the original quantization algorithms

in signal processing effectively perform clustering on training sets taken from the signal stream [12].

Braverman *et al.* appeal explicitly to this formalism, but the direct density estimation methods discussed earlier do not. Nonetheless, it is interesting to contemplate how the rate-distortion paradigm could be used to evaluate these techniques. For instance, data squashing suggests that distortion should be evaluated not with respect to the original data, but with respect to the quality of a particular estimator. Polynomial histograms, quick kernel density estimators, and Ridgeway and Madigan's ultimate estimates of posterior distributions could be evaluated with respect to a distortion measure for differences between distributions.

Understanding Large-Scale Structure

In many fields of study where massive data sets exist, one goal is to understand or describe what the data have to say about general relationships among important variables on large scales. The task is difficult because a key feature of massive data sets is that they provide details of *local* structure on a global scale. That is, massive data sets are not just big – they are complex, providing fine-scale information globally. Simple forms of data reduction that throw away fine-scale information thwart the effort.

As an example, consider the application with which the author is most familiar: massive remote sensing data sets. These data sets provide detailed information at high spatial and temporal resolution about structure and changes in the atmosphere. They typically include a large number of variables, and relationships among these variables are of primary importance (e.g., for the study of climate change). Moreover, phenomena of interest exist at various scales: hurricanes are transient in time and space, but oscillations like El Nino cover much larger spatial areas and occur over longer time periods.

The simplest way to reduce such a data set is to perform what is known as *gridding* in geoscience. One fixes attention on some spatial and temporal scale of interest, say monthly time resolution and 1° latitude by 1° longitude spatial resolution, and populates each space–time grid cell with the average values of the variables of interest. Gridding has the advantage of being simple and easy, but also has obvious drawbacks. Distributional information within grid cells is lost, including the very information that

might be most important: information about high-order relationships among variables.

Nonetheless, the essential idea of stratifying data according to values of well-understood variables and then describing these data distributions within strata is useful. We call this *chunking* the massive data set. In the remote sensing example, chunking is done using space and time. The relationship between two space–time grid cells is intuitive, and it’s easy to understand a comparison between data distributions defined in this way. Whether one uses simple means or more sophisticated data reduction methods, the idea is the same. Further, if the initial chunking is done at a fine enough level (e.g., daily, one-half degree latitude and longitude), then data distributions at coarser levels can be formed by aggregation provided the statistics chosen to describe the data are amenable.

In view of this strategy, large-scale structure in massive data sets may be understood as the evolution of the empirical distributions across suitably chosen chunks, including chunks that span different resolutions of aggregation. While this perspective follows naturally from working with spatial and temporal data, the idea can found elsewhere [13, 14] as well. Many methods have been proposed to estimate or describe empirical distributions. Some of these were discussed earlier in this article. Others include clustering algorithms [15], regression models [13], and sketches [16, 17].

Understanding large-scale structure in massive data sets requires compromise between data reduction and the need not to presume too much so the data may speak for themselves. Of course, this is the classic tension in all of statistics, so in many ways massive data sets are just the latest challenge in the ongoing evolution of the field. Some argue that the inevitable increase in computer power will take care of the problem eventually. However, data collection will always be easier than analysis, and so data set size will always outstrip computational resources. Massive data sets and analysis techniques for them will continue to be part of the statistical landscape no matter how large and powerful computers become.

What Challenges and Benefits Do Massive Data Sets Pose for Risk Analysis?

Massive data sets pose the same challenges for risk analysis that they pose generally in other areas, but

they also offer the possibility of mitigating risk by virtue of their complexity and comprehensiveness. The general pitfalls of massive data sets include difficulties in analyzing them because of their size, and the possibility that some important piece of information will be overlooked. On the other hand, the notion that massive data sets allow one to study local or fine-scale behavior on a global scale offers tantalizing opportunities to reduce risk.

For example, in climate studies a basic objective is to improve understanding of the behavior of the various components of the climate system. One such component is the transport of aerosol particulate matter around the world. Aerosol transport models codify the best understanding of how this complex system works, and how it changes in response to changes in other parts of the climate system. At the same time, massive observational data sets are being acquired by satellites to be used to help improve the models.

How do we determine whether the models are correctly representing the movement of aerosols around the world? This is an important question because the radiative effects of aerosols are among the most uncertain of the entire climate system [18]. Aerosol transport models can be used to produce massive synthetic data sets, and a natural starting point for model improvement is to ask whether the large-scale structures present in the satellite data are reproduced by the models.

Perhaps more importantly, can the satellite data sets be used to uncover why discrepancies exist between the model predictions and the data? The fine-scale detail in the observations play a crucial role: it is the ability to drill down to fine scales that makes massive data sets so useful. They preserve both large-scale context and fine-scale information. The trick is to get at both, so methods for understanding large-scale structure in massive data sets must preserve this link. If they do, then analysts can grasp the overall landscape and assimilate it as a guide to the more detailed information at finer levels.

In this example, improved models indicate better physical understanding and a consequent ability not just to make better predictions in the future but also to improve adaptation and mitigation strategies. This in turn reduces health risks in areas heavily impacted by aerosols, and increases ability to respond to climate impacts of changing distributions of various aerosol types.

This is just one example of how massive data sets can reduce risk in areas of fundamental importance to society. There are, of course, many others drawn from other fields. In any case, our ability to understand our world and improve our quality of life rests in no small part on data collection and analysis. Massive data set analysis plays a very important role in this endeavor, and will continue to be an area at the forefront of statistical research.

Acknowledgment

The research described in this article was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

- [1] National Research Council, Committee on Applied and Theoretical Statistics (1996). *Massive Data Sets: Proceedings of a Workshop*, National Academies Press, Washington, DC.
- [2] Wegman, E.J. (2005). *Lecture Notes on Data Mining*, at http://www.galaxy.gmu.edu/stats/lecture_notes.html.
- [3] Kettenring, J.R. (1997). Shaping statistics for success in the 21st century, *Journal of the American Statistical Association* **92**(440), 1229–1234.
- [4] Scott, D.W. & Sagae, M. (1997). Adaptive density estimation with massive data sets, *Proceedings of the Statistical Computing Section, Joint Statistical Meetings*, American Statistical Association, 104–108.
- [5] Cheng, K.F., Chu, C.K. & Lin, D. (2006). Quick multivariate kernel density estimation for massive data sets, *Applied Stochastic Models in Business and Industry* **22**, 533–546.
- [6] Ridgeway, G. & Madigan, D. (2002). Bayesian analysis of massive datasets via particle filters, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 5–13.
- [7] Braverman, A., Fetzer, E., Eldering, A., Nittel, S. & Leung, K. (2003). Semi-streaming quantization for remote sensing data sets, *Journal of Computational and Graphical Statistics* **12**(4), 759–780.
- [8] DuMouchel, W., Volinsky, C., Johnson, T., Cortes, C. & Pregibon, D. (1999). Squashing flat files flatter, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 6–15.
- [9] Berg, E., Dimakos, X. & Huseby, R.B. (2000). Squashing massive data sets: an overview of existing methods and ideas for further research, Technical Report No. 961, Norwegian Computing Center.
- [10] Pan, F., Wang, W., Tung, A.K.H. & Yang, J. (2005). Finding representative set from massive data, *Proceedings of the Fifth IEEE International Conference on Data Mining*, IEEE Computer Society, Washington, DC.
- [11] Cover, T.M. & Thomas, J.A. (1991). *Elements of Information Theory*, John Wiley & Sons, New York.
- [12] Gersho, A. & Gray, R.M. (1992). *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Boston.
- [13] Downing, D.J., Fedorov, V., Lawkins, W.F., Morris, M.D. & Ostrouchov, G. (1996). Large datasets: segmentation, feature extraction, and compression, Technical Report No. ORNL-13114, Computer Science and Mathematics Division, Mathematical Sciences Section, Oak Ridge National Laboratory.
- [14] McDermott, J.P., Babu, G.J., Liechty, J.C. & Lin, D.K. (2007). Data skeletons: simultaneous estimation of multiple quantiles for massive streaming datasets with applications to density estimation, *Statistics and Computing* **17**(4), 311–321.
- [15] Murtagh, F. (2002). Clustering in massive data sets, in *Handbook of Massive Data Sets*, J. Abello, P.M. Pardalos & M.G.C. Resende, eds, Kluwer Academic Publishers, New York.
- [16] Gibbons, P. & Matias, Y. (1999). Synopsis structures for massive data sets, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, A.
- [17] Cormode, G. & Muthukrishnan, S. (2003). Improved data stream summary: the count-min sketch and its applications, DIMACS Technical Report 2003-20.
- [18] Intergovernmental Panel on Climate Change (2001). *Climate Change 2001: Synthesis Report*, Cambridge University Press, London.

AMY J. BRAVERMAN

Use of Decision Support Techniques for Information System Risk Management

Identifying and understanding security risks for information system design or evaluation is very challenging. Making decisions about accepting or mitigating these risks through a rational, traceable, and understandable process is even harder. One reason is because adversaries are numerous – talented and motivated to attack information systems. Their motives can include harming the missions or functions supported by the information system, making a profit, enacting revenge, or just having fun. Another reason that analyzing security risks is hard is that adversaries observe the actions of the information system users and defenders and change their tactics in response to enhancements made to the system. In addition, information systems are becoming more complex and interdependent. Rapid changes in hardware and software make it nearly impossible to find and fix all system vulnerabilities, which can be introduced or exploited at many points in the system's life cycle.

Confounding the difficulty of security risk analysis are the numerous stakeholders involved in any information system. Stakeholders are individuals or organizations that have a vested interest in the information security and its solution [1]. Stakeholders such as users, system administrators, and system designers care about the system and have multiple and often conflicting objectives. System certifiers and accreditors are stakeholders too, and they need the appropriate risk information to make sound decisions about the system. Other information system stakeholders can include customers, maintainers, bill payers, regulatory agencies, sponsors, and manufacturers [2]. Ideally, one would like to include the points of view of all stakeholders in any decision about the security risks associated with an information system.

Quantitative risk assessments for information systems have many benefits. However, they also have their disadvantages. Both the advantages and disadvantages are discussed below.

The Advantages of Quantitative Risk Analysis for Information Systems

The largest benefit of a quantitative information system risk analysis is the ability to help stakeholders make risk-informed decisions by explicitly and clearly measuring risk and design trade-offs. Some quantitative methods, such as the multiple objective decision analysis (MODA) techniques described later, provide a rigorous, traceable, repeatable, and understandable process for communicating quantitative risk and design decisions to stakeholders and decision makers. Another advantage of quantitative risk analysis techniques is that they account for uncertainty [3], which is always a concern when making risk-related decisions. There are several good examples of quantitative information system risk assessment methods including Buckshaw *et al.* [4], Butler [5], and Hamill [6, 7]. These techniques have been successfully used to support multiple, large-scale network designs.

The Disadvantages of Quantitative Risk Analysis for Information Systems

The biggest disadvantage of quantitative security risk assessments for information systems is that they are hard to correctly develop, analyze, and explain. Quantitative risk assessments require experienced quantitative risk analysts, resources (including access to information system experts and threat experts), and time to complete a risk analysis. There are numerous pitfalls that a risk analyst needs to avoid to ensure that his/her quantitative risk recommendations are meaningful and valid [8]. One major pitfall is mistaking semiquantitative risk models for being quantitative. Semiquantitative models often translate qualitative scales into numeric scales. For example, the words “high”, “medium”, and “low” might be given the numbers “1”, “2”, and “3”. Often, these scales are ordinal, meaning that the only information they contain is a numeric rank order of benefit. This limits the legitimate mathematical operations allowed on these numbers [9]. These models can present a false appearance of precision and objectivity, and may encourage a level of confidence that may not be fully justified [10]. A majority of the information system risk assessments observed by the author can be described as semiquantitative. Some additional pitfalls include misunderstanding

of the meaning of weights used in the quantitative assessment [11], improperly measuring the effect of risk on the stakeholders' mission, and treating high-impact, low-probability events as the same as low-impact, high-probability events.

The remainder of this paper assumes that the reader values a quantitative risk assessment approach, and describes more about the people, processes, fundamental concepts, and mathematical interpretations of risk. The first section reviews the different stakeholders and their risk preferences. To help the stakeholders make risk-informed decisions, they need to understand some basic risk concepts, which are described in the second section. A formal process to analyze and adjudicate risk decisions is discussed in the third section. Finally, the fourth section describes different methods for quantitative risk modeling and their advantages and disadvantages.

Decision Support Stakeholders

Information systems have many stakeholders who are interested in the successful design, certification, and operation of the system. However, each individual stakeholder has his/her own wants and needs from the information system, and often these objectives conflict with the objectives of other stakeholders. For example, a user might want a system that is easy to use and has a great deal of functionality. However, a system administrator might want to enforce strong passwords (which users find harder to use) and remove instant messaging (a capability that users may desire) in order to remove vulnerabilities that can be exploited by network attackers. The major concerns and needs of all of the stakeholders need to be considered in the risk model. Individual stakeholders may not get the system that is best for them, but they may better understand why design decisions and trade-offs had to be made to satisfy the other stakeholders (*see Experience Feedback; Multiattribute Modeling*).

Users

Users are the individuals for whom the information system was, is, or will be designed. Users have a mission to perform and the information system should enable them to perform their mission in an effective and efficient manner. The information system should

be available and provide accurate information to only authorized users at all times, even when under attack. A typical user is primarily interested in an easy-to-use, functional solution. Confidentiality of data, availability of data and services, and integrity of data are desired, or even required, but only in a transparent manner.

System Administrators

System administrators are individuals who actively manage and maintain the services and configuration of the information system. They want to keep their system available and as easy to use as possible, but not at the cost of network security. Administrators are very busy and in high demand. They look for security solutions that do not take a lot of time to review or maintain. Security solutions that do require too much attention (e.g., audit logs that require manual review or intrusion detection systems that provide too many false positives) may be ignored or not used at all.

System Owners

System owners are individuals who are accountable for accomplishing a specified mission using the services of an information system that they own and operate. They have to balance the needs of the operating organization with the efficiency of the user. They have one eye on the cost and the other on the schedule. The owner not only wants to make the user happy but is also willing to give up a little functionality and ease of use, if it means that the network will be safe from attack. Owners are often the decision makers for information system risk and design decisions.

Decision Support Basic Concepts

There are four basic concepts for considering the risk to a complex, interconnected information system:

1. **Interdependency of system components**

The first concept involves the fact that in networks there are complex interdependencies of components that are not easily captured in system descriptions and models. Small perturbations in the complex networks, channeled through these subtle interdependencies, can create conditions for cascading failures that can

lead to failures throughout the system, such as loss of network connectivity.

2. Value measurement

The second concept is that we can measure the value of an information system to the multiple stakeholders. Values are simply things that we care about [12] or desired characteristics of the system and its services. Simple mathematical techniques exist that allow one to measure value across many performance and security measures.

3. Value trade-offs

The third concept is that in most cases an increase in performance in one function of the system often leads to a decrease in value of another function. The most common example is that, in most cases, any increase in performance comes at a monetary cost. Thus increasing the value in one area (performance) decreases the value in another area (cost). Often, increases in the level of security of an information system are accompanied by the loss of a functionality. One might remove an instant messenger feature in order to remove an exploitable vulnerability, but this will result in the loss of a rapid, user-friendly form of communication. The system owner or designer will have to make a trade-off decision, carefully weighing the difference in value of one design choice over another.

4. Return on investment

The fourth basic concept is return on investment (ROI). ROI measures the benefit that is returned for the investment (the costs of purchasing or implementing risk-mitigating strategies). This means that we need to consider the benefits and the costs.

Interdependency of System Components

An important concept that should be considered in any risk and design effort is the concept of cascading failures. A cascading failure is a phenomenon where a single disrupting event, such as the failure of a router, causes the entire network to fail. When the first node fails, its load is transferred automatically to other nodes. If one of these replacement nodes becomes too overloaded, it also fails, causing its assigned load, plus the additional load of the first node, to be assigned to other nodes. This continues until an entire network collapses. Cascading failures can and do happen. Examples include a 1986 Internet congestion

collapse, the 1996 power outage in western United States, and probably the August 2003 power outage in the northeastern United States [13]. Cascading failures are often not recognized by designers because of the use of static models that might only examine the effects of a nearest neighbor. However, dynamic models and simulations exist that can be used to stress the network under various attack scenarios.

Value Measurement of an Information System

A key concept for integrating risk analysis and design analysis is the idea of maximizing the benefit of an information system. An information system has value because it provides important information and services. Risk-mitigating designs provide value by ensuring a secure and available system in a hostile and malicious environment. In some cases, the decision might be easy if there is only one objective to worry about, such as minimizing cost. However, often there are several competing objectives, such as security risk, cost, schedule, and performance. In addition, there are many stakeholders, each with his/her own sets of assumptions and preferences across those objectives. One can create a value model to help determine the overall worth of any design choice, given the values and objectives of an organization. By using a value model, one can easily trade-off concepts that are greatly different such as security risk and system's ease of use.

Value Trade-offs

Most information system design choices require some form of a trade-off. The simplest example would be that an additional capability or enhancement to an information system will incur a cost or schedule increase. A trade-off analysis can help determine if the benefit of the choice is worth the cost considering all stakeholders. However, some trade-off decisions are not so easy to make since they involve complex issues besides cost or schedule. For example, the network owner might be faced with a decision on how to implement the rules of an intrusion detection system. Increasing the sensitivity of an automated intrusion detection system will help ensure that more adversary attacks are detected and stopped. However, increased sensitivity might mean that there are many more false positives. This can cause a decrease in network performance and can inure system security engineers to

real attacks. A decision analysis framework is tailored to help decision makers explicitly measure the trade-offs required, especially if the trade-offs are hard to measure.

Return on Investment

ROI can be defined as the profit from an investment divided by the money invested. For an information system, ROI can be more generally defined as the value provided by investments in the information system (e.g., capabilities gained) divided by its cost. It is possible to measure the ROI of increased security to an information system by determining the cost of the countermeasures and estimating the reduced operating and insurance costs from finding and fixing a network intrusion. However, some organizations are not able to measure ROI using monetary gains. For example, government organizations need to measure the benefit (value) of the network in terms of keeping secrets safe, ensuring the availability of critical data and services, and assuring that sensitive data has not been altered. These organizations can measure ROI by determining the overall value of the architecture (availability, accuracy, protection, and confidence in the information and services of the system) before and after a risk-changing design. Essentially, the analysis is a system-level trade-off study where the value of the architecture takes into account several individual benefits and cost trade-offs.

Decision Support Processes

Once the information system stakeholders understand the basic decision concepts, they need to fit the concepts into a decision process. Many information systems of large organizations must undergo a certification and accreditation process before they become operational. Certification is the comprehensive evaluation of technical and nontechnical features of an information system to establish the extent to which a particular design meets a specified security requirement. Often the certifier of the system will mandate technical evaluations of the system to document the certification. An accreditor is the official with the authority to formally assume responsibility at an acceptable level of risk. However, the certification and accreditation processes are often viewed as “a

picayune process where auditors inspect reams of security documentation on an agency’s IT system and infrastructure” [14]. However, quantitative risk modeling can greatly help with the certification and accreditation processes by providing a clear statement of information system residual risk and provide a recommendation for an acceptable level of risk. Additionally, a quantitative risk model can be used to prioritize recommended system design changes or countermeasures on a risk and cost basis.

Certification

The certifier determines whether a system is ready for certification and conducts the certification process. At the completion of the certification effort, the certifier reports the status of the certification and recommendation to the accreditation official for an accreditation decision. Often the certifier will evaluate the security compliance and assess residual risk. A quantitative risk assessment can be used by the certifier to identify the overall system risk and determine the best ways to reduce the risk while keeping costs down and having minimal negative effects on the users’ capabilities or the performance of the network. A prevalent philosophy is that the certifier should be independent from both the users and developers. However, any design effort could greatly benefit from having a team member with certification experience who can help design security into the system. Including a certifier on the design team can also help ensure that a system is fully accredited by the time it is deployed.

Evaluation

Certifiers use a variety of evaluation techniques to ensure compliance with security policies. Security test and evaluation teams examine and analyze the safeguards required to protect an information system as they have been applied in an operational environment. Penetration testing and “Red Teams” are techniques used to simulate the activities of a hostile adversary attacking the information system from inside or outside of the network boundary. Tests can also be performed to verify the integrity of the hardware and software throughout the design, development and delivery, to determine if there are excessive emanations from the equipment and to

verify that no unauthorized connections exist. The evaluators are the primary source of data for the quantitative risk assessment and the risk-management review.

Accreditation

The accreditor assesses the vulnerabilities and residual risk of the system and decides to accredit, give an interim authority to operate, or terminate system operations. Accreditation is the official decision by an information technology management official to authorize operations of an information system and to explicitly accept the risk to organizational operations, assets, or individuals based upon the implementation of agreed-upon security controls. Quantitative risk assessment methods can be of great value to the accreditor by providing a framework for assessing the risk of any system, and across systems, using the same metrics.

Defining Risk for Information Systems

Risk can be mathematically defined as a measure of the probability and severity of adverse effects [15]. Making information system design decisions under uncertainty is a complex undertaking. Compounding the problem is a general lack of historical risk data for most information system problems, as compared to the historical data that is available for an environmental risk assessment. However, in general, most information system risk assessments claim that risk is some function of “threat”, “vulnerability”, and “consequence” (sometimes called *impact*). Different models interpret and measure these three parameters differently. Three techniques seen in risk modeling are qualitative, semiquantitative, and quantitative. Well constructed qualitative models can be very useful, especially for quick, high-level assessments, but are beyond the scope of this section. See Smith *et al.* for a discussion on the advantages and disadvantages of qualitative risk assessments [16].

Quantitative Risk Models

Quantitative risk models for malicious activity can determine risk by considering the “threat” as a probability that an adversary will attempt an attack ($p(a)$),

“vulnerability” as the probability of adversary success given that they attempt the attack ($p(s|a)$), and consequence as the negative effects given a successful attack ($C|s, a$). Two quantitative risk methods include probabilistic risk analysis and MODA.

Probabilistic Risk Analysis. Probabilistic risk analysis assigns probabilities to the $p(a)$ and $p(s|a)$ elements in the quantitative risk model. Consequences are either known or can be assigned a probability over a range of consequences. One can analyze the risk distributions and calculate expected values of the risk. The concept of expected value risk is similar to that of expected value in probability theory – the multiplication of the probability of an event and its outcome value. Essentially, this is the average outcome that one would expect from an attempt of a particular attack. Since attacks generally have a binary outcome (success or failure) one might not even realize the “expected” risk. Mathematically, the expected risk to a system for a particular attack on a given information system, over a given period of time is shown by equation (1).

$$Risk = p(a) \times p(s|a) \times C|s, a \quad (1)$$

The risk of a portfolio of attacks might be the sum of risks over the entire attack space. There are several problems with this formula. First, there is the desire to translate the consequences of risk into a measure of expected annual financial loss. This might work well for some; however, many organizations cannot translate their mission into meaningful financial terms. Even private firms would have difficulty estimating the consequences of a successful penetration of their information systems.

Second, obtaining valid estimates of $p(a)$ for a given time frame is very difficult, often impossible. Adversaries do not often act in predictable ways and will change their behavior, based on changes to the information system’s security.

A third problem is that data for $p(s|a)$ is often historical in nature and biased to attacks that were detected and reported. Ideally, the data for $p(s|a)$ should be forward-looking to address future concerns, instead of reporting past problems that most likely have already been mitigated.

A fourth problem is that the formula above is the *expected* risk. In an expected risk formulation,

rare events that have catastrophic consequences are considered equal in risk to common events that have relatively small consequences. Haimes states that “the expected value of risk fails to represent a measure that truly communicates the manager’s or decision maker’s intentions and perceptions [17]”. In the author’s experience, many decision makers do not agree with expected value risk decisions. This is especially true in military and government organizations. These organizations do not follow expected value decision making, because the impact of a low probability but high consequence computer attack could cripple a nation’s defense or economy. Also, a successful, undetected attack against an information system can have long-term consequences since an adversary can cover his/her tracks and possibly create multiple hidden entry and exfiltration routes inside the system. A quantitative method that was created to mitigate these three problems is a MODA approach [18–20].

Haimes created a method to mitigate the effects of expected value risk analysis called the *partitioned multiobjective risk method* (PMRM) [21]. In PMRM risk is measured, given that the damage falls within specific ranges of probabilities. Instead of just providing the expected value of risk, PMRM allows the analyst to show many ranges of risk, such as “low severity, high exceedance probability; moderate severity, medium exceedance probability; and high severity, low exceedance probability”.

Multiple Objective Decision Analysis (MODA) Risk Analysis. MODA risk assessments, such as the mission-oriented risk and design analysis (MORDA) method [22], were created because of the difficulties in collecting data and assessing risk in a purely probabilistic manner. Instead of providing probabilities of attempt, MORDA assessments seek to characterize attacks in terms of their value [12] or impact to both the attacker and the defender. System security engineers and adversary experts work together to define what makes an attack valuable to an adversary considering measures such as probability of detection, consequences of detection, probability of success given detection, resources expended, probability of attribution, consequences of attribution and payoff given a successful attack. Attacks are scored on these value measures using an interval or ratio scale, which are then translated into attack values. The total attack value to the

adversary is calculated using a weighted sum of the attack measures. If the model meets the required independence assumptions and the modeler uses ratio weights [23], the value is measured using a simple weighted average as shown in the equation (2).

$$v(x) = \sum_{i=1}^n w_i v_i(x_i) \quad (2)$$

where $v(x)$ is the value of the attack to the adversary, $i = 1$ to n is the number of the value measure, e.g., “payoff” or “resources expended”, x_i is the score of the attack on the i th value measure, e.g., “high” or “low” payoff, $v_i(x_i)$ is the single dimensional value of a score of x_i , e.g., “high” payoff has a value of “1”, w_i is the weight of the i th value measure, e.g., weight for payoff = 0.25, and $\sum_{i=1}^n w_i = 1$ (all weights sum to 1).

With either the probabilistic risk model or the MORDA model, risk to an information system can be analyzed and reported in a traceable and understandable manner. More importantly, these models can be adapted for use as a design tool for assisting with secure system design and cost–benefit trade studies (*see Reliability Optimization; Clinical Dose–Response Assessment*).

Semiquantitative Risk Analysis

Semiquantitative risk analysis for information systems is a widely used method to simplify the risk assessment process using easy to understand scales and mathematics. A commonly used model is a variant of the probabilistic risk calculation discussed earlier. However, instead of probabilities, ordinal or interval scales are used to measure threat, vulnerability, and impact. For example, these three characteristics might be measured on a 1–100 scale. The risk to the system is determined by taking the product of the three risk scores as shown in equation (3).

$$Risk = threat \times vulnerability \times impact \quad (3)$$

This technique has one major benefit – it is very simple because it allows for data collection using natural language. However, there is great debate as to the meaning of the resulting number and the validity of the calculation. Often semiquantitative risk assessments use interval scales. Multiplication and division

on the individual measures using interval scales is not permitted [24]. MODA risk assessments avoid this issue because interval scales can be multiplied by a constant, which is MODA's ratio weight. Semiquantitative risk methods that perform mathematical operations on ordinal scales are of great concern. Risk assessments using ordinal scales cannot even use addition and subtraction. In one study, risk calculations using ordinal scales created an error of over 600% in the model [25]. Semiquantitative risk assessments should be avoided on account of these concerns.

Summary

In this paper, we have discussed the advantages and disadvantages of quantitative risk assessments for information systems. We have also discussed the stakeholders, basic decision support concepts, decision support processes, and measurement of risk. There are great benefits in performing a quantitative risk assessment. Mathematical techniques can account for the multiple, and often conflicting desires of many stakeholders and provide the certifiers and accreditors with the necessary data to make complex design trade-off decisions. These decisions can be made through a rational, traceable, and understandable process. Semiquantitative risk assessments have mathematical problems that prohibit their reliable use for risk assessments.

References

- [1] Sage, A. & Armstrong, J. (2000). *Introduction to Systems Engineering*, John Wiley & Sons, New York.
- [2] Parnell, G.S., Driscoll, P.J. & Henderson, D.L. (eds) (2006). *Systems Decision Making for Systems Engineering and Management*, Fall 2006 Edition, John Wiley & Sons.
- [3] Beauregard, J., Deckro, R. & Chambal, S. (2002). Modeling information assurance: an application, *Military Operations Research* 7(4), 35–55.
- [4] Buckshaw, D., Parnell, G., Unkenholz, W., Parks, D., Wallner, J. & Saydjari, O. (2005). Mission-oriented risk and design analysis of critical information systems, *Military Operations Research* 2(10), 19–38.
- [5] Butler, S. (2002). Security attribute evaluation method: a cost-benefit approach, *Proceedings of the 24th International Conference of Software Engineering*, Orlando.
- [6] Hamill, T. (2000). *Modeling Information Assurance: A Value Focused Thinking Approach*, MS Thesis, AFIT/GOR/ENS/00M-15, Air Force Institute of Technology, Wright-Patterson AFB, OH.
- [7] Hamill, T., Deckro, R., Kloeber, J. & Kelso, T. (2002). Risk management and the value of information in a defense computer system, *Military Operations Research* 7(2), 61–82.
- [8] Dillon-Merrill, R., Parnell, G., Buckshaw, D., Hensley, W. & Caswell, D. (2006). Avoiding common pitfalls in decision support frameworks for department of defense analyses, *Military Operations Research Society Journal* submitted for publication in.
- [9] Stevens, S. (1946). On the theory of scales of measurement, *Science* 103, 677–680.
- [10] Pariseau, R. & Oswalt, I. (1994). Using data types and scales for analysis and decision making, *Acquisition Review Quarterly* 1, 145–159.
- [11] Watson, S. & Buede, D. (1987). *Decision Synthesis: The Principles and Practice of Decision Analysis*, Cambridge University Press, Cambridge.
- [12] Keeney, R. (1992). *Value-Focused Thinking: A Path to Creative Decisionmaking*, Harvard University Press, Cambridge.
- [13] Crucitti, P., Latora, V. & Marchiori, M. (2004). Model for cascading failures in complex networks, *Physical Review E* 69, 045–104.
- [14] Taylor, L. (2004). *Security Certification and Accreditation 101* Security Certification and Accreditation 101, http://www.intranetjournal.com/articles/200406/pij_06_23_04a.html.
- [15] Lowrance, W. (1976). *Of Acceptable Risk*, William Kaufmann, Los Altos.
- [16] Smith, G., Scouras, J. & DeBell, R. (2007). Qualitative representation of risk, *Wiley Handbook of Science and Technology for Homeland Security*, Wiley Handbook of Science and Technology for Homeland Security, New Jersey, John Wiley & Sons, New York.
- [17] Haimes, Y. (1998). *Risk Modeling, Assessment, and Management*, John Wiley & Sons, New York.
- [18] Keeney, R. & Raiffa, H. (1976). *Decision Making with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley & Sons, New York.
- [19] Kirkwood, C. (1997). *Strategic Decision Making: Multi-objective Decision Analysis with Spreadsheets*, Belmont, Duxbury Press, California.
- [20] Parnell, G. (2007). Value-focused thinking, *Methods for Conducting Military Operational Analysis: Best Practices in Use Throughout the Department of Defense*, Military Operations Research Society and the LMI Research Institute, Chapter 19.
- [21] Haimes, Y. (1998). *Risk Modeling, Assessment, and Management*, John Wiley & Sons, New York.
- [22] Buckshaw, D., Parnell, G., Unkenholz, W., Parks, D., Wallner, J. & Saydjari, O. (2005). Mission-oriented risk and design analysis of critical information systems, *Military Operations Research* 2(10), 19–38.

- [23] Keeney, R. & Raiffa, H. (1976). *Decision Making with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley & Sons, New York.
- [24] Warren, L. (2004). *Uncertainties in the Analytic Hierarchy Process*, Australia Defence Science and Technical Organisation Technical Note, DSTO-TN-0597.
- [25] Conrow, E. (2003). *Effective Risk Management: Some Keys to Success*, 2nd Edition, American Institute of Aeronautics and Astronautics.

DONALD L. BUCKSHAW

Utility Function

Suppose you are asked to choose between two projects as in Figure 1.

This is a difficult choice as we must consider both probabilities and associated outcomes to identify the preferred project.

Utility function and expected utility theory [1] aid a decision maker in making choices among projects (alternatives or options) whose outcomes are uncertain at the time of decision making. The basic idea is to use the utility function of the decision maker to convert outcomes into utilities and then to compute expected utility for each project. The project with the highest expected utility is then chosen. We will answer the following questions:

1. What is the justification for using expected utility as a criterion for choosing among alternatives?
2. How do we assess utility function?
3. What are some behavioral properties of utility function?

Expected Utility

Consider a set of outcomes or consequences, X . Let P be the set of probability distributions on X . Let us assume X is finite or P is the set of all simple probability measures on X . An element of P is denoted by lower case letters p, q, r , etc. and that of X by x, y, z , etc. In our example, project A is represented by p and project B by q . Let us use the notation $\succsim$ to denote “preferred or indifferent to”. We need a concept of compound lottery, $\alpha p + (1 - \alpha)q$, which also belongs to P and denotes that the lottery returns p with chance α and returns q with chance $(1 - \alpha)$, $\alpha \in [0, 1]$.

The following axioms, which address the decision maker’s preferences for lotteries in P , justify the use of expected utility as a criterion for choosing among lotteries in P .

Axiom 1 For all $p, q, r \in P$, (a) $p \succsim q$ or $q \succsim p$ and (b) if $p \succsim q$ and $q \succsim r$, then $p \succsim r$. The first part of this axiom shows completeness and the second part shows transitivity.

Axiom 2 For all $p, q, r \in P$ and $\alpha \in [0, 1]$, $p \succsim q$ implies that $\alpha p + (1 - \alpha)r \succsim \alpha q + (1 - \alpha)r$. This is

the key axiom of expected utility theory and requires that if one prefers p over q , then one should prefer the compound lottery with α chance of yielding p and $(1 - \alpha)$ chance of yielding r to the compound lottery with α chance of yielding q and $(1 - \alpha)$ chance of yielding r . Since the common outcome r is obtained with common probability $(1 - \alpha)$ in both compound lotteries, the choice between two such compound lotteries is determined only by the preference between p and q . This axiom is called the substitution or independence axiom. Normatively, this axiom is very appealing, but people often violate it and some extensions of expected utility theory are based on relaxing this axiom.

Axiom 3 For all $p, q, r \in P$, if $p \succsim q \succsim r$, then there exists $\alpha \in [0, 1]$ such that

$$\alpha p + (1 - \alpha)r \sim q \quad (1)$$

This is called the continuity or Archimedean axiom. This axiom ensures that every intermediate lottery can be made equivalent (in preference) to a compound lottery between the more desirable and less desirable lotteries by suitably choosing α . The above three axioms lead to a remarkable result.

Theorem 1 A binary relation $\succsim$ on P satisfies Axioms 1–3 if and only if there exists a function $u : X \rightarrow \mathbb{R}$ such that $p \succsim q$ if and only if:

$$\sum_{x \in X} u(x)p(x) \geq \sum_{x \in X} u(x)q(x) \quad (2)$$

Further, u is unique up to a positive linear transformation. In our example, project A is preferred or indifferent to project B if and only if $0.35u(42) + 0.23u(209) + 0.42u(611) \geq 0.61u(153) + 0.39u(502)$. The choice between complicated lotteries is reduced to a simple calculation if we can somehow specify the utility function, u .

Assessing Utility Function

We now discuss two methods for eliciting a utility function from a decision maker. Broadly speaking, a utility function reflects the risk attitude of a decision maker and therefore will be unique to each individual. Consider monetary payoffs in the range x_0 to x^* , where x_0 is the worst possible payoff and x^* is the best possible payoff in a given decision context.

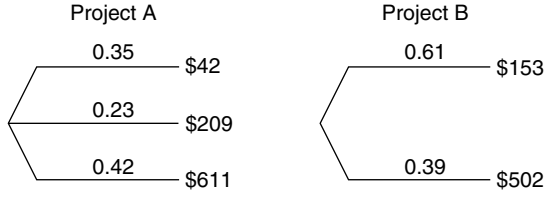


Figure 1 A sample probability choice problem

Certainty Equivalent Method

The certainty equivalent of a lottery p is the amount, x_p , such that a decision maker is indifferent between receiving the guaranteed payoff x_p or the lottery p . We denote $x_{0.5}$ as the certainty equivalent of a lottery with $(x^*, 0.5; x_0, 0.5)$, which is a lottery that yields x^* with a 0.5 chance and x_0 with a 0.5 chance.

We set $u(x^*) = 1$ and $u(x_0) = 0$. Therefore, $u(x_{0.5}) = 0.5u(x^*) + 0.5u(x_0) = 0.5$ by the expected utility rule. Now, we have three points on a utility curve. To obtain $u(x_{0.75})$, we use a clever trick. We know that $(x^*, 0.75; x_0, 0.25) \sim (x^*, 0.5; x_{0.5}, 0.5)$ as $0.75u(x^*) + 0.25u(x_0) = 0.5u(x^*) + 0.5u(x_{0.5})$. Therefore, we simply seek the certainty equivalent of a 50–50 lottery between x^* and $x_{0.5}$ and this certainty equivalent provides us with $x_{0.75}$. Similarly, we seek $x_{0.25}$ by eliciting the certainty equivalent of the lottery $(x_{0.5}, 0.5; x_0, 0.5)$.

We now have utilities of five outcomes – $u(x_0) = 0$; $u(x_{0.25}) = 0.25$; $u(x_{0.5}) = 0.5$; $u(x_{0.75}) = 0.75$; $u(x^*) = 1$ – in the desired range $[x_0, x^*]$. We can now find a curve through these points. Notice that the certainty equivalent of a 50–50 lottery between two outcomes is easier to obtain than a direct choice between two complicated lotteries, such as projects A and B in our example. Thus, by seeking answers to some simple questions, we can calculate the preference for more complicated lotteries. We cannot do away with the preferences entirely, but by using expected utility theory we can greatly simplify the choice process.

Probability Equivalent Method

For the certainty equivalent method using 50–50 lotteries between two outcomes, we had essentially fixed the utility levels at 0, 0.25, 0.5, 0.75, and 1 and sought the sure outcomes that had those utilities. For the probability equivalent method, we fix the outcomes $x_0, x_1, x_2, x_3, \dots, x^*$ and seek utilities for

each x_i . Thus, if outcomes are on the x axis and utilities are on the y axis, then the certainty equivalent method fixes some points on the y axis and seeks corresponding points on the x axis. The probability equivalent method fixes points on the x axis and seeks corresponding points on the y axis. To obtain the utility of outcome x_i we simply seek p_i such that $x_i \sim (x^*, p_i; x_0, 1 - p_i)$. Now by the expected utility rule $u(x_i) = p_i$. Depending on the accuracy desired, we elicit utilities for several points and then plot a curve through these points.

Behavioral Properties

Suppose you are offered a choice between a lottery p and a sure outcome equal to the expected monetary value of lottery p . If you prefer the expected monetary value to the lottery and do so for all lotteries defined over the range of interest $[x_0, x^*]$, then you are *risk averse* in the range of interest. An immediate implication of risk aversion is that your utility function will be concave. For a risk averse decision maker, knowledge of the utility function at a few assessed points restricts the possible utility values at other intermediate points (see [2]).

A special case of risk aversion called *constant risk aversion* is often used in applications. Suppose a strictly risk averse decision maker is indifferent between receiving a sure outcome x or a lottery p . Now let us add a constant amount Δ to all outcomes of lottery p . If the decision maker is indifferent between the modified lottery $p + \Delta$ and the sure outcome $x + \Delta$ for all Δ , then the decision maker is constantly risk averse within the relevant range of outcomes. The only utility function appropriate for this hypothetical person is exponential. Suppose $x_{0.5} \sim (x^*, 0.5; x_0, 0.5)$, then this one judgment is sufficient to specify the utility function for the constantly risk averse decision maker. Suppose, $x^* = 100$, $x_0 = 0$ and $x_{0.5} = 40$; then,

$$u(x) = \frac{1 - e^{-0.8223x/100}}{1 - e^{-0.8223}} \quad (3)$$

Note that at $x = 40$, $u(40) = 0.5$ as desired.

Suppose u is twice continuously differentiable. Let u' be the first derivative of u and u'' be the second derivative of u . Note that $u' > 0$ if u is strictly increasing (more money is preferred to less) and $u'' \leq 0$ if u is concave (or < 0 if u is strictly

concave). Pratt [3] defined a measure of risk aversion as follows:

$$R(x) = -\frac{u''(x)}{u'(x)} \quad (4)$$

For the constant risk aversion case in which the utility function is exponential, $R(x)$ is a constant. For decreasing risk aversion, relative risk aversion, and comparison of risk aversion for two decision makers, see [3].

The risk aversion property though useful in analysis and applications is not universally accepted as an accurate description of people's behavior. Kahneman and Tversky [4] have demonstrated that people are, in fact, risk seeking for losses, as they do not want to accept a sure loss. Thus, people will not accept a sure loss of \$45 and would rather play a lottery in which they may lose \$100 with 0.5 chance, but with an equal chance to lose nothing (\$0). Risk aversion also does not hold for lotteries with small probabilities of very large payoffs.

In summary, if a decision maker subscribes to Axioms 1–3, then he can use maximization of expected utility as a criterion to evaluate choices under uncertainty. The utility function can be assessed by the certainty equivalent or probability equivalent methods. The assessment of utility function is greatly simplified if some further assumptions such as constant risk aversion are made about decision maker's preferences.

Challenges to Expected Utility

Allais [5] provided perhaps the best known challenge to expected utility. Allais's dramatic example involved large sums of money and extreme probabilities. We provide an example from [4] with moderate outcomes and reasonable probabilities.

Situation 1. Which of the following options do you prefer?

- A. a sure gain of \$30; or
- B. an 80% chance to win \$45 and a 20% chance to win nothing.

Situation 2. Which of the following options do you prefer?

- C. a 25% chance to win \$30 and a 75% chance to win nothing; or
- D. a 20% chance to win \$30 and an 80% chance to win nothing.

The model preference in experimental setting is: $A \succ B$ and $D \succ C$. Such a preference pattern is inconsistent with the expected utility model. To see this observe that $A \succ B$ implies that $u(30) > 0.8u(45) + 0.2u(0)$. However, $D \succ C$ implies that $0.25u(30) + 0.75u(0) < 0.2u(45) + 0.8u(0)$, or $u(30) < 0.8u(45) + 0.2u(0)$.

Since $u(30)$ cannot simultaneously be greater as well as less than the number $0.8u(45) + 0.2u(0)$, there can be no utility assignment that will permit such a preference pattern.

The above example represents a violation of the substitution principle (Axiom 2). To see this, we define option A as p and option B as q . Let α be 0.25 and r be 0. By Axiom 2, $p \succsim q$ if and only if $0.25p + 0.75r \succsim 0.25q + 0.75r$, which is the same as requiring $A \succsim B$ if and only if $C \succsim D$.

There has been a debate in the literature as to whether a rational person would change his mind upon introspection and modify his preferences to be consistent with the expected utility model. It can, however, be asserted that unaided choices of even reasonable and well-trained people do not satisfy axioms of expected utility. Thus, as a descriptive model of behavior, the expected utility model needs modification. Most, if not all, scholars agree that Axioms 1–3 of expected utility theory are normatively compelling. The expected utility model has been used in numerous applications and provides a foundation for decision and risk analysis.

References

- [1] von Neumann, J. & Morgenstern, O. (1953). *Theory of Games and Economic Behavior*, 2nd Edition, John Wiley & Sons, New York.
- [2] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley & Sons, New York.
- [3] Pratt, J.W. (1964). Risk aversion in the small and in the large, *Econometrica* **32**, 122–136.
- [4] Kahneman, D. & Tversky, A. (1979). Prospect theory: an analysis of decision under risk, *Econometrica* **47**, 263–291.
- [5] Allais, M. (1953). Le comportement de l'homme rationnel devant de risqué: critique des postulats et axiomes de l'école Americaine, *Econometrica* **21**, 503–546.

Related Articles

Subjective Expected Utility

Value at Risk (VaR) and Risk Measures

Generally speaking, risk is the possibility of missing an organization's objectives because of external events or internal actions. Hence, risk is strongly related to uncertainty and randomness. This is the reason why quantitative risk management mainly derives from mathematical stochastics. A risk measure is a function that quantifies the size of risk contained in a business unit, company, or trading position. The main reason for using risk measures is to support business decisions. An important figure, therefore, is the risk-return ratio that describes how many additional units of risk we have to take to get one additional unit of return. Another application is to determine how many capital reserves a company needs to stay solvent in case of an occurring risk. This especially affects banks and insurance companies for which taking risk is, in fact, their main business.

In this essay, we discuss risk measures with different applications. In the section titled "Traditional Risk Measures", we present traditional risk measures like volatility, shortfall probability (SP) (*see Risk Measures and Economic Capital for (Re-)insurers; Extreme Value Theory in Finance*), and other lower partial moments (LPMs) (*see Axiomatic Models of Perceived Risk; Axiomatic Measures of Risk and Risk-Value Models*) that are used since the mid-twentieth century. In the section titled "Value at Risk", we discuss the value at risk (VaR), which is, at the moment, used to control the market risk of almost every major financial institution in the world. The section titled "Coherent Measures of Risk" describes some mathematical requirements that a so-called coherent risk measure should fulfill and how the presented risk measures meet them. Finally, we give a brief summary in the section titled "Summary".

Traditional Risk Measures

In the sequel, we will study a portfolio with a given price P_0 today and a random price P in the future, e.g., in 1 day. Furthermore, we will continuously use compounded returns $R = \ln(P \cdot P_0^{-1})$, e.g., daily

returns if P is the price in 1 day, to measure the performance of the portfolio.

Volatility and Standard Deviation

The first milestone in risk measurement was made by Markowitz [1, 2]. Before his path-breaking publication, the quality of an investment was mainly measured by its expected return $\mu = E[R]$. Markowitz proposed to use the variance $\sigma^2 = \text{Var}[R]$ to quantify the risk of an investment. Every suitable portfolio should lie on the so-called efficient frontier (*see Asset-Liability Management for Nonlife Insurers*), where no higher return can be generated at the same level of risk. For a further characterization of a probability distribution, we use the concept of central moments. The k th central moment of the random variable R is defined by

$$m_k = E[(R - \mu)^k] \quad (1)$$

For $k = 2$ this is the variance, its square root, the so-called standard deviation or volatility σ measures how widespread the possible returns are. Figure 1 shows the charts of two important indices – the S&P 500 Total Return Index and the JP Morgan Global Government Bond Unhedged USD Index – in the period from January 01, 1988 until December 31, 2005. For the ease of comparison, the starting values at January 01, 1988 are set to 100. We use this data sample in the whole article to calculate the corresponding daily returns and different risk measures.

It is obvious that overall the S&P 500 not only showed a better performance than the bond index, but also had some bad years. On the other hand, the bond index never fell over a long period and shows much less fluctuation. This can be seen in Table 1 showing not only a higher than expected return but also a higher volatility for the stocks. In summary, mean-variance analysis of risky variables is easily calculated and can be very useful to compare different investment alternatives.

However, this method also has disadvantages: variance includes risk and chance to the same degree and thus does not distinguish between gains and losses. Owing to this fact, variance is only appropriate for symmetric distributions. In practice, however, many return distributions are asymmetric. An appropriate measure to check for an asymmetric return

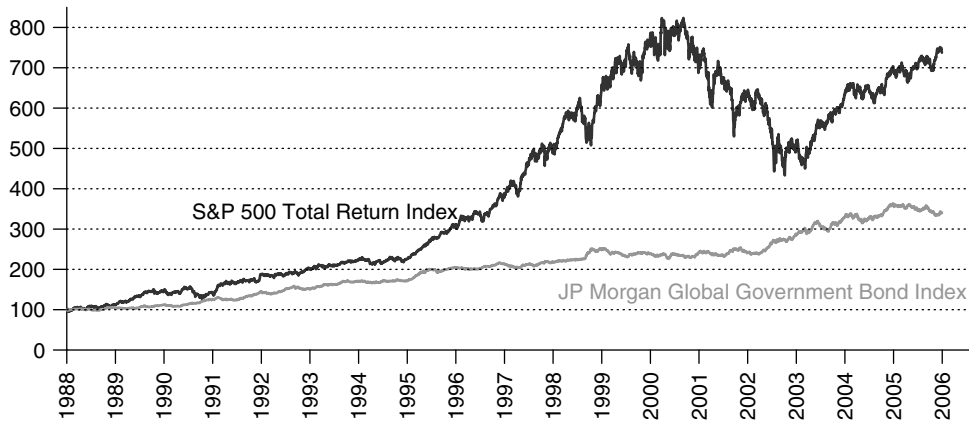


Figure 1 Index charts (January 1, 1988 = 100)

Table 1 Risk measures for sample data (daily returns) based on a portfolio price of $P = 100$

	S&P 500 Total Return		JP Morgan Global Government Bond Unhedged USD, Global	
		p. a.		p. a.
Mean return (%)	0.045	11.105	0.028	6.798
Standard deviation (%)	1.011	15.889	0.376	5.912
Skewness	-0.171	—	0.0326	—
Excess kurtosis	3.962	—	1.771	—
Semivariance	0.004	—	0.001	—
VaR (99%)	2.659	—	0.978	—
VaR (95%)	1.604	—	0.593	—
CVaR (99%)	3.576	—	1.178	—
CVaR (95%)	2.314	—	0.820	—

distribution is called *skewness* and is the third standardized moment of the return distribution, i.e.,

$$s = E \left[\left(\frac{R - \mu}{\sigma} \right)^3 \right] \quad (2)$$

By using an uneven exponent, we keep the algebraic sign of each standardized return leading to a skewness of zero for perfect symmetric distributions. Figure 2 shows the daily returns of the indices in our sample. The S&P 500 has a negative skewness, while the bond index has a smaller positive one. One reason for this is that there are extraordinary high share losses because of natural disasters, political events, or just psychological effects, but rarely gains with the

same dimension. On the other hand, bonds can benefit from panic selling on stock markets. For example, on the first trading day in New York after the terrorist attack on September 11, 2001, the S&P 500 lost over 5% in a single day while the bond index increased by more than 2%.

Another key figure when analyzing market risk is the kurtosis that is similarly calculated as the skewness in equation (2) but with the power of four. The kurtosis gives information about the shape of the return distribution. Most often, we want to compare a variable with the normal distribution that has a kurtosis of three. Therefore we subtract this number from the kurtosis to get the excess kurtosis or just excess. While a negative excess means a low peak with wider “shoulders” (platycurtic), a positive

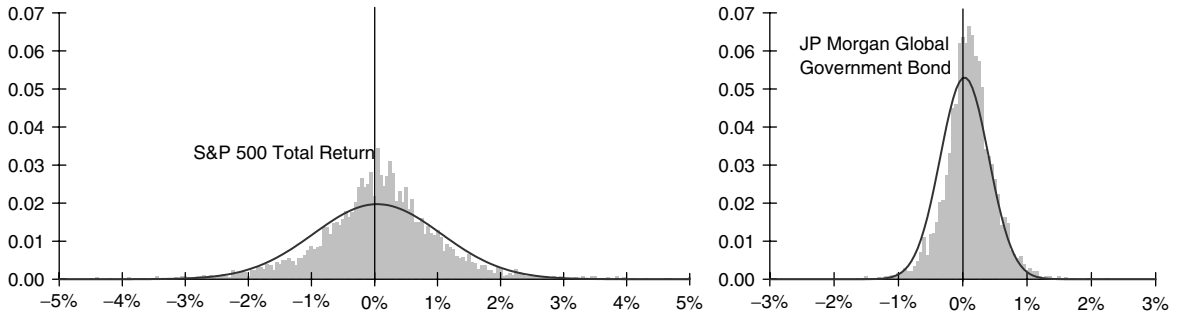


Figure 2 Histogram of daily returns (January 1, 1988 to December 31, 2005)

value means a narrow high peak and fatter “tails” (leptokurtic). The returns of financial markets are often found to be leptokurtic. Skewness and kurtosis are mainly used to check if an unknown random variable is normally distributed.

Lower Partial Moments – Shortfall Probability and Expected Shortfall

We recognized that there are distributions that are not symmetric and therefore we need a measure that examines only the downside risk of the return distribution, without regarding the chance. For this reason, we discuss a concept that only considers the lower part of the return distribution, the LPMs. They are similar to the central moments in equation (1) but only consider the downside deviation from a given barrier, benchmark return, or target b :

$$LPM_k(R, b) = E[1_{(-\infty, b)}(R) \cdot (b - R)^k] \text{ with} \quad (3)$$

$$1_{(-\infty, b)}(R) = \begin{cases} 1 & \text{if } R < b \\ 0 & \text{else} \end{cases}$$

For $k = 0$ we get the so-called shortfall probability. This simple risk measure was first proposed by Roy [3]. $SP(R, b)$ is the likelihood for the return to fall short of the barrier b and has big relevance in situations where the occurrence of a certain event has an extraordinary effect, e.g., the likeliness of not being able to fulfill contractual duties like the payment of a guaranteed return. From our sample we learn that there has been no trading day where the bond index lost more than 2%. On the other hand, the stock index lost even more than 5% in about 1 day out of thousand.

However, the SP contains no information on the amount of loss in case of missing a target. To do

this, we can use the so-called expected shortfall (ES), which shows the expected amount of loss, given that we fall short of the benchmark b . In mathematical terms we have $ES(R, b) = -E[R|R < b]$ given that $SP(R, b) > 0$. Expected shortfall and LPM are connected by the formula

$$ES(R, b) = SP(R, b)^{-1} \cdot LPM_1(R, b) - b \quad (4)$$

Applied to our example, we recognize that in the few cases where the S&P 500 lost more than 5% in a single day, its expected loss is about 6.3% (see Table 2). That means if there is a crash, it can go down even much further than only the 5%.

Another interesting risk measure is the lower semivariance given as $LPM_2(R, \mu)$. It is calculated exactly like the variance but only considers returns below the mean. Its applications are similar to those of the variance but the problem of asymmetric return distributions is reduced. We get an idea of the size of the return variation below average. However, the meanings of SP and ES are more intuitive and therefore these measures are often preferred. Another reason for this is the fact that these concepts are strongly related to the most famous risk measure – the VaR – which is discussed in the following chapter.

Value at Risk

In the financial sector, managing extreme events plays a major role. Banks and insurance companies can lose immense amounts of money in a single day because of a market crash or natural disaster. Managing extreme risks means building capital reserves to stay solvent in the event of damage. In the banking industry, there exist strong regulations to avoid “bank

Table 2 Downside risk measures for sample data (daily returns)

b (%)	S&P 500 Total Return				JP Morgan Global Government Bond Unhedged USD, Global			
	0.00	−1.00	−2.00	−5.00	0.00	−1.00	−2.00	−5.00
SP (%)	46.152	11.611	2.633	0.113	46.602	0.878	0.000	0.000
LPM_1 (%)	0.340	0.083	0.021	0.002	0.127	0.002	0.000	0.000
ES (%)	0.736	1.718	2.787	6.291	0.272	1.207	–	–

runs” that would destabilize the whole economy. VaR (see **Large Insurance Losses Distributions; Credit Scoring via Altman Z-Score**) is the most widely used measure to quantify extreme risks. We discuss its definition in the first part of this chapter. The second part shows different methods to calculate the VaR. In the last section, we discuss the relations between VaR, the newer concept of conditional value at risk (CVaR), and the traditional concepts of shortfall probability and expected shortfall.

Definition

In the early 1990s, of the last century, financial derivatives became more and more important raising the need for a proper risk management. At the same time the top management requested a 1-day one-page summary of the overall risk position to control the business more properly. The solution was introduced by JP Morgan in 1994 (see e.g. [4] for more details) developing the concept of VaR. Under the name RiskMetrics, the idea was distributed and became standard for almost every major bank in the world. The VaR is based upon the value of a portfolio measured in market prices. It regards the maximum loss that can occur within a certain time horizon Δt (e.g., 1 or 10 business days) with a given confidence level $1 - \alpha$ (e.g., 95 or 99%). The other way round, we experience a loss bigger than the VaR within that period only with a probability of α . Hence, VaR is no more than a quantile of the portfolio return distribution multiplied by the portfolio price. For the ease of exposition we set the portfolio price to 100 in our numerical example. There are several approaches to determine VaR in practice and we will discuss the most common in the next section.

Implementation Methods

All calculation methods for the VaR are based on the distribution of the portfolio return. This return

may be driven by several risk factors, e.g., interest rates, market indices, stock prices, or currencies, and obviously depends on the portfolio weights. The different methods mainly differ by the required knowledge about the portfolio return distribution (see e.g. [5] for more details).

Historical Method. The historical method requires empirical data for the portfolio price $P_t, t = 0, \dots, T$. As already mentioned, these empirical prices may be calculated on the basis of time series of several risk factors. We derive the time series for the portfolio return by

$$R_t = \ln \frac{P_t}{P_{t-1}}, \quad t = 1, \dots, T \quad (5)$$

We get an estimate of the expected return μ and the variance σ^2 per period by

$$\mu \approx \frac{1}{T} \cdot \sum_{t=1}^T R_t \quad (6)$$

and

$$\sigma^2 \approx \frac{1}{T-1} \cdot \sum_{t=1}^T (R_t - \mu)^2 \quad (7)$$

Furthermore, we can derive the frequency distribution F_R of the return in the observed time horizon Δt . In the case that Δt is equal to the frequency in which the empirical data is collected, we get the empirical distribution by simply counting the number of returns smaller or equal than x , i.e.,

$$F_R(x) \approx \frac{1}{T} \cdot \sum_{t|R_t \leq x} 1 \quad (8)$$

In any other case, we first need to transform the time horizon in an appropriate way. Now we can

calculate the maximum loss that might happen in Δt with a certain confidence level α , the VaR per dollar portfolio value (VaR_R) as follows:

$$VaR_R(\alpha) = -\sup\{x | F_R(x) < \alpha\} \quad (9)$$

Consequently, the portfolio VaR is calculated as the VaR per dollar portfolio value multiplied by the portfolio value P , i.e.,

$$VaR(\alpha) = VaR_R(\alpha) \cdot P \quad (10)$$

Note that we always get a positive VaR for a fairly small α . If we let $c_R(\alpha)$ denote the α -quantile of the standardized distribution of the portfolio return, we derive

$$VaR_R(\alpha) = -(\mu + c_R(\alpha) \cdot \sigma) \quad (11)$$

and

$$VaR(\alpha) = -(\mu + c_R(\alpha) \cdot \sigma) \cdot P \quad (12)$$

Now, we use our example again and identify the VaR for mixed portfolios consisting of equity and bonds (Figure 3). As already mentioned, we set $P = 100$ for the ease of exposition.

We see that the portfolio with the lowest risk mainly consists of bonds. The historical method is easy to understand and intuitive but needs some preconditions. First, we need the required data, which is no problem for major indices, but can be difficult for small or illiquid assets and is almost impossible for new ones. Furthermore, we

implicitly assume that we can predict the future only by looking into the past. However, over time, the environmental conditions change and therefore the chosen time period must not be too long, but needs to be a representative sample. In comparison to the historical method, the Monte Carlo simulation (*see Nonlife Loss Reserving; Simulation in Risk Management; Reliability Optimization*) approach uses stochastic processes to generate a multiple of portfolio returns based on statistical or forecasted parameters and may help to overcome the problem of being limited to one historical scenario.

Variance–Covariance Approach. A rather simple and very common method to calculate VaR is the variance–covariance approach, also known as *Risk-Metrics approach* [6]. This method to calculate VaR assumes that the returns of all assets are normally distributed with zero mean and the returns of all assets R_i , $i = 1, \dots, N$ add up linearly to the portfolio return, i.e.,

$$R(\varphi) = \sum_{i=1}^N \varphi_i R_i \text{ with } \varphi_i \text{ denoting}$$

the weight of asset i in the portfolio. (13)

Note that this formula only applies for simple returns and is not true in general for continuously compounded returns. However, if the returns are small, simple, and continuously compounded returns

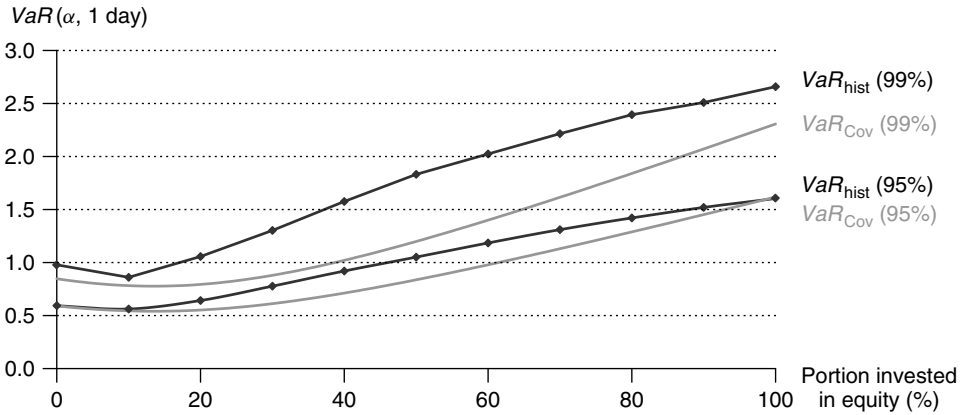


Figure 3 One day VaR with historical method and variance covariance approach based on a portfolio price of $P = 100$ and daily returns

are approximately the same. Furthermore, the variance of the portfolio return is given by

$$\begin{aligned}\sigma^2(\varphi) &= \text{Var}[R(\varphi)] \\ &= \sum_{i=1}^N \sum_{j=1}^N \varphi_i \cdot \varphi_j \cdot \text{Cov}[R_i, R_j]\end{aligned}\quad (14)$$

with $\text{Cov}[R_i, R_j]$ denoting the covariance of R_i and R_j . Using this information we know that the standardized portfolio return is standard normally distributed and we simply get the VaR by

$$\text{VaR}_R(\alpha) = -c(\alpha) \cdot \sigma(\varphi) \quad (15)$$

and

$$\text{VaR}(\alpha) = -c(\alpha) \cdot \sigma(\varphi) \cdot P \quad (16)$$

with $c(\alpha)$ denoting the α -quantile of the standard normal distribution. Since for huge portfolios it can be quite cumbersome to store, estimate, or update all the required data, RiskMetrics proposed a special method to map the returns of the different assets to a given set of risk factors (see [7] for more details). For these, RiskMetrics and specialized data vendors provide volatility and correlation data. The application of the variance–covariance approach is fast and easy and therefore often applied. Furthermore, there also exist extensions of this approach like the delta normal method and the delta-gamma method for assets that are not normally distributed. In our example, there are deviations between the different methods because the returns of our time series are not exactly normally distributed as we have seen in the section titled “Volatility and Standard Deviation”. Therefore, the variance–covariance approach can only give an approximation for the correct VaR. Nevertheless, it is useful because it often is the only technique that is applicable in the required time frame and which can be handled at a reasonable cost.

Further Implementation Methods. Besides the methods described above, there are quite a number of extensions and generalizations, which can be found in the literature. Some focus on the modeling of volatility using methods like exponentially weighted moving averages or generalized autoregressive conditional heteroscedasticity. Others relax the distributional assumptions using, e.g., a student’s t

distribution or normal mixtures. An overview on these methods can be found, e.g., in [8].

VaR and CVaR

For a better understanding of risk measures it is helpful to discuss the relations between the different approaches, e.g., between VaR and shortfall probability. While the VaR determines the loss value for a given probability, the shortfall probability gives the probability for a given loss. In case of a continuous distribution function there is an inverse relation as follows:

$$SP(R, \text{VaR}_R(\alpha)) = \alpha \quad (17)$$

Another very strong relationship exists between the expected shortfall and the concept of CVaR (*see Risk Measures and Economic Capital for (Re-)insurers*). The CVaR determines the expected loss under the condition of falling below the VaR. It is defined by

$$\text{CVaR}(\alpha) = -E[R | R \leq -\text{VaR}_R(\alpha)] \cdot P \quad (18)$$

Hence, according to equation (4),

$$\text{CVaR}(\alpha) = ES(R, -\text{VaR}_R(\alpha)) \cdot P \geq \text{VaR}(\alpha) \quad (19)$$

in case of a continuous distribution function.

Coherent Measures of Risk

As we have seen in the previous chapters, there is no one-and-only risk measure for investment decisions, but there exist many different approaches that can be useful for different situations and markets. We thus should answer the question, which characteristics a good risk measure should have. A list of such characteristics was published by Artzner *et al.* in 1999 who called it the four axioms for a coherent risk measure [9, 10].

The Axioms of Coherence

Let X and Y be two random variables and Δt a given time horizon. X and Y are assumed to indicate positive cash flows, so a higher value is better than a lower. Then, ρ is a coherent risk measure if the following four axioms hold (*see Imprecise Reliability*).

Axiom 1 (Translation Invariance) For every real number a we have

$$\rho(X + a) = \rho(X) - a \quad (20)$$

This axiom is necessary for the interpretation of ρ as risk capital. By adding a deterministic quantity a to the risky position, we reduce the required capital reserve by the same amount.

Axiom 2 (Subadditivity)

$$\rho(X + Y) \leq \rho(X) + \rho(Y) \quad (21)$$

The subadditivity reflects the idea of diversification. If we take several risks, the portfolio risk is smaller or equal than the sum of all partial risks.

Axiom 3 (Positive Homogeneity) For every $\lambda > 0$ we have

$$\rho(\lambda \cdot R) = \lambda \cdot \rho(X) \quad (22)$$

There is no netting or diversification if we change the intensity of a risky business. For example, if we buy two shares of a company, the risk of a decreasing stock price will be twice as high as with only one share. Note that subadditivity and positive homogeneity imply that the risk measure ρ is convex.

Axiom 4 (Monotonicity) If $X \leq Y$ for every possible state of nature then we have

$$\rho(X) \geq \rho(Y) \quad (23)$$

If one business unit has a better outcome than another in every possible case, then obviously this business should have a better risk indicator and therefore needs less capital reserves to compensate the risk.

Review of the Presented Measures of Risk

If we check the risk measures described in the sections titled “Traditional Risk Measures” and “Value at Risk” for coherence we can easily see that variance already fails the translation invariance because of $\text{Var}[X + a] = \text{Var}[X]$. Furthermore, VaR fulfils all axioms except subadditivity (see **Insurance Pricing/Nonlife; Premium Calculation and Insurance Pricing; Behavioral Decision Studies**). For an example see, e.g., Zagst [11, p. 257]. It can also be shown that the SP is not coherent.

However, as is shown in Acerbi and Tasche [12], CVaR is a coherent risk measure for continuous distributions.

Summary

Every company needs realistic but challenging targets to survive in competition. Having targets means taking the risk to fail them. Hence, doing successful business requires taking the right risks and controlling them properly. Therefore, we need an appropriate risk measure to quantify the risk and to compare different decision alternatives. All risk measures derive from mathematical stochastics with its basic concepts – mean, variance, skewness, and kurtosis – which we use to describe the probability distributions for future returns. With the VaR, we can determine the possible loss that will not be exceeded with a given probability and determine the required risk capital. This measure is most commonly used in practice. The implemented methods of calculation mainly depend on the availability and quality of the empirical information. The newer concept of CVaR, which is strongly related to the expected shortfall, is a coherent risk measure that is advantageous in the case of returns that are not normally distributed. All discussed concepts provide a toolbox that can be used for quantifying almost every possible risk. However, every measure has specific requirements and assumptions and therefore it is important to use the right risk measure for the specific application.

References

- [1] Markowitz, H. (1952). Portfolio selection, *Journal of Finance* **7**, 77–91.
- [2] Markowitz, H. (1959). *Portfolio Selection: Efficient Diversification of Investments*, John Wiley & Sons, New York.
- [3] Roy, A.D. (1952). Safety first and the holding of assets, *Econometrica* **20**, 431–449.
- [4] Pézier, J. (2004). Market risk management, in *The Professional Risk Managers' Handbook*, C. Alexander & E. Sheedy, eds, PRMIA Publications, Wilmington, Vol. 3, pp. 43–74.
- [5] Zagst, R. (1997). Value at risk (VaR): Viele Wege führen ans Ziel, *Solutions* **1**, 11–16; **2**, 13–21.
- [6] Jorion, P. (1997). *Value at Risk*, McGraw-Hill, New York.

- [7] Dowd, K. & Rowe, D. (2004). Introduction to value at risk models, in *The Professional Risk Managers' Handbook*, C. Alexander & E. Sheedy, eds, PRMIA Publications, Wilmington, Vol. 3, pp. 95–109.
- [8] Alexander, C. & Sheedy, E. (2004). Advanced value at risk models, in *The Professional Risk Managers' Handbook*, C. Alexander & E. Sheedy, eds, PRMIA Publications, Wilmington, Vol. 3, pp. 115–155.
- [9] Artzner, P., Delbaen, J., Eber, J.-M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**, 203–228.
- [10] Artzner, P., Delbaen, J., Eber, J.-M. & Heath, D. (1997). Thinking coherently, *Risk* **11**, 68–71.
- [11] Zagst, R. (2002). *Interest Rate Management*, Springer Finance, Springer Verlag, Heidelberg.
- [12] Acerbi, C. & Tasche, D. (2002). On the coherence of expected shortfall, *Journal of Banking and Finance* **26**, 1487–1503.

CHRISTIAN SCHMITT AND RUDI ZAGST

Value Function

A value function rank orders a set of objects (alternatives) X . For a finite X , the following two conditions for any $x', x'', x''' \in X$ guarantee the existence of a value function.

1. Completeness: $x' \succsim x''$ or $x'' \succsim x'$.
2. Transitivity: if $x' \succsim x''$ and $x'' \succsim x'''$, then $x' \succsim x'''$.

The first condition requires that a decision maker must be able to compare any two objects and not declare “I do not know.” Aumann [1] relaxes this condition of comparability. The second condition requires that preference including indifference be transitive. Fishburn [2] argues that the transitivity of indifference may not hold.

The above two conditions are necessary and sufficient for a numerical representation of preference $\succsim$; that is, a function $v : X \rightarrow R$ such that for all $x', x'' \in X$, $x' \succsim x''$ if and only if $v(x') \geq v(x'')$.

The function v is commonly called an *ordinal utility function*. We use the term *value function* to clearly distinguish it from the utility function that is defined over the set of lotteries (choice under uncertainty).

Since the value function merely rank orders (including indifference) the objects in X , it is unique up to a monotone increasing transformation; that is, another value function constructed by a monotone increasing transformation of v will also provide the same rank ordering of objects in X .

In many real-world applications, the objects in X are multidimensional. Examples include consumer decisions, facilities location, new product introduction, and investment decisions. Thus $X = X_1 \times X_2 \times \dots \times X_n$, where X_i is the set of outcomes for criterion i . For a new product decision, cost performance, reliability, and time to market entry may be criteria relevant to ranking alternatives. In these relatively complex cases, additional conditions are needed to simplify the form of the value function. The most common form of the value function is additive:

$$v(x_1, x_2, \dots, x_n) = \sum_{i=1}^n v_i(x_i) \quad (1)$$

Krantz *et al.* [3] provide preference conditions that permit decomposition of the multidimensional value

function into the additive form. A key condition for the additive form is preference independence. Simply stated, preference independence implies that the trade-offs between any pair of criteria are unaffected by the fixed levels of the remaining criteria. Thus, the indifference curves for any pair of criteria do not change if the fixed levels of the remaining $(n - 2)$ criteria are changed.

The value function described above provides an ordinal ranking of alternatives or objects in X . We may wish to order the differences in the strength of preference between pairs of alternatives (see [4]). We shall use the term *measurable value function* for a function that permits ordering of preference differences between alternatives. For axiomatic systems that imply the existence of a measurable value function, see [3] and [5].

Dyer and Sarin [6] provide axioms for both additive and multiplicative measurable value functions. The key condition for an additive measurable value function is difference independence. Simply stated difference independence requires that the preference difference between two alternatives that differ only on one criterion does not depend on the common values of the remaining $(n - 1)$ criteria. An interesting result is that, along with some technical conditions, difference independence for any one criterion along with preference independence discussed earlier for all pairs of criteria implies that the additive value function is also measurable.

A value function is elicited from a decision maker's preferences and therefore reflects decision utility. Another concept that is both classical [7] and modern [8] is experienced utility, which reflects pain and pleasure associated with an experience. The distinction between a value function derived from preference and experienced utility is important for both research and application.

Another important extension of a value function is that the carrier of value is gain and loss from a reference point and not the absolute level [9]. In the area of well-being research, an implication of the reference dependence of a value function is that more money may not necessarily buy more happiness. This is because the reference point increases as wealth increases leaving the net value derived from money unchanged. The applications of value function are described in [10].

References

- [1] Aumann, R.J. (1962). Utility theory without the completeness axiom, *Econometrica* **30**(4), 445–462.
- [2] Fishburn, P.C. (1970). Intransitive indifference in preference theory: a survey, *Operations Research* **18**, 207–228.
- [3] Krantz, D.H., Luce, R.P., Suppes, P. & Tversky, A. (1971). *Foundations of Measurement*, Academic Press, New York, Vol. 1.
- [4] Stevens S. (1959). Measurement, psychophysics, and utility, in *Definitions and Theories*, C. Churchman & P. Ratoash, eds, John Wiley & Sons, New York.
- [5] Suppes, P. & Winet, M. (1955). An axiomatization of utility based on the notion of utility differences, *Management Science* **1**, 259–270.
- [6] Dyer, J.S. & Sarin, R.K. (1979). Measurable multi-attribute value function, *Operations Research* **27**(4), 810–822.
- [7] Bentham, J. (1948). *An Introduction to the Principle of Morals and Legislations*, Blackwell Science, Oxford.
- [8] Kahneman, D., Wakker, P.P. & Sarin, R.K. (1997). Back to Bentham? explorations of experienced utility, *The Quarterly Journal of Economics* **112**(2), 375–405.
- [9] Kahneman, D. & Tversky, A. (1979). Prospect theory: an analysis of decision under risk, *Econometrica* **47**, 263–291.
- [10] Keeney, R.L. & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley & Sons, New York.

RAKESH K. SARIN

Volatility Modeling

Financial asset returns are generally highly unpredictable over shorter horizons such as a day or a month. Put differently – the return standard deviation, or the *volatility*, is much larger than the mean expected return in the short run. Moreover, this extreme short-term risk is fundamental in finance theory as it precludes the existence of simple arbitrage trading strategies (*see Statistical Arbitrage*). So theory asserts that we should expect the price process to be fairly volatile. This is embedded in standard asset pricing theory *via* the requirement that the return process constitutes a so-called semimartingale.

Given the dominance of the variance in the distribution of the returns, at least locally, it is natural that much effort has been exerted in measuring and modeling volatility. In particular, research over the last couple of decades has established that return volatilities across all major asset classes are *time varying* and possess highly *persistent* dynamic components. As such, the volatility is to some degree predictable. This also has direct implications for a range of broader issues, including asset and derivatives pricing, portfolio choice, and risk management.

Although there are alternative ways in which to forecast volatility, the focus of this entry is exclusively on the time series modeling of volatility on the basis of observed market prices. This is also the most universally applicable approach as it merely requires historical price data. To set the stage, we formally define the return variance and standard deviation. We assume, as is commonplace, that the return process is sufficient regular (covariance stationary) that these quantities are well defined. For concreteness, we concentrate initially on a single financial return series. We define the (continuously compounded) return between time t and $t + 1$ as follows,

$$\begin{aligned} r_{t+1} &= \log(P_{t+1}) - \log(P_t) \\ &= p_{t+1} - p_t, \quad t = 1, 2, \dots, T \end{aligned} \quad (1)$$

There are $T + 1$ separate price observations available, resulting in T returns defined through the corresponding log-price differentials. Denoting the (unconditional) expectation of any random variable, say X , by $E[X]$, we can write the (unconditional

or average expected) return standard deviation and variance as, respectively,

$$\sigma = \sqrt{E[(r_t - E[r_t])^2]} \quad (2)$$

and

$$\sigma^2 = E[(r_t - E[r_t])^2] \quad (3)$$

The mean, standard deviation, and variance may be estimated consistently by the corresponding sample moments,

$$\begin{aligned} \hat{\mu} &= \frac{1}{T} \sum_{t=1}^T r_t; \\ \hat{\sigma} &= \sqrt{\frac{1}{T} \sum_{t=1}^T (r_t - \hat{\mu})^2}; \\ \hat{\sigma}^2 &= \frac{1}{T} \sum_{t=1}^T (r_t - \hat{\mu})^2 \end{aligned} \quad (4)$$

This refers, of course, only to the long-run values of these moments. The evidence of persistent and time-varying volatility dynamics implies that the return variation is not constant but fluctuates in a partially predictable manner. That is, if we consider the volatility at time t taking into account the recent return fluctuations and all other relevant information, we should instead focus on the *conditional* volatility and variance given as,

$$\sigma_t = \sqrt{E_{t-1}[(r_t - E_{t-1}[r_t])^2]} \quad (5)$$

and

$$\sigma_t^2 = E_{t-1}[(r_t - E_{t-1}[r_t])^2] \quad (6)$$

where $E_{t-1}[X]$ denotes the expected value of X given all information at time $t - 1$.

The subject of volatility modeling is now readily defined. It is dedicated to the generation of sequential forecasts of the conditional volatility, σ_t or σ_t^2 , in real time as new information emerges and asset volatility evolves. Moreover, for short horizons it is typically safe to ignore the (conditional) mean altogether or, for slightly longer horizons, treat it as a constant, μ , because of the semimartingale property of the return series.

Historically Filtered Volatility Estimates and Forecasts

The simplest possible approach to volatility estimation is to exploit the strong persistence in the series and use recent squared returns as an indicator of the current volatility level. This *ad hoc* method is often labeled *historical volatility*. If daily returns are available, the current volatility is typically estimated through a backward looking window of 1–3 months. If the window contains K daily squared returns, the corresponding volatility and variance estimates are,

$$\hat{\sigma}_t = \sqrt{\frac{1}{K} \sum_{j=1}^K r_{t-j}^2}; \quad \hat{\sigma}_t^2 = \frac{1}{K} \sum_{j=1}^K r_{t-j}^2 \quad (7)$$

More generally, however, one may want to estimate volatility through a weighted average of past squared returns allowing for more weight to be assigned to the more recent observations. In addition, this readily avoids the choice of an arbitrary cutoff associated with the selection of a window length. In this case, the general representation of the variance estimator takes the form,

$$\begin{aligned} \hat{\sigma}_t^2 &= \sum_{j=1}^{\infty} w_j r_{t-j}^2; \\ \sum_{j=1}^{\infty} w_j &= 1; \quad w_j \geq 0, \forall j \geq 1 \end{aligned} \quad (8)$$

The historical volatility (variance) estimator is obtained for $w_j = 1/K$ for $1 \leq j \leq K$ and $w_j = 0$ for $j > K$. A more widespread approach is the so-called *exponential smoothing*. This is achieved by letting $w_j = (1 - \lambda)\lambda^{j-1}$ for all $j \geq 1$, and $0 < \lambda < 1$. It produces a sequential updating equation,

$$\begin{aligned} \hat{\sigma}_t^2 &= (1 - \lambda) \sum_{j=1}^{\infty} \lambda^{j-1} r_{t-j}^2 \\ &= \lambda r_{t-1}^2 + (1 - \lambda) \hat{\sigma}_{t-1}^2 \end{aligned} \quad (9)$$

Since only a finite number of past observations are available, there is a need to complement the formula with an initial condition but the impact of this initial condition wears off quite quickly and does not typically pose a practical problem.

A prominent example of exponential smoothing is Risk Metrics (RM) which applies the method for large sets of conditional asset volatilities and, with minor modifications, conditional covariances. RM does not estimate λ but fixes it uniformly across all assets at 0.06. The advantage is robust and sensible measures that are unaffected by idiosyncratic events or outliers. Avoiding the latter is of primary concern for an automated computational approach, which serves to support portfolio choice applications. On the other hand, there is inevitably a degree of specification error present when using a fixed representation across all assets. Moreover, the approach has some troubling and counterfactual implications, mainly concerning longer-term volatility forecasts. If the return residuals are symmetric, one may iterate the conditional one-step-ahead forecasts forward to generate multi-period forecasts. For RM, the k -period-ahead conditional forecast is, with the convention that $\text{Var}_{t-1}[X]$ denotes the conditional variance of X given all information at time $t - 1$,

$$\begin{aligned} \text{Var}_{t-1}(r_{t+k-1} + r_{t+k-2} + \cdots + r_t) &\equiv \sigma_{t:t+k-1|t-1}^2 \\ &= k\sigma_t^2 \end{aligned} \quad (10)$$

Hence, RM implies linear scaling of the variance forecasts, so that the term *structure of volatility* is flat. If current volatility is low the forecasts project this low state into the future without any adjustments reflecting mean reversion toward a long-run level. Likewise, RM will mechanically translate high current volatility into equally high forecasts for the indefinite future. This is clearly counterfactual as the volatility process appears stationary and displays distinct mean reversion at medium and longer horizons. Another weakness of the RM procedure is the absence of a formal metric for assessing model and forecast fit.

GARCH Volatility Modeling

A rigorous approach to volatility estimation, related to the above filtering techniques, is given by the generalized autoregressive conditional heteroskedasticity (GARCH) model of Bollerslev [1], with GARCH(1,1) being particularly successful and popular. It is a parametric procedure relying on statistical inference for model estimation and forecasting. As such, it represents a full-blown model of the conditional return distribution and not just of the second

moments. Hence, it is necessary to introduce the assumptions invoked on the return generating process explicitly. It is stipulated to take the following general form,

$$\begin{aligned} r_t &= \mu_t + \sigma_t z_t; \quad z_t \sim i.i.d.; \\ E(z_t) &= 0; \quad \text{Var}(z_t) = 1 \end{aligned} \quad (11)$$

The specification of the conditional mean is not critical for practical inference regarding the volatility so we simply fix it at zero, but complex mean dynamics can readily be incorporated into the models below. For now, we postpone the discussion of the distributional assumption on the standardized innovation, z_t , but for concreteness one may assume it to be normal.

The only remaining feature of the return generating process is the evolution of the conditional variance. The symmetric GARCH(1,1) model stipulates that,

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (12)$$

where the parameters ω , α , and β all are assumed strictly positive and we again require that an initial condition for the conditional variance at $t = 0$ is provided. One may extend the model by including higher-order lags in either squared returns or the conditional variance, but we focus on GARCH(1,1) for brevity. Repeated substitution in equation (9) readily yields, for $0 < \beta < 1$,

$$\sigma_t^2 = \frac{\omega}{1 - \beta} + \alpha \sum_{j=1}^{\infty} \beta^{j-1} r_{t-j}^2 \quad (13)$$

so, apart from a constant, current volatility is an exponentially weighted moving average of past squared returns. If, in addition, as is almost invariably true, $0 < \alpha + \beta < 1$, then the variance process is covariance stationary with unconditional variance σ^2 . We have,

$$\begin{aligned} \sigma_t^2 &= \sigma^2 + (\alpha + \beta)(\sigma_{t-1}^2 - \sigma^2) \\ &\quad + \alpha \sigma_{t-1}^2 (z_{t-1}^2 - 1) \end{aligned} \quad (14)$$

The last term in this representation has mean zero and is serially uncorrelated so it constitutes a genuine volatility innovation. Hence, this is reminiscent of an AR(1) model for the conditional variance with heteroskedastic errors. The size of $\alpha + \beta$ is a critical determinant of the persistence of the volatility

process which constantly drifts toward σ^2 . Typical values for $\alpha + \beta$ with daily data fall in the interval $[0.92, 0.995]$. This corresponds to a half-life of volatility shocks ranging from about 1 week to 6 months. Moreover, note that α governs the volatility of volatility. Hence, a variety of distinct features of the return-volatility process may be accommodated by different parameter constellations within the basic GARCH model. If $\alpha + \beta \geq 1$, the volatility process is sufficiently ill-behaved and the return distribution sufficiently fat tailed that the second return moment no longer exists. For a range of values beyond unity, the volatility process does remain strictly stationary, but for even higher values of $\alpha + \beta$ the process diverges, indicating that the latter scenario is empirically uninteresting.

As mentioned, one advantage of GARCH is that the implied dynamics secure reversion in volatility to a long-run value, enabling interesting and realistic forecasts. Specifically, the GARCH(1,1) forecast for k steps ahead is,

$$\sigma_{t+k|t}^2 = \sigma^2 + (\alpha + \beta)^{k-1} (\sigma_{t+1}^2 - \sigma^2) \quad (15)$$

Since the daily returns are assumed serially uncorrelated, the variance of the k -day cumulative returns, which provide the basic input to the calculation of the volatility term structure, becomes,

$$\begin{aligned} \sigma_{t:t+k|t}^2 &= k \cdot \sigma^2 + (\sigma_{t+1}^2 - \sigma^2) \\ &\quad \times (1 - (\alpha + \beta)^k)(1 - \alpha - \beta)^{-1} \end{aligned} \quad (16)$$

It is evident from equation (15) that volatility shocks have a persistent effect on the forecasts but the effect decays at a geometric rate governed by $\alpha + \beta$. This contrasts sharply with the RM approach. In fact, RM style exponential smoothing is optimal only if the squared returns follow a “random walk plus noise” model, in which case the minimum mean squared error (MSE) forecast at any horizon is simply the current smoothed value. The evidence blatantly contradicts the random walk scenario, implying that the flat forecast function associated with the RM smoothing is unrealistic and undesirable for volatility modeling and forecasting purposes.

The GARCH(1,1) volatility measurement has other advantages. First, the GARCH parameters, and hence the associated volatility estimates, are obtained using statistical methods that facilitate rigorous inference, thus avoiding *ad hoc* selection of

parameters. Typically, one estimates the GARCH parameter vector $\theta = (\omega, \alpha, \beta)$ by maximizing the log-likelihood function,

$$\log L(\theta; r_T, \dots, r_1) \propto - \sum_{t=1}^T [\log \sigma_t^2(\theta) - \sigma_t^{-2}(\theta) r_t^2] \quad (17)$$

This implicitly assumes that the standardized return innovations are normally distributed, but this is largely a matter of convenience. Even if the conditional return distribution is nonnormal, the associated quasi-maximum likelihood estimates (QMLE) of the parameters remain consistent and asymptotically normal under weak conditions. The optimization of the log-likelihood must be done numerically, but GARCH models are parsimonious and the appropriate estimation procedures have been included in many commercially available software packages.

Second, the dynamics associated with GARCH (1,1) afford intuitive interpretations that readily are generalized to accommodate further realistic features. Without providing details, we mention a few important extensions. One, an asymmetric return-volatility relation may be induced through a direct dependence of the volatility dynamics on the sign of the return innovation as in, e.g., the exponential generalized autoregressive conditional heteroskedasticity (EGARCH) model of Nelson [2]. Two, one may incorporate so-called long memory features in the volatility process by having multiple GARCH(1,1) type components in the model, as illustrated by Engle and Lee [3], or directly through a fractionally integrated EGARCH model, as considered by Baillie *et al.* [4]. In either case, volatility shocks will then eventually die out at a slower (approximate) hyperbolic rate than the geometric rate implied by a standard GARCH model. Three, one may allow for fat tails and asymmetries in the standardized return innovation through alternative distributional assumptions. Typical examples are the Student t distribution and the generalized error distribution (GED) as well as extensions thereof. Four, it is feasible to generalize the approach to multivariate systems and model the conditional variance-covariance matrix but parameters proliferate so it is critical to develop parsimonious representations for higher dimensional systems. A variety of approaches has been developed

with the dynamic conditional correlation (DCC) model of Engle [5] and Tse and Tsui [6] being the most tractable and popular. Finally, one may introduce additional explanatory (macroeconomic and financial) variables into the volatility equation and one may even accommodate latent regime shifts through clever parameterizations as in Gray [7]. For surveys providing more details on these issues, see, e.g., Andersen, Bollerslev, Christoffersen, and Diebold (ABCD) [8, 9] on volatility forecasting and risk management, respectively.

Stochastic Volatility Models

GARCH models are examples of stochastic volatility (SV) models. However, GARCH models imply, for given model parameters, that volatility is conditionally deterministic over the next period given the history of returns. Genuine SV models allow volatility for period t to be stochastic, even conditional on all information through period $t - 1$. This is more in line with the continuous-time paradigm of theoretical finance where volatility typically is driven, at least in part, by its own process separate from the return innovations. Moreover, it is natural to associate SV innovations to other concurrent market activity variables, such as trading volume, information flow or bid-ask spreads, see, e.g., Andersen [10]. A generic discrete-time SV model, imposing the mean zero constraint, may take the following form,

$$\begin{aligned} r_t &= \sigma_t z_t; \quad z_t \sim i.i.d. \ N(0, 1); \\ \sigma_t &= h(\sigma_{t-1}, r_{t-1}) + \eta_t; \\ \eta_t &\sim i.i.d. \ (0, \sigma_\eta^2) \end{aligned} \quad (18)$$

This is a generalization of the GARCH return equation (11). The decomposition of the return process into a standardized innovation, z_t , and a volatility process, σ_t , is as before. However, the volatility process is not only a function, $h(\cdot, \cdot)$, of past volatility and returns but also subject to a contemporaneous shock, η_t . The latter feature renders volatility latent or unobserved even conditional on all information through time $t - 1$. SV models are furthermore compatible with the *mixture of distributions hypothesis* (MDH) popularized by Clark [11]. In this theory, the return innovation is conditionally Gaussian, but with a variance which is governed by a separate stochastic process. As such, the full return distribution is a

mixture of the standard normal innovation and the associated variance process. This type of mixture process can rationalize the conditional fat-tailed innovations invariably found in return series. Besides this basic feature, SV models possess a great deal of flexibility through alternative specifications of the h function governing the volatility dynamics and inclusion of additional lags and/or explanatory variables.

Because the SV models imply a latent volatility process, they can be harder to estimate and analyze than GARCH models even if dramatic progress has been made recently in terms of estimation and inference *via* different simulation based techniques. Nonetheless, the basic features of the estimated models and the ability to empirically fit the properties of the volatility processes are not very different across the two model classes. Hence, we do not explore these models further and instead refer to existing surveys such as Andersen *et al.* (ABCD) [9] or Shephard [12].

Realized Volatility Modeling

The increased availability of high-frequency intraday data has spurred a novel approach to volatility measurement and modeling, in some sense taking us full circle and bringing back into focus a concept that refines the notion of historical volatility discussed in the introduction. The motivation is that if volatility evolves continuously then a rich history of tick-by-tick transaction prices and quotes should provide an opportunity to measure the process with good precision, even over relatively short intervals, as argued in Andersen and Bollerslev [13, 14] by extension of the ideas of Merton [15] and also independently developed by Zhou [16]. Here, we follow Andersen *et al.* [17] (henceforth ABDL) and define the *realized variation* (informally also denoted *realized volatility* (RV)), on day t using returns constructed at the Δ intraday frequency as

$$RV_t \equiv \sum_{j=1}^{1/\Delta} r_{t-1+j\Delta, \Delta}^2 \quad (19)$$

where $1/\Delta$ is, for example, 144 for 10-min returns in 24-h markets or 78 for 5-min returns over a 6.5-h trading day. In theory, letting Δ go to zero, which implies continuous sampling, RV will approach

the true *integrated variation*, IV , of the underlying return process under weak assumptions related to the absence of arbitrage. Hence, we may obtain direct and observable indicators of realized return variation for each day in our sample, independently of any specific model. This simplifies matters greatly as we avoid issues such as model adequacy and inference regarding latent volatility processes.

A simple scenario illustrates how the RV relates to actual return volatility. Assume the true (mean zero) price process evolves according to a semimartingale in continuous time and does not display discontinuities (or jumps), and that the volatility process evolves independently of the price process. Then the return over period t , conditional on the full path of the volatility process over the period, is normally distributed and characterized through the simple summary statistics, the integrated variance IV_t ,

$$r_t | \sigma_{[t-1, t]} \sim N(0, IV_t),$$

$$\text{where} \quad IV_t = \int_{t-1}^t \sigma_u^2 du \quad (20)$$

This result is directly in line with the discrete-time return distribution in SV models as stated in equation (18). The return is conditional Gaussian with a variance governed by the (stochastic) realization of the integrated variance for period t . Hence, the RV approach extends the SV models naturally to the continuous-time setting. The return distribution is a mixture of the standard normal and the distribution of the (integrated) variance, where the IV_t measure is genuinely random and thus not fully predictable from information available at time $t-1$. It is also worth noting that the IV concept occupies a central role in the theory of option pricing under SV, see Hull and White [18]. Of course, we never observe IV_t perfectly, even *ex post*, but only estimate it, e.g., *via* the corresponding cumulative sum of squared intraday returns, RV_t . The discretization error incurred because RV is constructed from a finite sample of intraday returns rather than a continuous price record may be assessed using the asymptotic theory developed by Barndorff *et al.* [19].

If the assumptions above are violated because of price jumps or correlation between price and volatility innovations, the return process will display more complex distributional characteristics (*see Lévy Processes in Asset Pricing*). Some of these features are

subject of current work in the literature. For example, tests for the presence of jumps are available using alternative cumulative return variation measures. One may thus decompose RV into contributions from jumps *vis-a-vis* the continuous part of the sample path. This has proven useful for volatility forecasting as the persistence of the volatility process predominantly is driven by the continuous piece of the return distribution. As a consequence, filtering out the jump component helps improve forecast accuracy, as documented in Andersen *et al.* [20], henceforth ABD.

Another complication is the presence of market microstructure features and/or noise that tend to obscure the properties of the true return process at the very highest sampling frequencies. This includes issues such as bid-ask bounce, the existence of a price grid, and the lack of a truly continuous sampling and trading/quoting record. These complications suggest that it is better to compute RV from somewhat lower frequencies such as 1- to 10-min intervals rather than tick-by-tick data or including explicit corrections for the presence of noise. A nice overview of this literature is given by Hansen and Lunde [21]. In either case, the measure of realized variation includes the true underlying return variation plus an error term, which is in general, serially uncorrelated. A related issue is the presence of a strong intraday volatility pattern which induces an artificial dependence structure in RV measures spanning less than a full trading day. A simple and robust response is to only construct RV measures for periods that comprise (a multiple of) one trading day, and this is now common practice in the literature.

The availability of actual return-volatility observations with only an uncorrelated measurement error disguising the true underlying realizations provides the impetus for direct time series modeling of volatility. For example, ABDL pursue a standard ARMA style model, only extended to allow for long memory features, for the logarithmic RV of foreign exchange rates. The associated fit for the future return volatility is found to dominate that of traditional volatility GARCH models based on daily data. An even simpler and quite effective approach is to pursue a simple component based regression model. The long memory features of the volatility process is accommodated well in this fashion. Specifically, using the so-called heterogeneous autoregressive (HAR) model of Corsi [22], ABD propose a simple HAR-RV specification

where $R_{t-h:t-1}$ denotes the average daily RV over the interval $[t-h, t-1]$,

$$RV_t = c_0 + c_d RV_{t-1} + c_w RV_{t-5:t-1} + c_m RV_{t-22:t-1} + u_t \quad (21)$$

The regression coefficients may be estimated directly by ordinary least squares (OLS) from historical data, and forecast then generated for future volatility in a straightforward manner. The inclusion of the past daily, weekly, and monthly RV is sufficient to provide a reasonable approximation to the slow hyperbolic delay in the (realized) volatility persistence documented in numerous studies.

Obviously, the simple model in equation (21) may be refined and extended in many directions and much current work is dedicated to doing so. Besides the jump filtering mentioned previously, examples include incorporation of overnight and holiday volatility indicators, the treatment of scheduled economic news announcements, and the addition of alternative explanatory variables such as volatility forecasts implied by option prices or volatility swap contracts (*see Volatility Smile*). Finally, it is of interest to document the economic, as opposed to statistical, value of RV forecasts for practical applications. An initial illustration is given by Fleming *et al.* [23].

In principle, it is also straightforward to extend the realized variation concept to encompass realized covariation. This is achieved by cumulating concurrent high-frequency return cross products rather than squared returns. The ability to provide direct measures of time-varying return covariances are of direct interest for factor-based asset pricing models, including the capital asset pricing model (CAPM) and the arbitrage pricing theory. Promising and intriguing results have been obtained for liquid assets, see, e.g., Andersen *et al.* [24] documenting systematic shifts in the exposure of value *versus* growth portfolios over the business cycle. Unfortunately, the covariance measures are even more susceptible to market microstructure distortions than the basic RV measure. Consequently, this issue is subject to intense research efforts, but nobody has yet proposed a method that is both tractable and robust. However, this work is still in its infancy and much progress will surely occur over the next few years.

Finally, there is the issue of how to apply RV-like approaches if markets are not quite as liquid

as the major financial markets. One possibility is to lower the sampling frequency and measure volatility over longer time intervals, e.g., weekly or monthly rather than daily. Another approach uses more robust volatility indicators such as the high–low price range to deal with the worse signal-to-noise ratio in less liquid markets, see, e.g., Alizadeh *et al.* [25]. A particularly compelling piece detailing the links between RV and intraday range–based volatility measures is given by Dobrev [26].

Conclusion

In summary, volatility modeling continues to pose interesting and important practical challenges that motivate the development of new econometric theory and implementation techniques. Moreover, the current frontier research dealing with volatility and correlation measurement via intraday data is finding immediate applications in trading firms and investment banks. This direct interaction between theory and practice is sure to invigorate further research in the coming years.

References

- [1] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**, 307–327.
- [2] Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: a new approach, *Econometrica* **59**, 347–370.
- [3] Engle, R.F. & Lee, G.G.J. (1999). A permanent and transitory component model of stock return volatility, in *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W.J. Granger*, R.F. Engle & H. White, eds, Oxford University Press, Oxford, pp. 475–497.
- [4] Baillie, R.T., Bollerslev, T. & Mikkelsen, H.O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **74**, 3–30.
- [5] Engle, R.F. (2002). Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models, *Journal of Business and Economic Statistics* **20**, 339–350.
- [6] Tse, Y.K. & Tsui, A.K.C. (2002). A multivariate generalized autoregressive conditional heteroskedasticity model with time-varying correlations, *Journal of Business and Economic Statistics* **20**, 351–362.
- [7] Gray, S. (1996). Modeling the conditional distribution of interest rates as a regime-switching process, *Journal of Financial Economics* **42**, 27–62.
- [8] Andersen, T.G., Bollerslev, T., Christoffersen, P. & Diebold, F.X. (2005). Practical volatility and correlation modeling for financial market risk management, in *Risks of Financial Institutions*, M. Carey & R. Stulz, eds, University of Chicago Press for National Bureau of Economic Research, Chicago.
- [9] Andersen, T.G., Bollerslev, T., Christoffersen, P. & Diebold, F.X. (2006). Volatility and correlation forecasting, in *Handbook of Economic Forecasting*, G. Elliott, C.W.J. Granger & A. Timmermann, eds, North-Holland, Amsterdam, Vol. 1, pp. 777–878.
- [10] Andersen, T.G. (1996). Return volatility and trading volume: an information flow interpretation of stochastic volatility, *Journal of Finance* **51**, 169–204.
- [11] Clark, P.K. (1973). A subordinated stochastic process model with finite variance for speculative prices, *Econometrica* **41**, 135–156.
- [12] Shephard, N. (2005). *Stochastic Volatility: Selected Readings*, Oxford University Press, Oxford.
- [13] Andersen, T.G. & Bollerslev, T. (1997). Heterogeneous information arrivals and return volatility dynamics: uncovering the long run in high frequency returns, *Journal of Finance* **52**, 975–1005.
- [14] Andersen, T.G. & Bollerslev, T. (1998). Deutsche Mark–Dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies, *Journal of Finance* **53**, 219–265.
- [15] Merton, R.C. (1980). On estimating the expected return on the market: an explanatory investigation, *Journal of Financial Economics* **8**, 323–361.
- [16] Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates, *Journal of Business and Economic Statistics* **14**, 45–52.
- [17] Andersen, T.G., Bollerslev, T., Diebold, F.X. & Labys, P. (2003). Modeling and forecasting realized volatility, *Econometrica* **71**, 529–626.
- [18] Hull, J. & White, A. (1987). The pricing of options on assets with stochastic volatilities, *Journal of Finance* **42**, 281–300.
- [19] Barndorff-Nielsen, O.E. & Shephard, N. (2002). Estimating quadratic variation using realised variance, *Journal of Applied Econometrics* **17**, 457–477.
- [20] Andersen, T.G., Bollerslev, T. & Diebold, F.X. (2007). Roughing it up: including jump components in measuring, modeling, and forecasting asset return volatility, *Review of Economics and Statistics* **89**(4).
- [21] Hansen, P.R. & Lunde, A. (2006). Realized variance and market microstructure noise, *Journal of Business and Economic Statistics* **24**, 127–161.
- [22] Corsi, F. (2003). *A Simple Long Memory Model of Realized Volatility*, Working Paper, University of Southern Switzerland.
- [23] Fleming, J., Kirby, C. & Ostdiek, B. (2003). The economic value of volatility timing using realized volatility, *Journal of Financial Economics* **67**, 473–509.
- [24] Andersen, T.G., Bollerslev, T., Diebold, F.X. & Wu, J. (2005). A framework for exploring the macroeconomic

determinants of systematic risk, *American Economic Review* **95**, 398–404.

- [25] Alizadeh, S., Brandt, M.W. & Diebold, F.X. (2002). Range-based estimation of stochastic volatility models, *Journal of Finance* **57**, 1047–1092.
- [26] Dobrev, D. (2006). *Capturing Volatility from Large Price Moves: Generalized Range Theory and Applications*, Working Paper, Kellogg School of Management, Northwestern University, Evanston.

Related Articles

Model Risk

Value at Risk (VaR) and Risk Measures

TORBEN G. ANDERSEN

Volatility Smile

The Black and Scholes [1] model lays the foundation to the modern option pricing theory. It prescribes the option price as a function of the volatility of the underlying asset and some other pertinent and observable variables. One important assumption of the Black–Scholes model is that the price of the underlying asset follows a geometric Brownian motion with constant drift and volatility; that is,

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t^P \quad (1)$$

where W_t^P is the standard Brownian motion with respect to the physical probability measure P . By employing a continuous hedging argument, they then derived the following European call option pricing formula:

$$C_t = S_t N(d_t) - X e^{-r(T-t)} N\left(d_t - \sigma \sqrt{T-t}\right) \quad (2)$$

where $d_t = \frac{\ln(S_t/X) + (r + 0.5\sigma^2)(T-t)}{\sigma \sqrt{T-t}}$, $N(\cdot)$ is the standard normal distribution function, X is the strike price, $T - t$ is the remaining maturity, and r is the risk-free interest rate (see **Numerical Schemes for Stochastic Differential Equation Models**).

Subsequently, several researchers [2–4] proposed a powerful way of viewing option pricing as a risk-neutral valuation problem (see **Risk-Neutral Pricing: Importance and Relevance**). After transforming the probability measure P to Q , one can price options by acting as if one were risk-neutral and employing the following price process with an altered drift:

$$\frac{dS_t}{S_t} = r dt + \sigma dW_t^Q \quad (3)$$

where W_t^Q is the standard Brownian motion with respect to the risk-neutral probability measure Q . This approach yields the same Black–Scholes option pricing formula. In fact, the idea is applicable to any continuous-time continuous sample path semimartingale so that the discounted asset value becomes a martingale under measure Q .

Empirically, the Black–Scholes model has been shown to have rather unsatisfactory performance. Its deficiency is most notable when the Black–Scholes implied volatility is related to strike price and

maturity. For a European option, one can easily compute the implied volatility by using the Black–Scholes formula to back out a unique volatility that equates the formula value to the market price of the option. Since the implied volatility is meant for the underlying asset, options on the same underlying should not have their implied volatilities exhibiting any systematic pattern in relation to strike price or maturity. The empirical evidence extensively reported in the literature, however, suggests otherwise. Figure 1 reports the implied volatilities of the S&P 500 index options (out-of-the-money calls and puts) on April 7, 2006 corresponding to different maturities and strike prices. The patterns in Figure 1 are typical of index options. It is evident that the implied volatility corresponding to a maturity is generally downward sloping when the strike price is less than the S&P 500 index value, i.e., out-of-the-money puts (corresponding to in-the-money calls by the put-call parity). When the strike price goes beyond the index value, it begins to slope upward. The implied volatility curve in relation to the strike price is commonly referred to as the volatility smile (sometimes volatility smirk). If we fix the attention at a strike price and examine them in terms of maturity, it is evident that the implied volatility will be mostly decreasing in maturity except for the strike price at around the index value. Such kind of phenomenon is known as the term *structure of implied volatilities*.

Understanding what causes volatility smile is crucial to the financial economists' search for a better option pricing model. At the practical level, devising a workable way of incorporating volatility smile is also vital to industry practitioners in their daily tasks of pricing options and risk manage their trading books. Researchers and practitioners have sought to address volatility smile along two paths. The first path is to incorporate volatility smile through nonparametric techniques, such as implied-tree models, kernel regressions, neural networks, and entropy principle, without actually focusing on the cause of volatility smile. Naturally, such solutions without subjecting to economic constraints are likely to work well for the tasks that are in essence interpolation, say for example, pricing a European option with a different strike price. Among them, the implied-tree models are different and deserve further comments later.

The second path is through parametric approaches such as stochastic volatility models, jump-diffusion

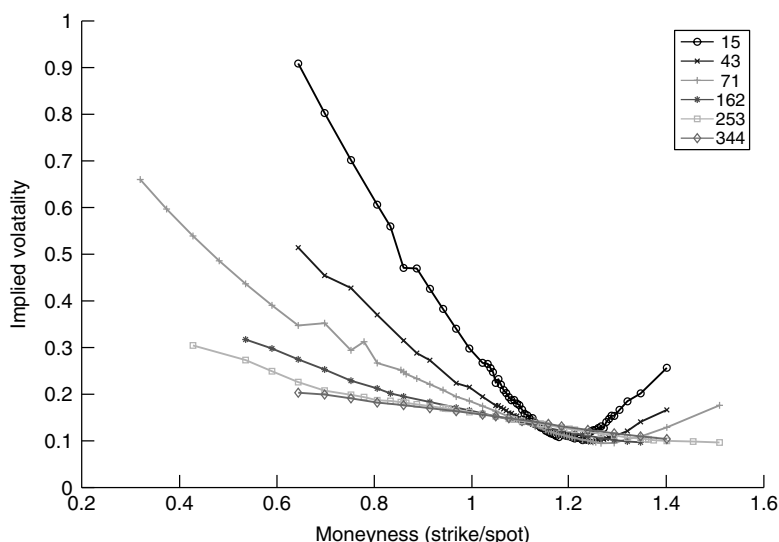


Figure 1 Implied volatilities of the S&P 500 index options (out-of-the-money calls and puts) on April 7, 2006 corresponding to different maturities and strike prices

models (see **Lévy Processes in Asset Pricing**) and generalized autoregressive conditional heteroskedasticity (GARCH) models (see **Statistical Arbitrage**). These are pricing theories comparable to the Black–Scholes model in the sense that they first postulate an asset price dynamic and then derive the option pricing model through an economic argument. These models can, in principle, price any option – European, American, and exotic. Although the economic constraints may lead them to perform not as well for specific interpolation tasks, they have a much wider applicability for a whole spectrum of pricing and risk-management tasks. In this article, we will cover both modeling approaches and offer a brief description and discussion on specific methods in both categories. In particular, we will demonstrate by examples the use of the kernel regression method and the GARCH option pricing model.

Nonparametric Methods

Nonparametric methods leave specific function forms open and let data speak for themselves. The performance of nonparametric methods will, of course, be better when the assumptions of parametric models are violated. Flexibility inevitably comes at a cost. Nonparametric techniques typically require a large

quantity of data in implementation, which limits their applicability or renders them completely unworkable in some circumstances. Nevertheless, nonparametric techniques offer a valuable set of tools that can be extremely useful for some applications in option pricing and risk management. The well-known nonparametric methods that can handle volatility smile include implied-tree models [5–7], learning networks [8], principle of maximum entropy [9–11], and kernel regressions [12].

In this article, we use kernel regression as an example to demonstrate the use of nonparametric techniques. Kernel regression is a popular tool for figuring out the nonlinear relationship among variables. The main assumption is that the unknown functional form is sufficiently smooth. Kernel regression was first applied to option pricing by Ait-Sahalia and Lo [12]. They assumed that the Black–Scholes’ implied volatility is a smooth function of two contract attributes – moneyness (strike price over asset price) and maturity. However, the functional form is unknown. Ait-Sahalia and Lo [12] employed the Nadaraya–Watson kernel estimator, which in this case is the weighted average of observed implied volatilities for the existing options with the weight depending on how close the target point’s attributes are to those of the existing options. If the distance to an option is large (small), the weight will

be small (large). The kernel function is a two-dimensional Gaussian density with a bandwidth to be determined by the data. (The bandwidth determines how fast the weight declines in relation to the distance).

Since the kernel regression is applied to the implied volatility directly, the Black–Scholes formula is needed to recover the option value corresponding to the implied volatility predicted by the kernel regression. Needless to say, one can also apply the kernel regression directly on option prices. In principle, neither way dominates, but in practice, the choice should depend on the specific application and/or one's preferences in weighting in- and out-of-the-money options; for example, using option prices as opposed to implied volatilities amounts to committing to a different weighting scheme on the regression errors.

The kernel estimator is a smooth function by construction, and can thus be used to interpolate in moneyness and maturity. But the kernel estimator should not be used for extrapolation, because no reference points can be used to anchor the unknown functional relationship beyond the data range. In fact, the quality of Nadaraya–Watson kernel estimator is known to be poor for points within but close to the boundary of the data range. An alternative with a better boundary performance is a local linear kernel regression [13]. Kernel regressions are much more data intensive than implied-tree models, which are a category of nonparametric techniques based on precommitting to a tree structure in fitting the observed option prices. Furthermore, the kernel regression function constructed from one type of options cannot be used to price other types of options, whereas implied-tree models can be. Evidence suggests, however, implied-tree models are too restrictive to match volatility smiles over time without frequent recalibrations. Thus, kernel regressions stand to be a very useful nonparametric tool for option pricing. Now we demonstrate the performance of a kernel regression in matching volatility smiles both in-sample and out-of-the-sample.

In deciding what enter into the kernel function, we note the importance of adding historical volatility as the third attribute. Including historical volatility is important because the asset volatility is known to change over time. Thus, our kernel regression consists of three regressors – moneyness, maturity, and historical volatility. The historical volatility is computed from the 252 daily returns

immediately preceding any given day in the sample. For simplicity, we use the Nadaraya–Watson kernel estimator with a three-dimensional Gaussian kernel function. The bandwidth determines how fast the weight declines. Too small a bandwidth can result in an in-sample overfit, whereas too large a bandwidth can cause over smoothing and large biases. The bandwidth in our demonstration is determined with the standard leave-one-out cross validation.

All strike prices and maturities during the period March 1, 2006 to March 31, 2006 (daily) are included if market prices of options give rise to valid implied volatilities. Option data are taken from option metrics. The root mean squared error of annualized implied volatilities for the sample is 0.0024. Fitting errors (the market annualized implied volatility minus the kernel estimate) are smaller for longer-term options. This is not surprising as Figure 1 indicates that the volatility smile phenomenon is least pronounced for long-term options.

Naturally, one would like to know how the kernel regression function performs out-of-the-sample. For this, we check the performance one week ahead on April 7, 2006. The root mean squared error turns out to be 0.019764, which is about eight times the in-sample error magnitude. This suggests that the in-sample kernel regression has likely been overfitted. One plausible cause is that historical volatility, one of our three regressors, fails to properly reflect market volatility but the kernel regression function has been forced to adapt to the in-sample option data.

Figure 2 plots the out-of-the-sample fitting errors of the kernel regression for the implied volatilities of the S&P 500 index options on April 7, 2006. Fitting errors of short-term options are larger and exhibit a smilelike shape, indicating that kernel regression has not succeeded in removing volatility smile completely.

Parametric Models

Parametric option pricing models begin by postulating a stochastic process for the underlying risk factor(s), typically the underlying asset price. Economic arguments (arbitrage (*see Statistical Arbitrage*) or equilibrium (*see Risk-Neutral Pricing: Importance and Relevance*)) are then employed to develop the functional relationship linking the price of an option to the underlying asset price and the parameters which govern the stochastic process and

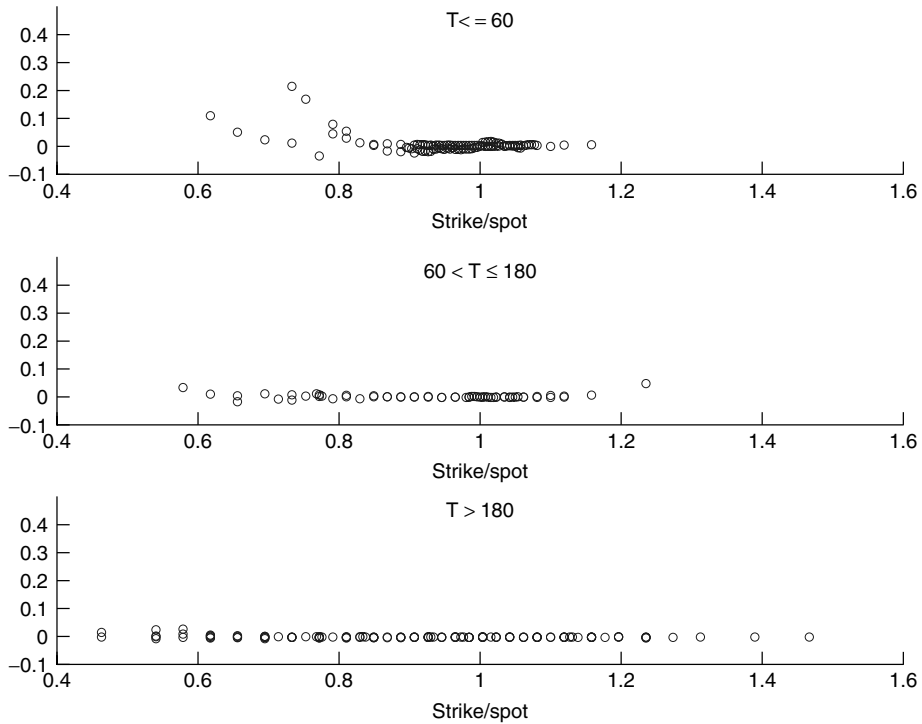


Figure 2 Out-of-the-sample fitting errors of the kernel regression for the implied volatilities of the S&P 500 index options (April 7, 2006)

the option contract. The parametric option pricing model can be implemented with first estimating the parameters of the stochastic process using the time series of the underlying asset prices. Alternatively, one can directly apply the functional relationship to calibrate the model to the observed option prices without using the price time series. The former implementation scenario is uniquely available to parametric option pricing models because they offer an explicit link between the risk-neutral and physical price dynamics for the underlying asset. In principle, one can also combine both the underlying asset and option prices to perform a joint estimation of a parametric model.

Benefiting from the model restrictions, implementing parametric option pricing models is less data intensive. In addition, parametric option pricing models can be used to extrapolate for options with strike prices and maturities outside the data range. In principle, these models can also be used to price different types of derivatives, because different pricing functions share the same model parameters.

A parametric pricing model thus provides a complete solution to option pricing. In practice, however, the restrictions placed on the system for deriving the model may be too constraining, and thus render the resulting model a poor performer empirically. The well-known parametric option pricing models are stochastic volatility [14, 15], jump-diffusion [16, 17], and GARCH [18].

We use the GARCH option pricing model to show how a parametric pricing model can be used to deal with volatility smile. Autoregressive conditional heteroskedasticity (ARCH) modeling was pioneered by Engle [19] and its generalized version, known as *GARCH*, has been widely adopted in modeling financial time series with proven success. In a nutshell, the GARCH model treats the volatility as a deterministic function of past innovations. Different GARCH specifications simply employ different deterministic functions in linking the past innovations to the current volatility. The GARCH volatility is not a latent variable because volatility is simply a deterministic transformation of the past innovations. This stands

in sharp contrast to stochastic volatility models and has practical significance for option pricing, because volatility in the stochastic volatility model, being a latent variable, has a nondegenerate distribution which makes model implementation difficult.

The option pricing theory under GARCH was first developed by Duan [18] using an equilibrium argument. Later, Kallsen and Taqqu [20] showed that the same GARCH option pricing model can be obtained by an arbitrage-free argument (*see Statistical Arbitrage*) using a properly constructed continuous-time GARCH process. Estimating the GARCH model using the time series of asset prices is an easy task, and most statistical packages, in fact, have built-in modules for this. Valuing derivatives using the GARCH option pricing model is generally speaking computer intensive. For European options, some approximation formulas are available, however. Empirical studies show that the GARCH option pricing model performs well particularly with the nonlinear asymmetric generalized autoregressive conditional heteroskedasticity (NGARCH) specification of Engle and Ng [21]; see, for example, Christoffersen and Jacobs [22].

We now use the GARCH option pricing model to demonstrate its application to option data. Specifically, the asset price under the physical probability measure P is assumed to follow the NGARCH(1,1) specification of Engle and Ng [21]:

$$\begin{aligned} \ln \frac{S_{t+1}}{S_t} &= r - q + \lambda \sigma_{t+1} - \frac{1}{2} \sigma_{t+1}^2 + \sigma_{t+1} \varepsilon_{t+1}^P \\ \sigma_{t+1}^2 &= \beta_0 + \beta_1 \sigma_t^2 + \beta_2 \sigma_t^2 (\varepsilon_t^P - \theta)^2 \end{aligned} \quad (4)$$

where ε_{t+1}^P is a P -standard normal random variable conditional on the time t information; and q is the continuously compounded dividend yield. A positive value for parameter θ leads to an asymmetric volatility response to the return innovation, an empirical phenomenon commonly known as *leverage effect*.

By the local risk-neutral valuation relationship of Duan [18], the corresponding risk-neutral pricing system becomes

$$\begin{aligned} \ln \frac{S_{t+1}}{S_t} &= r - q - \frac{1}{2} \sigma_{t+1}^2 + \sigma_{t+1} \varepsilon_{t+1}^Q \\ \sigma_{t+1}^2 &= \beta_0 + \beta_1 \sigma_t^2 + \beta_2 \sigma_t^2 (\varepsilon_t^Q - \theta - \lambda)^2 \end{aligned} \quad (5)$$

where $\varepsilon_{t+1}^Q = \varepsilon_{t+1}^P + \lambda$ is a Q -standard normal random variable conditional on the time t information. Because both θ and λ are expected to be positive, this

transformation from measure P to Q increases the volatility persistence, which in turn raises the stationary volatility and causes the cumulative asset return to be more negatively skewed.

There are five parameters in this pricing model: $\beta_0, \beta_1, \beta_2, \theta$, and λ . Maximum-likelihood estimation is performed using the time series of the S&P 500 index values from April 1, 2001 to March 31, 2006. It is well-known in the time series literature that at daily frequency it is hard to estimate the conditional mean parameters (in this case λ) with satisfactory accuracy. Thus the estimates are kept for $\beta_0, \beta_1, \beta_2$, and θ but λ is updated by calibrating the GARCH option pricing model to the option data over the period March 1, 2006 to March 31, 2006, which is the same period used previously in the kernel regression analysis. The calibration is conducted by matching the implied volatility of the GARCH option price to the implied volatility of the market price. The GARCH option prices are computed with 5000 sample path empirical martingale simulation of Duan and Simonato [23]. This two-step procedure yields the following result: $\beta_0 = 5.34 \times 10^{-7}$, $\beta_1 = 0.88863$, $\beta_2 = 0.02599$, $\theta = 1.76056$, and $\lambda = 0.17159$. The root mean squared error of the calibration step is 0.03090, which is much larger than 0.0024, the root mean squared error of the kernel regression on the same data set. Similar to the kernel regression, the fitting errors are larger for short-term options. Moreover, the fitting errors for short-term options continue to exhibit a smilelike shape, indicating that the GARCH option pricing model has not succeeded in removing volatility smile completely.

The fitting result suggests that a better risk-neutral distribution will be the one with heavier tails. The inability of the GARCH option pricing model of Duan [18] may be attributed to its imposition of conditional normality. The GARCH option pricing model with a compound Poisson innovation such as in [24] will likely fit these options better.

When the GARCH option pricing model is applied out-of-the-sample, it produces a fairly stable performance result. The performance one week ahead on April 7, 2006 yields the root mean squared error of 0.05418. This stands in sharp contrast to the kernel regression method, which yields a drastically worse out-of-the-sample performance.

Figure 3 plots the out-of-the-sample fitting errors for three maturity categories. Compared to those of kernel regression, the fitting errors are larger but

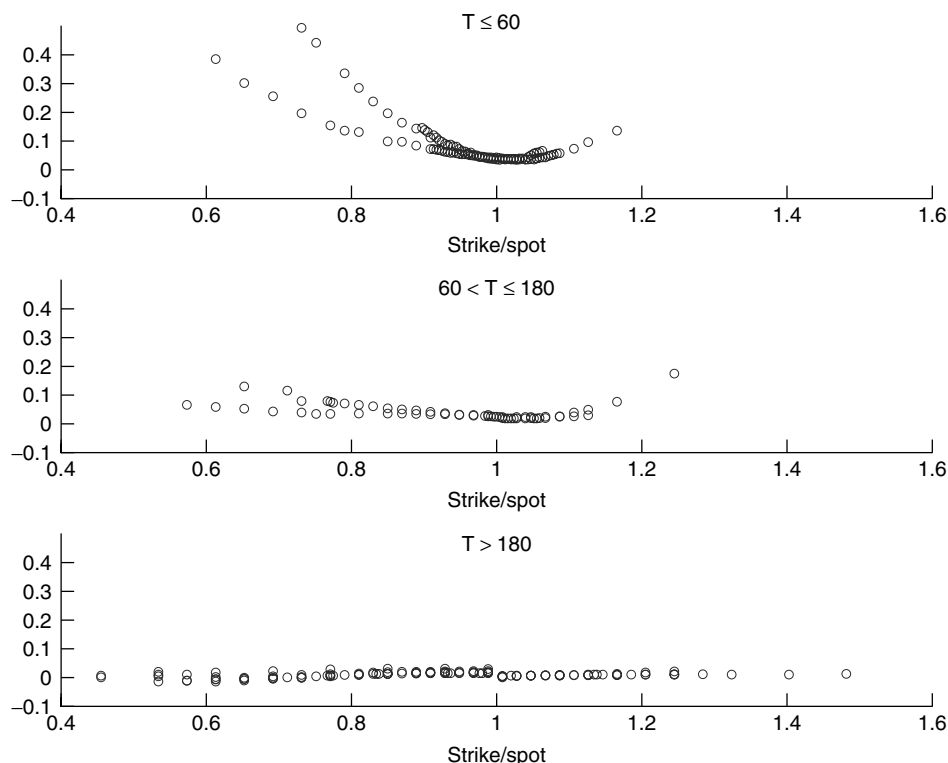


Figure 3 Out-of-the-sample fitting errors of the GARCH option pricing model for the implied volatilities of the S&P 500 index options (April 7, 2006)

have similar patterns, i.e., fitting errors of short-term options are larger and exhibit a smilelike shape.

References

- [1] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- [2] Cox, J. & Ross, S. (1976). The valuation of options for alternative stochastic processes, *Journal of Financial Economics* **3**, 145–166.
- [3] Harrison, J. & Kreps, D. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- [4] Harrison, J. & Pliska, S. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and Their Applications* **11**, 261–271.
- [5] Derman, E. & Kani, I. (1994). Riding on a smile, *Risk* **7**, 32–39.
- [6] Dupire, B. (1994). Pricing with a smile, *Risk* **7**, 18–20.
- [7] Rubinstein, M. (1994). Implied binomial trees, *Journal of Finance* **49**, 771–818.
- [8] Hutchinson, J., Lo, A. & Poggio, T. (1994). A non-parametric approach to pricing and hedging derivative securities via learning networks, *Journal of Finance* **49**, 851–889.
- [9] Buchen, P. & Kelly, M. (1996). The maximum entropy distribution of an asset inferred from option prices, *Journal of Financial and Quantitative Analysis* **31**, 143–159.
- [10] Stutzer, M. (1996). A simple nonparametric approach to derivative security valuation, *Journal of Finance* **51**, 1633–1652.
- [11] Duan, J. (2002). *Nonparametric Option Pricing by Transformation*, University of Toronto working paper.
- [12] Ait-Sahalia, Y. & Lo, A. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices, *Journal of Finance* **53**, 499–547.
- [13] Fan, J. (1992). Design-adaptive nonparametric regression, *Journal of the American Statistical Association* **87**, 998–1004.
- [14] Hull, J. & White, A. (1987). The pricing of options on assets with stochastic volatility, *Journal of Finance* **42**, 281–300.
- [15] Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond

- and currency options, *Review of Financial Studies* **6**, 327–344.
- [16] Merton, R. (1976). Option pricing when underlying stock returns are discontinuous, *Journal of Financial Economics* **3**, 125–144.
- [17] Bates, D. (2000). Post-'87 crash fears in the S&P 500 futures option market, *Journal of Econometrics* **94**, 181–238.
- [18] Duan, J. (1995). The GARCH option pricing model, *Mathematical Finance* **5**, 13–32.
- [19] Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation, *Econometrica* **50**, 987–1008.
- [20] Kallsen, J. & Taqqu, M. (1998). Option pricing in ARCH-type models, *Mathematical Finance* **8**, 13–26.
- [21] Engle, R. & Ng, V. (1993). Measuring and testing the impact of news on volatility, *Journal of Finance* **48**, 1749–1778.
- [22] Christoffersen, P. & Jacobs, K. (2004). Which GARCH model for option valuation? *Management Science* **50**, 1204–1221.
- [23] Duan, J. & Simonato, J. (1998). Empirical martingale simulation for asset prices, *Management Science* **44**, 1218–1233.
- [24] Duan, J., Ritchken, P. & Sun, Z. (2005). *Jump Starting GARCH: Pricing and Hedging Options with Jumps in Returns and Volatilities*, University of Toronto working paper.

Related Articles

Decision Trees

Volatility Modeling

JIN-CHUAN DUAN AND YUN LI

Vulnerability Analysis for Environmental Hazards

What makes people and places vulnerable to environmental hazards and threats? Broadly speaking, vulnerability is the potential for loss or other adverse impacts, or the capacity to suffer harm from an environmental hazard [1, 2]. Hazards, disasters, or extreme events arise from the interaction between human systems and natural systems (natural hazards or natural disasters); but they also can arise from technological failures such as chemical accidents, and from human-induced or willful acts such as terrorism [3]. Vulnerability can be examined at the individual level (for example, the health sciences look at individual people, while the engineering sciences look at the vulnerability of individual buildings or structures). However, hazards vulnerability is most often used to describe the susceptibility of social systems (human communities) or environmental systems (ecosystems) to environmental threats (*see Environmental Health Risk; Environmental Hazard*).

Vulnerability science is in its intellectual infancy and builds on the existing interdisciplinary and integrated traditions of risk, hazards, and disasters research, incorporating both qualitative and quantitative approaches, local to global geography, historic to future temporal domains, and best practices [4]. Spatial social science is critical in advancing vulnerability science through improvements in geospatial data, basic science, and application. It adds technological sophistication and analytical capabilities, especially in the realm of the geospatial and computational sciences, and integrates perspectives from the natural, social, health, and engineering sciences.

Vulnerability analysis seeks to understand the circumstances that place people and localities at risk (their susceptibility), in particular, those circumstances that enhance or reduce the ability of people and places to respond to them (their resilience). For example, some places are simply more vulnerable to hazards than others by virtue of their location (e.g., along coastlines, or along rivers) [5]. In other places, the vulnerability may be a function of the characteristics of people who live there (i.e., the homeless, the poor), where some social groups are more at risk and least able to recover from disasters than others.

Hazard Vulnerability

Physical hazard vulnerability is determined by examining the areal extent of exposure, and the specific characteristics of the potential hazards including magnitude, intensity, frequency, duration, temporal spacing, and rate of onset. For example, to examine the vulnerability of a location to flooding, the likely area of impact must first be determined. The geometry of the impact can be mapped, such as a floodplain (an area prone to flooding). Following this, either historic frequency data or probabilistic data can be applied to the delineated area to measure vulnerability. In this way, a spatial measure of the vulnerability is identified and the resultant product is the map of a 100-year floodplain (or the floodplain that will experience an annual 1% chance flood). Unfortunately, not all types of hazards have quantitative measures for all the characteristics, and in those instances qualitative descriptors are used to describe the events such as strong or weak, fast or slow, small area or large area. The mapping of the hazard zones has a long history [6]. However, the use of more sophisticated geospatial data and technologies such as geographic information systems (GISs) has led to improvements in hazard mapping and vulnerability analyses [7, 8].

Once these areas of exposure are delineated, then they can be used to determine the likely population at risk. The number of people and or structures in harm's way provides one set of measures for understanding vulnerability. Yet, individuals and societies have varying coping capacities, and vulnerability analysis needs to consider the differential capacity that social groups have in the face of disasters.

Social Vulnerability

Social vulnerability is a product of social inequalities. Social vulnerability is a well-developed scientific concept that transcends the expedient descriptors such as race, ethnicity, or poverty [9]. Although it is still not well-understood how each social factor works to produce vulnerability, or necessarily how factors come to work together as amplifiers or attenuators, several broad drivers of social vulnerability have found acceptance within the hazards literature. Generally, these include access to resources and political power; social capital and social networks; beliefs, culture, and customs; building stock and age; frail and

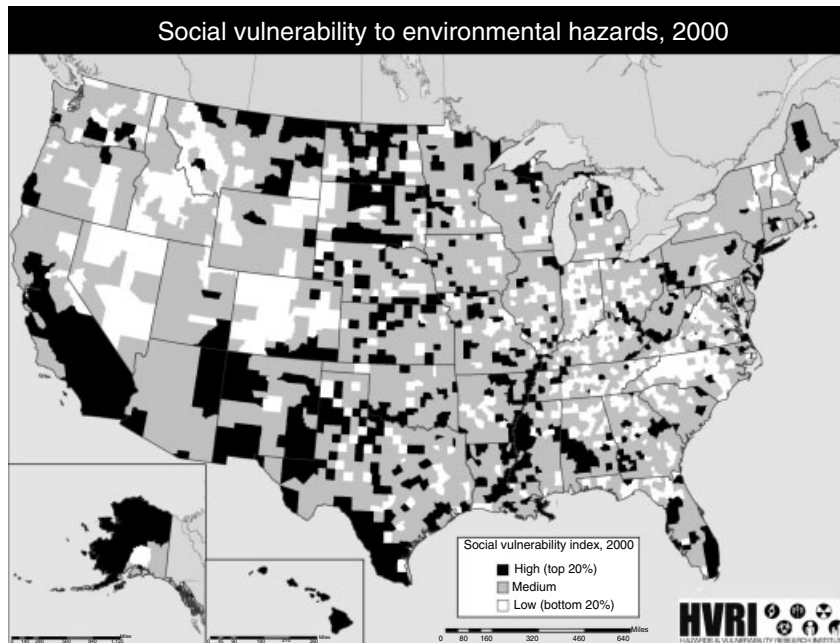


Figure 1 The Social Vulnerability Index (SoVI) for 2000. The cartographic representation of social vulnerability illustrates the extremes in the distribution of the scores. Those counties in black represent the most socially vulnerable places, while the counties in white are the least socially vulnerable. The underlying factors contributing to these patterns are socioeconomic status, age, and the density of the built environment. *Source:* <http://www.cas.sc.edu/geog/hrl/sovi.html> [Reproduced from [14]. © Blackwell Publishing, 2006.]

physically limited individuals; and type and density of infrastructure and lifelines [10–13].

An index of social vulnerability has been generated for US counties, which provides a quantitative assessment of the driving forces [14]. This metric (Social Vulnerability Index or SoVI) permits the identification of regional differences in the relative level of vulnerability across the United States (Figure 1). The SoVI helps to understand the preexisting conditions that contribute to social vulnerability and those circumstances that may influence the ability of the place to respond to and recover from disasters. While initially developed at the county scale for a single time period, SoVI has now been reproduced for various time periods (1960–2000) and for different geographical scales (census tracts) [15].

Policy Perspectives

In many places, it is the combination of social and physical vulnerabilities that potentially undermine a community's ability to respond and recover

from disaster [16, 17]. It is for this reason that vulnerability analyses play an important role in hazards and disaster planning and policy. As part of the US Disaster Mitigation Act of 2000, all state and local governments must develop hazard mitigation plans, and those plans must be based on all hazards vulnerability analyses. Multihazard maps are now part of the public risk and hazard communication effort (www.hazardmaps.gov/atlas.php). The implementation of a conceptually derived place-based vulnerability analysis [7] has led to more science-based hazards assessments at state and local levels. On the international level, there is also considerable interest in vulnerability assessments as part of global efforts to reduce the impacts of disasters [18, 19]. The use of vulnerability analysis in public policies and as the scientific basis for disaster reduction shows great promise. However, the choice of indicators, the availability of place-specific data, and the methodological approaches are still in their infancy, but with continued advancements in vulnerability science these should be overcome in short order [20].

References

- [1] Alexander, D. (2002). *Principles of Emergency Planning and Management*, Oxford University Press, New York.
- [2] Cutter, S.L. (1996). Vulnerability to environmental hazards, *Progress in Human Geography* **20**(4), 529–539.
- [3] Cutter, S.L. (2005). Hazards measurement, in *Encyclopedia of Social Measurement*, K. Kempf-Leonard, ed, Academic Press, New York, Vol. 2, pp. 197–202.
- [4] Cutter, S.L. (2003). The science of vulnerability and the vulnerability of science, *Annals of the Association of American Geographers* **93**(1), 1–12.
- [5] Cutter, S.L. (ed) (2001). *American Hazardscapes: The Regionalization of Hazards and Disasters*, Joseph Henry Press/National Academy of Sciences, Washington, DC.
- [6] Burton, I., Kates, R.W. & White, G.F. (1993). *The Environment as Hazard*, 2nd Edition, The Guilford Press, New York.
- [7] Cutter, S.L., Mitchell, J.T. & Scott, M.S. (2000). Revealing the vulnerability of people and places: a case study of Georgetown County, South Carolina, *Annals of the Association of American Geographers* **90**(4), 713–737.
- [8] Rashed, T. & Weeks, J. (2003). Assessing vulnerability to earthquake hazards through spatial multicriteria analysis of urban areas, *International Journal of Geographical Information Science* **17**(6), 547–576.
- [9] Cutter, S.L. & Emrich, C.T. (2006). Moral hazard, social catastrophe: the changing face of vulnerability along the hurricane coasts, *Annals of the American Academy of Political and Social Science* **604**, 102–112.
- [10] Heinz Center for Science Economics and the Environment (2002). *Human Links to Coastal Disasters*, Washington, DC.
- [11] Pelling, M. (2003). *The Vulnerability of Cities: Natural Disasters and Social Resilience*, Earthscan, London.
- [12] Bankoff, G., Frerks, G. & Hilhorst, D. (eds) (2004). *Mapping Vulnerability: Disasters, Development and People*, Earthscan, London.
- [13] Blaikie, P., Cannon, T., Davis, I. & Wisner, B. (2004). *At Risk: Natural Hazards, People's Vulnerability and Disasters*, 2nd Edition, Routledge, New York.
- [14] Cutter, S.L., Boruff, B.J. & Shirley, W.L. (2003). Social vulnerability to environmental hazards, *Social Science Quarterly* **84**(1), 242–261.
- [15] Cutter, S.L., Emrich, C.T., Mitchell, J.T., Boruff, B.J., Gall, M., Schmittlein, M.C., Burton, C.G. & Melton, G. (2006). The long road home: race, class, and recovery from Hurricane Katrina, *Environment* **48**(2), 8–20.
- [16] Boruff, B.J., Emrich, C. & Cutter, S.L. (2005). Hazard vulnerability of US coastal counties, *Journal of Coastal Research* **21**(5), 932–942.
- [17] Chakraborty, J., Tobin, G.A. & Montz, B.E. (2005). Population evacuation: assessing spatial variability in geophysical risk and social vulnerability to natural hazards, *Natural Hazards Review* **6**(1), 23–33.
- [18] United Nations, International Strategy for Disaster Reduction (2004). *Living with Risk: A Global Review of Disaster Reduction Initiatives*, New York, Geneva.
- [19] Dilley, M., Chen, R.S., Deichmann, U., Lerner-Lam, A.L., Arnold, M., Agwe, J., Buys, P., Kjekstad, O., Lyon, B. & Yetman, G. (2005). *Natural Disaster Hotspots: A Global Risk Analysis Synthesis Report*, The World Bank, Washington, DC.
- [20] Cutter, S.L. (2006). *Hazards, Vulnerability and Environmental Justice*, Earthscan, London, Sterling.

SUSAN L. CUTTER

Warranty Analysis

New products have been appearing on the market at an ever-increasing rate due to the ever-increasing pace of technological innovation. Each generation of products typically incorporates new features and is more complex than the one it replaces. With the increase in complexity and the use of new materials and design methodologies, product reliability becomes an increasingly important issue. One way for the manufacturer to reduce the buyer's risk is to offer a warranty with the sale of the product. In very simple terms, the warranty assures the buyer that the manufacturer will either repair or replace items that do not perform satisfactorily or refund a fraction or the whole of the sale price in the event of product failure. Offering a warranty results in a significant risk to the manufacturer in the form of the expected cost of servicing the warranty. This cost must be assessed very early in the product development stage in order to make decisions with regard to desired product reliability and appropriate product pricing.

This article deals with warranty cost analysis for some simple warranty policies. This involves building models for claims under warranty, taking into account the reliability of the product. The outline of the article is as follows. The section titled "Product Warranty" looks at the basic concept of a warranty and provides a few examples of commonly used consumer warranties. The cost of warranty is impacted by product reliability in an obvious fashion. Some key reliability results needed in assessing this impact are summarized in the section titled "Product Reliability". Probabilistic results used in modeling warranty costs are given in the section titled "Product Reliability" and are applied to the analysis of cost per unit sold and life cycle cost of warranty in the sections titled "Modeling Warranty Cost per Unit Sold" and "Warranty Cost Over Life Cycle", respectively.

Product Warranty

Warranty Concept

A warranty is a contractual agreement between buyer and manufacturer (or seller) that is entered into upon sale of the product or service. The purpose of a

warranty is to establish liability of the manufacturer in the event that an item fails or is unable to perform its intended function when properly used. The warranty (often referred to as the *base warranty*) is an integral part of the sale (a service bundled with the product) and the risk assumed by the manufacturer in offering the warranty is factored into the price. In contrast, extended warranty is a service contract that a buyer might purchase separately to extend the warranty coverage once the base warranty expires.

Warranties are an integral part of nearly all consumer and commercial purchases and also of many government transactions that involve manufactured products. In such transactions, warranties serve a somewhat different purpose for buyer and manufacturer. For a more detailed discussion of the role of warranty from buyer and manufacturer perspectives, see [1]. There are many aspects of warranty and these have been studied by researchers from many different disciplines. A discussion of these can be found in [2].

Classification of Warranties

A taxonomy for warranty classification was first proposed by Blischke and Murthy [3]. Figure 1 is a chart showing the classification scheme. The first criterion for classification is whether or not the manufacturer is required to carry out further product development (for example, to improve product reliability) subsequent to the sale of the product. Policies that do not involve such further product development can be further divided into two groups: the first consisting of policies applicable for single item sales and the second consisting of policies used in the sale of groups of items (called *lot* or *batch sales*).

Policies in the first group can be subdivided into two subgroups based on whether the policy is renewing or nonrenewing. For renewing policies, the warranty period begins anew with each replacement, while for nonrenewing policies the replacement (or repaired) item assumes the remaining warranty time of the item it replaced. A further subdivision comes about in that warranties may be classified as "simple" or "combination". The commonly used consumer warranties featuring free-replacement warranty (FRW) and replacement at pro rata warranties (PRW) (discussed in the next section) are simple policies. A combination policy is a simple policy combined with

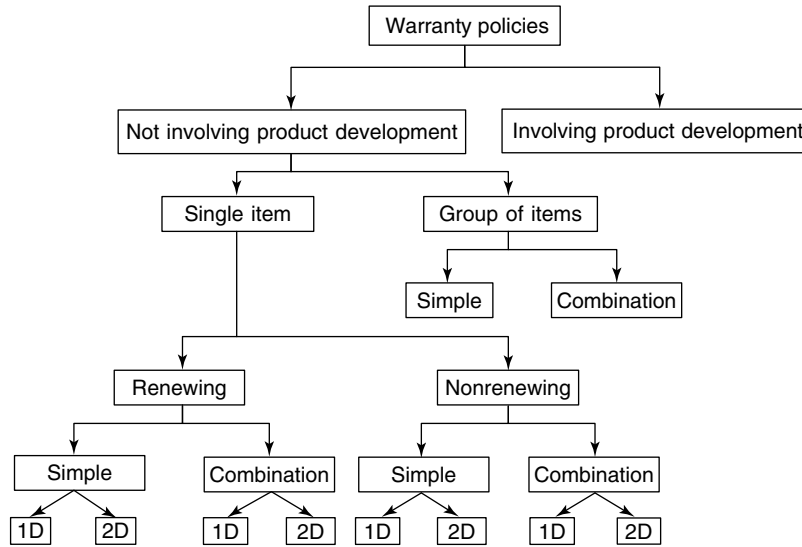


Figure 1 Classification of warranties

some additional features or a policy that combines the terms of two or more simple policies. Each of these four groupings can be further subdivided into two subgroups based on whether the policy is one dimensional or two (or higher) dimensional. A one-dimensional policy is one that is most often based on either time or age of the item, but could instead be based on usage. An example is a 1-year warranty on television with the option for buyers to purchase extended warranty coverage. In contrast, a two-dimensional policy is based on time or age as well as usage. An example is an automobile warranty with a time limit of 3 years and a usage limit of 30 000 miles. The warranty expires when one of the limits is exceeded.

Some Simple Warranty Policies

Three simple warranty policies are defined in this section. The cost analysis for each of these is given in later sections.

Policy 1: one-dimensional (1D) nonrenewing free-replacement warranty (FRW)

The manufacturer agrees to repair or provide replacements for failed items free of charge up to a time W from the time of the initial purchase. The warranty expires at time W after purchase.

Typical applications of this warranty are consumer products, ranging from inexpensive items (*see Reliability of Consumer Goods with “Fast Turn Around”*) such as alarm clocks and small radios to relatively expensive repairable items (*see Degradation and Shock Models*) such as automobiles and nonrepairable items such as microchips and other electronic components.

Policy 2: one-dimensional nonrenewing pro rata warranty (PRW)

The manufacturer agrees to refund a fraction of the purchase price if the item fails before time W from the time of the initial purchase. The buyer is not constrained to buy a replacement item.

Under the PRW, the refund given depends on the age of the item at failure. In the most common form of PRW, the refund is determined by the linear relationship $q(x) = [(W - x)/W]P$, where W is the warranty period, x is the age at failure, P is the sale price, and $q(\cdot)$, called the *rebate function*, is the amount of refund. Typically, these policies are offered on relatively inexpensive nonrepairable products such as batteries, tires, ceramics, etc.

Policy 3: two-dimensional (2D) nonrenewing FRW

The manufacturer agrees to repair or provide a replacement for failed items free of charge up to a time W from the time of initial purchase (limit on

time) or up to a usage U (limit on usage), whichever occurs first.

Product Reliability

The modeling of first failures is different from that of subsequent failures. In the latter case, the result depends on whether the failed item is repaired or replaced by a new item.

One-Dimensional Failure Modeling

First Failure. The time to first failure, T , is a random variable that can assume values in the interval $[0, \infty)$, with distribution function

$$F(t) = P(T \leq t), \quad 0 \leq t < \infty \quad (1)$$

The *density function* (if $F(t)$ is differentiable) is given by $f(t) = dF(t)/dt$. The reliability (of the item) is given by the survivor function

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) \quad (2)$$

The failure rate function (or hazard function) (see **Hazard and Hazard Ratio**) $h(t)$ associated with $F(t)$ is defined as

$$h(t) = \frac{f(t)}{1 - F(t)} \quad (3)$$

Subsequent Failures. When a repairable component fails, it can either be repaired or replaced by a new component. In the case of a nonrepairable component, the only option is to replace the failed component by a new item. Since failures occur in an uncertain manner, the number of components needed over a given time interval is a non-negative, integer-valued random variable. The distribution of this variable depends on the failure distribution of the component, the actions (repair or replace) taken after each failure, and, if repaired, the type of repair. The number of failures $N(t)$ over the interval $[0, t)$, starting with a new component at $t = 0$, can be modeled as a *counting process*.

Nonrepairable item. In the case of a nonrepairable component, every failure results in the replacement of the failed component by a new item. If the time to replace a failed component by a new one is

small relative to the mean time to failure (MTTF), then it can be ignored. In this case, the number of failures (replacements) over time (see **Recurrent Event Data**) is given by a *renewal process* and the mean number of failures over $[0, t)$ is given by the *renewal integral equation*

$$M(t) = F(t) + \int_0^t M(t - t') dF(t') \quad (4)$$

Repairable item. There are different types of repairs (see **Repairable Systems Reliability**). With *minimal repair*, the failure rate after repair is the same as the failure rate of the item immediately before it failed [4]. With *imperfect repair*, the failure rate after repair is less than that just before failure but not as good as that for a new component.

With minimal repair and the repair time negligible in relation to the mean time between failures, failures over time occur according to a *nonhomogeneous Poisson process* [5]. The mean number of failures over $[0, t)$ is given by

$$E[N(t)] = \Lambda(t) = \int_0^t h(x) dx \quad (5)$$

For the case of imperfect repair, the analysis is more complicated. See [6] and references therein for additional details.

Two-Dimensional Failure Modeling

There are two approaches to modeling failures in the 2D case.

Approach 1 (1D Approach). The two-dimensional problem is effectively reduced to a one-dimensional problem by treating usage as a function of age.

First failure. Let $Z(t)$ denote the usage of the item at age t and assume that it is a linear function of t of the form

$$Z(t) = Rt \quad (6)$$

where $R, 0 \leq R < \infty$, represents the usage rate and is a non-negative random variable with a distribution function $G(r)$ and density function $g(r)$.

The hazard function, conditional on $R = r$, is denoted by $h(t|r)$. Various forms of $h(t|r)$ have been proposed, including the linear case given by

$$h(t|r) = \theta_0 + \theta_1 r + \theta_2 t + \theta_3 Z(t) \quad (7)$$

The conditional distribution function of time to first failure is given by

$$F(t|r) = 1 - \exp \left\{ - \int_0^t h(u|r) du \right\} \quad (8)$$

On removing the conditioning, the distribution function for the time to first failure is given by

$$F(t) = \int_0^\infty \left(1 - \exp \left\{ - \int_0^t h(u|r) du \right\} \right) \times g(r) dr \quad (9)$$

Subsequent failures. Under minimal repair, the failures over time occur according to a nonhomogeneous Poisson process [7] with intensity function $\lambda(t|r) = h(t|r)$. Under perfect repair, the failures occur according to the renewal process associated with $F(t|r)$.

The bulk of the literature deals with a linear relationship between usage and age. See, for example [1, 8, 9]. For applications, see [10], which deals with motorcycle failures under warranty, and [8, 11], which deal with automobile warranty data analysis.

Approach 2 (2D Approach).

First failure. Let T and X denote the item age and usage at first failure. Both of these are random variables. T and X are modeled jointly by the bivariate distribution function

$$F(t, x) = P(T \leq t, X \leq x); \quad t \geq 0, x \geq 0 \quad (10)$$

The survivor function is given by

$$\begin{aligned} \bar{F}(t, x) &= \Pr(T > t, X > x) \\ &= \int_t^\infty \int_x^\infty f(u, v) dv du \end{aligned} \quad (11)$$

If $F(t, x)$ is differentiable in both arguments, the bivariate failure density function is given by

$$f(t, x) = \frac{\partial^2 F(t, x)}{\partial t \partial x} \quad (12)$$

Subsequent failures. In the case of nonrepairable items, failed items are replaced by new ones. In this case, failures occur according to a 2D renewal function [4] and the expected number of failures over $[0, t) \times [0, x)$ is given by the solution of the two-dimensional integral equation

$$\begin{aligned} M(t, x) &= F(t, x) + \int_0^t \int_0^x M(t - u, x - v) \\ &\quad \times f(u, v) dv du \end{aligned} \quad (13)$$

Warranty Cost Analysis

In a simple risk analysis, a manufacturer may consider only short- and long-term risks. In warranty analysis, these are represented by (a) expected cost per unit sold and (b) expected cost over the product life cycle (*see Subjective Expected Utility*).

Expected Cost per Unit Sold

Whenever a failed item is returned for rectification action under warranty, the manufacturer incurs various costs – handling, material, labor, facilities, etc. These costs are random variables. The total cost of servicing all warranty claims for an item over the warranty period is the sum of a random number of such individual costs, since the number of claims over the warranty period is also random. The expected warranty service cost per unit is needed in deciding the selling price of the item.

Expected Cost over the Life Cycle of a Product

From a marketing perspective, the product life cycle, L , is the period from the instant a new product is launched to the instant it is withdrawn from the market. Over the product life cycle, product sales (first and repeat purchases) occur over time in a dynamic manner. The manufacturer must service the warranty claims with each such sale. In the case of products sold with one-dimensional nonrenewing warranty, this period is $L + W$, where W is the warranty period.

The expected cost incurred over $[t, t + \delta t)$ depends on the sales over the interval $[t - \psi, t)$, where

$$\psi = \max(0, t - W) \quad (14)$$

Sales over time can be modeled either by a continuous function (appropriate for large volume sales) or a point process (appropriate for small volume sales). Assume that the sales rate is given by $s(t)$, $0 \leq t \leq L$, so that total sales S over the life cycle is

$$S = \int_0^L s(t) dt \quad (15)$$

Assumptions

In developing tractable and reasonable warranty cost models, various assumptions regarding the process are made. These typically include the following:

1. All consumers are alike in their usage. This can be relaxed by dividing consumers into two or more groups of similar usage patterns.
2. All items are statistically similar. This can be relaxed by including two types of items (conforming and nonconforming) to take into account quality variations in manufacturing.
3. Whenever a failure occurs, it results in an immediate claim. Relaxing this assumption involves modeling the delay time between failure and claim.
4. All claims are valid. This can be relaxed by assuming that a fraction of the claims are invalid, either because the item was used in a mode not covered by the warranty or because it was a bogus claim.
5. The time to rectify a failed item (either through repair or replacement) is sufficiently small in relation to the mean time between failures that it can be approximated as being zero.
6. The manufacturer has the logistic support (spares and facilities) needed to carry out rectification actions without delays.
7. The cost of each rectification under warranty (repair or replacement) is modeled by the following cost parameters:
 C_m : cost of replacing a failed unit by a new unit (manufacturing cost per unit)
 C_h : handling cost per claim
 C_r : average cost of each minimal repair.
8. The product life cycle, L , is modeled as a deterministic variable. One can easily relax this assumption and treat it as a random variable with a specified distribution.

9. All model parameters (costs and the various distributions involved) are known.

Modeling Warranty Cost per Unit Sold

In this section and in the following one, the cost analysis for Policies 1–3 is provided. Numerical examples for Policies 1 and 2 are given in the section titled “Examples”. Additional examples for these as well as the analysis and examples for many other warranties may be found in [1, 2, 8–11].

Cost Analysis of Policy 1

Replace by New. Since failures under warranty are rectified instantaneously through replacement by new items, the number of failures over the period $[0, t)$, $N(t)$, is characterized by a renewal process with time between renewals distributed according to $F(t)$. As a result, the cost $\tilde{C}(W)$ to the manufacturer over the warranty period W is a random variable given by

$$\tilde{C}(W) = (C_m + C_h)(N(W)) \quad (16)$$

The expected number of failures over the warranty period is

$$E[N(W)] = M(W) \quad (17)$$

where $M(t)$ is the renewal function given by equation (4). As a result, the expected warranty cost per unit to the manufacturer is

$$C(W) = (C_m + C_h)M(W) \quad (18)$$

Minimal repair. Failures over the warranty period occur according to a nonhomogeneous Poisson process with intensity function equal to the failure rate $r(t)$, and the expected warranty cost to the manufacturer is

$$C(W) = (C_r + C_h) \int_0^W r(x) dx \quad (19)$$

Cost Analysis of Policy 2

In the case of linear rebate function

$$q(x) = \begin{cases} (1 - x/W)P & 0 \leq x < W \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

the warranty cost to the manufacturer is

$$\tilde{C}(W) = C_h + q(T) \quad (21)$$

where T is the lifetime of the item supplied. As a result, the expected cost per unit to the manufacturer is

$$\begin{aligned} C(W) &= E(\tilde{C}(W)) \\ &= C_h F(W) + \int_0^W q(x) f(x) dx \end{aligned} \quad (22)$$

Using equation (20) in equation (22) and carrying out the integration, we have

$$C(W) = C_h F(W) + P \left[F(W) - \frac{\mu_W}{W} \right] \quad (23)$$

where

$$\mu_W = \int_0^W t f(t) dt \quad (24)$$

is the *partial expectation* of T .

Cost Analysis of Policy 3

As mentioned in the section titled “Approach 1 (1D Approach)”, there are two approaches to modeling failures. Approach 1 is more appropriate for the case of repairable products (with minimal repair) and Approach 2 for nonrepairable products.

Approach 1. The product is repairable and the manufacturer rectifies all failures under warranty through minimal repair. In this case, the distribution of failures over the warranty period can be characterized by an intensity function $\lambda(t|r)$. Conditional on $R = r$, we assume a linear intensity function given by equation (7).

Note that conditional on $R = r$, the warranty expires at time W if the usage rate r is less than γ , where

$$\gamma = U/W \quad (25)$$

and at time t_r given by

$$t_r = U/r \quad (26)$$

if the usage rate $r > \gamma$. The expected numbers of failures over the warranty period for these two cases are given by

$$E(N(W, U)|r) = \int_0^W \lambda(t|r) dt \quad (27)$$

and

$$E(N(W, U)|r) = \int_0^{t_r} \lambda(t|r) dt \quad (28)$$

respectively. On removing the conditioning, we have the expected number of failures during warranty and, using this, we find the expected warranty cost to be

$$\begin{aligned} C(W, U) &= [C_r + C_h] \left\{ \int_0^\gamma \left[\int_0^W \lambda(t|r) dt \right] g(r) dr \right. \\ &\quad \left. + \int_\gamma^\infty \left[\int_0^{t_r} \lambda(t|r) dt \right] g(r) dr \right\} \end{aligned} \quad (29)$$

Approach 2. The product is nonrepairable and the manufacturer rectifies all failures under warranty through replacement by new. In this case, the number of failures over the warranty region is given by a 2D renewal process and the expected warranty cost is given by

$$C(W, U) = (C_m + C_h)M(W, U) \quad (30)$$

where $M(W, U)$ is obtained from equation (13).

Examples

The examples that follow illustrate the cost analysis of Policies 1 and 2. In both cases, $F(W)$ is assumed to have a simple form, the exponential distribution, which is appropriate when failures occur at a constant rate. In practice, this is very often not the case, and the Weibull and any of several other alternative distributions are more appropriate [1, 2].

Policy 1. A cell phone is sold with nonrenewing FRW with warranty period W . Failed items are scrapped and replaced by new phones. Suppose that time to failure is exponentially distributed with parameter λ , i.e., $F(t) = 1 - e^{-\lambda t}$ for $t \geq 0$. Then [4] the MTTF is $\mu = 1/\lambda$ and $M(t) = \lambda t$. We further assume that the manufacturer’s total cost of processing a claim is $C_m + C_h = \$37$. From equation (18), the manufacturer’s expected cost of warranty is

$\$37\lambda W$. Suppose the *MTTF* of the phone is 4 years (so $\lambda = 0.25$). If a 90-day warranty ($W = 0.25$ years) is offered, then the expected warranty cost is \$2.31 per unit sold. For a 6-month warranty, this cost would be \$4.63. If the phone is priced at \$100, these would usually be considered reasonable warranty costs.

Policy 2. An automobile battery is sold with a pro rata warranty. For purposes of illustration, we assume that it is a rebate warranty (Policy 2) with linear rebate function and that the time to failure of the battery is exponentially distributed. The model for manufacturer's cost of warranty is given in equation (23). For the exponential distribution [1, p. 174], $\mu_W = \lambda^{-1}(1 - (1 + \lambda W)e^{-\lambda W})$. Suppose that the *MTTF* of the battery is 5.77 years (which corresponds to a median lifetime of 4 years) and that $W = 2$. Then $F(2) = 0.293$ and $\mu_W = 0.275$.

If $C_h = \$10$ and the selling price is $P = \$100$, the cost of warranty is $\$10(0.293) + 100(0.155) = \18.43 . The corresponding cost if $W = 1$ is \$9.77. Both of these costs are very substantial. The cost can be reduced in a number of ways – increasing reliability, increasing the *MTTF*, decreasing the warranty period, selecting a different warranty policy, and so forth.

Warranty Cost over Life Cycle

Consider the case in which the product is non-repairable and the life cycle L exceeds the warranty period W . Since the manufacturer must provide a refund or replacements for items that fail before reaching age W , and since the last sale occurs at or before time L , the manufacturer has an obligation to service warranty claims over the interval $[0, L + W]$. Life cycle costs for Policies 1 and 2 will be considered. Some results for Policy 3 can be found in Blischke and Murthy [1, Sections 8.3.1 and 8.4.1].

Cost Analysis for Policy 1

Since failed items are replaced by new ones over the warranty period, the demand for spares in the interval $[t, t + \delta t)$ is due to failure of items sold in the period $[\psi, t)$, where ψ is given by equation (14). It can be shown (details of the derivation can be found

in Chapter 9 of [1]) that the expected demand rate for spares at time t , $\rho(t)$, is given by

$$\rho(t) = \int_{\psi}^t s(x)m(t-x) dx \quad (31)$$

where $m(t)$ is the renewal density function given by

$$m(t) = \frac{dM(t)}{dt} = f(t) + \int_0^t m(t-t')f(t') dt' \quad (32)$$

The expected total number of spares required to service the warranty over the product life cycle, *ETNS*, is given by

$$ETNS = \int_0^{L+W} \rho(t) dt \quad (33)$$

and the total expected warranty cost over the product life cycle is given by

$$LC(W, L) = (C_m + C_h) \int_0^{W+L} \rho(t) dt \quad (34)$$

Cost Analysis for Policy 2

Under Policy 2 the manufacturer refunds a fraction of the sale price upon failure of an item within the warranty period. In order to carry this out, the manufacturer must set aside a fraction of the sale price. This is called *warranty reserving*. We assume a linear rebate function. The rebate over the interval $[t, t + \delta t)$ is due to failure of items that are sold in the interval $[t - \psi, t)$, where ψ is given by equation (14). Let $v(t)$ denote the expected refund rate (i.e., the amount refunded per unit time) at time t . Then it is easily shown that

$$v(t) = P \int_{\psi}^t s(x) \left(\frac{t-t'}{W} \right) f(t-t') dt' \quad (35)$$

for $0 \leq t \leq (W + L)$. (For details, see Chapter 9 of [1].) The expected total reserve needed to service the warranty over the product life cycle is given by

$$LC(W, L) = \int_0^{W+L} v(t) dt \quad (36)$$

Concluding Remarks

The cost analysis of many other types of warranty policies can be found in [1] and in more recent

papers; references to some of these can be found in [12]. This type of analysis is carried out during the front-end (feasibility) stage of the new product development process. Model selection and parameter values are based on similar products, expert judgment, etc. For more on this, see [13], which looks at warranty management from a product life cycle perspective. Once the product is launched, warranty claims provide data for assessing the models and updating the warranty cost estimates. For more on this, relevant references can be found in [12, 13].

End Notes

This article is a revised version of “Warranty Cost Analysis,” by D. N. Prabhakar Murthy and Wallace R. Blischke, which appeared in *Encyclopedia of Statistics in Quality and Reliability*, copyright John Wiley & Sons, Limited, 2006. Reprinted with permission.

References

- [1] Blischke, W.R. & Murthy, D.N.P. (1994). *Warranty Cost Analysis*, Marcel Dekker, New York.
- [2] Blischke, W.R. & Murthy, D.N.P. (1996). *Product Warranty Handbook*, Marcel Dekker, New York.
- [3] Blischke, W.R. & Murthy, D.N.P. (1992). Product warranty management – I: a taxonomy for warranty policies, *European Journal of Operational Research* **62**, 127–148.
- [4] Blischke, W.R. & Murthy, D.N.P. (2000). *Reliability*, John Wiley & Sons, New York.
- [5] Murthy, D.N.P. (1991). A note on minimal repair, *IEEE Transactions on Reliability* **40**, 245–246.
- [6] Doyen, L. & Gaudoin, O. (2004). Classes of imperfect repair models based on reduction of failure intensity or virtual age, *Reliability Engineering and System Safety* **84**, 45–56.
- [7] Ross, S.M. (1983). *Stochastic Processes*, John Wiley & Sons, New York.
- [8] Lawless, J., Hu, J. & Cao, J. (1995). Methods for estimation of failure distributions and rates from automobile warranty data, *Lifetime Data Analysis* **1**, 227–240.
- [9] Gertsbakh, I.B. & Kordonsky, K.B. (1998). Parallel time scales and two-dimensional manufacturer and individual customer warranties, *IIE Transactions* **30**, 1181–1189.
- [10] Iskandar, B.P. & Blischke, W.R. (2002). Reliability and warranty analysis of a motorcycle based on claims data, in *Case Studies in Reliability and Maintenance*, W.R. Blischke & D.N.P. Murthy, eds, John Wiley & Sons, New York, Chapter 27.
- [11] Kalbfleisch, J.D., Lawless, J.F. & Robinson, J.A. (1991). Methods for the analysis and prediction of warranty claims, *Technometrics* **33**, 273–285.
- [12] Murthy, D.N.P. & Djamaludin, I. (2002). Product warranty – a review, *International Journal of Production Economics* **79**, 231–260.
- [13] Murthy, D.N.P. & Blischke, W.R. (2005). *Warranty Management and Product Manufacture*, Springer-Verlag, London.

Related Articles

Common Cause Failure Modeling

Distributions for Loss Modeling

Probabilistic Design

Product Risk Management: Testing and Warranties

D.N. PRABHAKAR MURTHY AND WALLACE R.
BLISCHKE

Water Pollution Risk

Water is a necessity for all forms of life. Water is used by plants and animals to sustain their lives. In addition, man needs water for domestic purposes (food, cooking, drinking, cleaning, etc.) and for his industrial and agricultural activities. Vegetation not only uses water to carry food materials but uses it as a nutrient for its growth. Water acts also as an environment in which aquatic plants and animals live, and it is the bearer of nutrients that produce organic materials from inorganic materials through primary production that forms the base of the food chain. To a large extent, water pollution is mainly the consequence of this diverse use because water becomes the recipient of human and animal waste. For example, it has been estimated by Gray [1] that approximately 1 million cubic meter of domestic and 7 million cubic meter of industrial wastewater is produced daily in the United Kingdom. In addition to this volume, runoff from paved areas and roads and infiltration water produce over 20 million cubic meter of wastewater requiring treatment each day. To cope with this immense volume, the commitment of substantial technical and financial resources to construct and operate large number of sewage and wastewater treatment plants is required. The list of threats to water quality related to human and ecosystem health is immense. Among the chief threats it is possible to identify the following: pathogens, biological toxins, chemical toxics, nutrients, acidification, endocrine disrupting substances, municipal and industrial waste, and water quantity changes affecting water quality due to climate change, diversions, and extreme events.

It is likely that humans have always been concerned about the quality of the water they use to meet their domestic needs. However, water contamination and water protection have become a major concern since John Snow's [2] famous conclusion that water from a public pump was the cause of the cholera epidemic in London in 1854. This was reported almost 30 years prior to the discovery of cholera's causative agent by Robert Koch in Egypt in 1883. The important aspect of John Snow's discovery is that a substance does not need to be necessarily visible or without discernible taste or odor to be identified as harmful. Indeed, the most serious contaminants (viruses and toxic substances) are those that are not

directly detectable by human senses. Technological advances in measurements have been responsible for discovering the presence of many unknown hazards and quantifying their levels in various environmental media (air, water, and soil). Examples include pathogenic bacteria, viruses, and trace toxic contaminants. With the awareness of these pollutants entering water systems, risk analysis has become an essential tool to detect, identify, quantify, and manage their existing and potential threats. As a result there is mounting evidence that exposure to higher levels of water pollution is associated with adverse effects on the health of human and all forms of aquatic life.

This article is not intended to discuss the risks from all water pollution sources; it is restricted to the risks posed particularly to human health. For this purpose, a pollutant is defined as any substance or agent when introduced to a water system at a level that causes or expected to cause adverse consequences to any of the water's beneficial functions. The goals of water pollution risk analysis are to assess, predict, and control existing and potential adverse impact of water contamination. The emphases here are on the quantitative statistical aspects. Specifically, the discussion centers on two classes of pollution types: persistent toxic substances and microbiological contamination (MC).

Persistent Toxic Substances (PTS)

Persistent toxic substances (PTS) are chemical agents that are able to retain their molecular integrity and hence their physical, chemical, and functional characteristics in the aquatic environment. Once a PTS chemical enters an environment, it will be transported and distributed within that environment for a long period of time. The Great Lakes Water has been found to contain thousands of PTS, including polychlorinated biphenyls (PCBs), (*see* **Polychlorinated Biphenyls**) dichloro-diphenyl-trichloroethane (DDT), mirex and mercury [3]. The health of millions of Canadian and United States citizens is at risk because they depend on the Lakes as a source for drinking water and for their recreational activities. Furthermore, these substances tend to bioaccumulate through the food chain and become a serious threat to those who diet on fish caught from the Lakes' contaminated water. Some PTS agents such as mercury, lead, and cadmium (*see* **Lead**; **Methylmercury**) occur naturally in the environment, while others are the product

of anthropogenic activities. Canada and United States have the goal of zero discharge of PTS in the Great Lakes but this will not eliminate the chemicals from the water: the historic burden of these substances in the Lakes' contaminated sediment and the long-range airborne transport of these chemicals helps motivate their primary moniker, "persistent". The approach used by Canada to control the impacts of PTS is the issuance of fish consumption advisories [1] that take into account the variations in the concentrations of the contaminants in different-sized fish of various species.

Figure 1 shows the empirical cumulative distribution function (EDF) of the concentration of total PCBs in the tissues of samples of trout caught from Lake Ontario and Superior in the summer of 2001 along with the Ontario government's [4] critical level advisory for PCBs consumption. This figure illustrates two points:

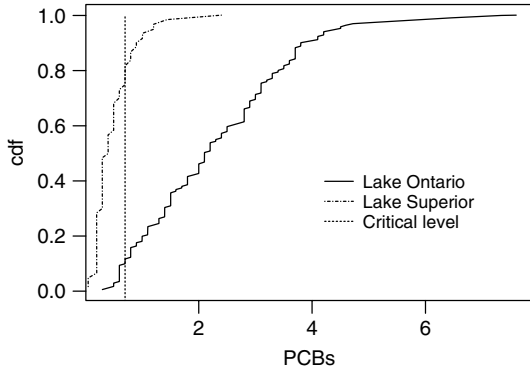


Figure 1 The EDF of PCBs concentration in fish tissue from Lakes Ontario and Superior

1. The concentration level is higher in Lake Ontario than that found in fish from Lake Superior, highlighting the persistence nature of PCBs and their accumulation from all upstream sources before reaching Lake Ontario.
2. There is a greater risk associated with consuming fish from Lake Ontario than that associated with consuming fish at the same age caught in Lake Superior.

Since it is impossible for individuals to measure levels of contaminants in the fish available for consumption, it is therefore necessary to relate the level of contamination to easily and inexpensive measurable fish characteristics. There are three possible surrogate characteristics that meet these requirements: age, weight, and length. These are positively correlated with the level of the bioaccumulation of contaminants. Usually, the advisory critical level is stated in terms of weight and/or length because the latter are easy to measure. Figure 2 shows the association between age and PCBs level in Lake Ontario. The advisory usually includes the specification of the maximum number of meals allowed during a standard time period and size of portion to be consumed during a meal. In addition, the advisory frequently states different restriction limits for different population group: for example, lower limits may be stated for children and pregnant women.

To describe a model for representing the dependence of PCBs concentration on age for Lake Ontario, let x_t be the observed concentration of PCBs in a fish at age t and let X_t be its corresponding random variable, then a possible model for X_t is

$$X_t = b(1 - \exp(-\lambda t^m)) + \varepsilon_t \quad (1)$$

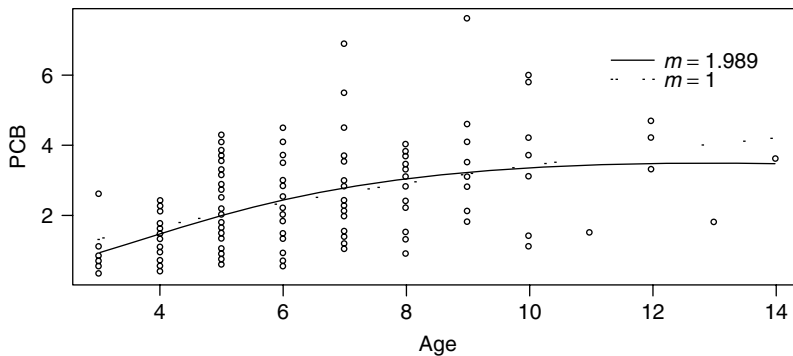


Figure 2 Measured and modeled PCBs concentration in fish from Lake Ontario

where b , λ , and m are unknown parameters and ε_t is a zero-mean random variable with constant variance. Note that b represents the maximum PCBs accumulation and λ represents the rate of accumulation. It should be noted that at $t = 0$, X_t will have mean zero, while when $m = 0$, the mean is a constant, independent of age. Figure 2 shows the fitted model when $m = 1$ (dotted curve) and when m is estimated along with the other two parameters (solid curve). The results indicate that m is significantly different from 1. The model could also be used to generate the PCBs distribution for a specified age. Figure 3 shows the model-based simulation of this distribution for ages 3, 7, and 14 years. The figure can be used to determine the degree of risk associated with each age and is thus useful for setting an appropriate advisory consumption level.

Microbiological Contaminants (MC)

The other example we consider is related to the risk associated with drinking and recreational use when the water source is contaminated by pathogenic organisms that cause illness and sometimes death in the exposed population. Examples of potentially lethal diseases caused by contaminated water include typhoid fever, typhus, cholera, and dysentery. Waterborne pathogens can pose threats to drinking water supplies, recreational waters, source waters for agriculture, and aquaculture, as well as to aquatic

ecosystems and biodiversity. Disease outbreaks are widespread in underdeveloped countries but also still occur in developed countries [5] despite the great efforts and resources taken to prevent microbial pollution and safeguard water supplies. Examples of outbreaks in developed countries are as follows:

1. The recent and fatal epidemic that occurred in Walkerton, Ontario, Canada, in May 2000 where, in that small town of 4800 residents, 2200 residents became seriously ill and 7 died. As a result a public inquiry was established by the Government of Ontario to address the causes of the outbreak. Mr. Justice Dennis O'Connor [6] headed the inquiry and his report summarizes the deficiencies involved from protecting water sources to the delivery of safe drinking water to the residents (O'Connor [6]). Hrudey *et al.* [5] compare the Walkerton epidemic to other outbreaks in the developed countries.
2. The Milwaukee outbreak that resulted in 54 deaths and over 400 000 cases of illness (Hoxie *et al.* [7]).
3. The spread of cholera across Iraq after it was detected earlier in the north of the country as reported in September 21, 2007 by *The New York Times*.

The catalog of waterborne pathogens is long, and it includes many well-known organisms as well as far larger numbers of more obscure organisms.

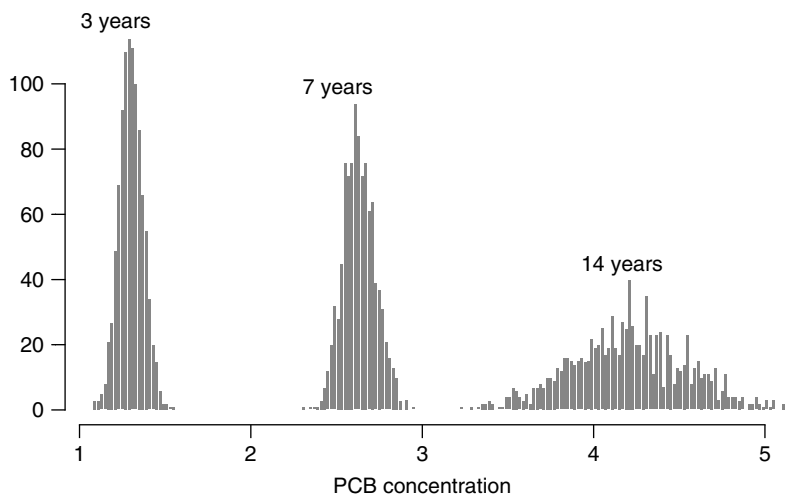


Figure 3 Simulated PCBs distribution in the trout tissue from Lake Ontario

Table 1 The 1986 EPA criteria for recreational waters

Water type	Indicator	Geometric mean	Single sample maximum
Marine	<i>Enterococci</i>	35	104
Fresh	<i>Enterococci</i>	33	61
	<i>E. coli</i>	126	235

Waterborne pathogens include viruses (e.g., hepatitis A, poliomyelitis); bacteria (e.g., cholera, typhoid, coliform organisms); protozoa (e.g., *cryptosporidium*, *amebae*, *giardia*); worms (e.g., *schistosomia*, guinea worm); etc. To enumerate all waterborne harmful organisms would require complex, time-consuming, and expensive procedures to be performed on each water sample analyzed. These difficulties led to the development of microbiological fecal pollution indicators. Coliform bacteria group was adopted as indicator in 1914 (Pipes *et al.* [8] and Olivieri *et al.* [9]) by the United States for monitoring drinking water. The use of coliforms as an index of water quality has been criticized, and currently researchers in Canada and United States are investigating the potential use of more appropriate indicators. The maximum microbiological contaminant levels (MCLs) approach was adopted by the United States Environmental Protection Agency (US EPA) as a legal requirement [6]. The MCLs state (Pipes [10]) that if a membrane filters method is used,

1. the arithmetic mean coliform count of all standard samples examined per month shall not exceed 1/100 ml and
2. the number of coliform bacteria shall not exceed 4/100 ml in
 - (a) more than one sample when less than 20 samples are examined or
 - (b) more than 5% of the samples if at least 20 samples are examined.

When the coliform count in a single sample exceeds 4/100 ml, at least two consecutive daily check samples shall be collected and examined until the results from two consecutive samples show less than one coliform bacterium per 100 ml.

El-Shaarawi and Marsalek [11] discussed various sources of uncertainties associated with the application of guidelines and criteria in water quality monitoring. Carbonez *et al.* [12] assessed the performance of the two rules of MCLs stated above and

concluded that it is sufficient to use only the first rule that requires the placing of an upper limit on the mean coliform count.

The 1986 EPA for primary contact recreational use criteria are listed in Table 1. It should be noted that the two rules are also involved here: the geometric mean rule and a single sample rule. Also there are two indicator organisms (*enterococci* and *Escherichia coli*) in the case of freshwater but for marine waters only *enterococci* is used. These limits were derived based on the epidemiological studies reported by Cabelli [13]. Another difference between the MCLs, which is designed for drinking water and for recreational water, is the placement of an upper limit on the geometric mean instead of setting the limit on the arithmetic mean.

A simple way to evaluate the need for the two rules is to assume that the lognormal distribution can be used as a model for the density of the indicator organism in water samples (El-Shaarawi [14]). It should be mentioned that the negative binomial distribution is also frequently used as a model for bacterial counts (El-Shaarawi [15] and El-Shaarawi *et al.* [16]). Let X_i be the random variable representing the log density, which is assumed to have a normal distribution with mean μ and standard deviation σ . The two rules require that when n observations are available their mean $\bar{X}$ should be smaller than the threshold b and the largest observation $X_{(n)}$ should be smaller than a .

Let $\xi = \frac{b-\mu}{\sigma}$, $\eta = \frac{a-\mu}{\sigma}$, and $g = \frac{\varphi(\xi)}{\Phi(\xi)}$, where $\varphi(\xi)$ and $\Phi(\xi)$ are the density and distribution functions of the standard normal. Then, as a criterion for comparing the two rules we use the log ratio of the probabilities of declaring that the water source is unfit for recreational activities, i.e.,

$$\rho = \log \frac{1 - \text{prob}(X_{(n)} < b)}{1 - \text{prob}(\bar{X} < a)} = \log \frac{1 - \{\Phi(\xi)\}^n}{1 - \Phi(\sqrt{n}\eta)} \quad (2)$$

The above criteria could be used to estimate the expected level of the indicator organism when the two

Table 2 Expected values of the density of indicator organism as a function of sample size

<i>N</i>	5	10	15	20	25	30	35	40	45	50
Mean	43.67	39.29	37.72	36.87	36.33	35.95	35.66	35.44	35.26	35.11

rules are equally effective in detecting the existence of water quality problems at a specified level. For example, if it is desirable to estimate the expected value that corresponds to a 0.95 level for various sample sizes, then we need to compute μ from the following expression

$$\lambda = \frac{\xi}{\eta} = \frac{\log(b) - \mu}{\log(a) - \mu} = \frac{\Phi^{-1}(0.95^{1/n})\sqrt{n}}{\Phi^{-1}(0.95)} \quad (3)$$

for various values of n . Then, the corresponding indicator level is $\exp(\mu)$. Table 2 gives these values for several sample sizes.

References

- [1] Gray, N.F. (2004). *Biology of Wastewater Treatment*, 2nd Edition, Imperial College Press.
- [2] Snow, J. (1855, 1936). *On the Mode of Communication of Cholera*, in *Snow on Cholera*, 2nd Edition, The Commonwealth Fund, New York.
- [3] The International Joint Commission (2000). *Tenth Biennial Report on Great Lakes Water Quality*, Windsor.
- [4] Ontario Ministry of the Environment (2007–2008). *Guide to Eating Ontario Sport Fish*, Toronto.
- [5] Hrudey, S.E., Payment, P., Huck, P.M., Gillham, R.W. & Hrudey, E.J. (2003). A fatal waterborne disease epidemic in Walkerton, Ontario: comparison with other waterborne outbreaks in the developed world, *Journal of Water Science and Technology* **47**(3), 7–14.
- [6] O'Connor, D.R. (2002). *Report of the Walkerton Inquiry*, Ontario Ministry of the Attorney General, Toronto, p. 504.
- [7] Hoxie, N.J., Davis, J.P., Vergeront, J.M., Nashold, R.D. & Blair, K.A. (1997). Cryptosporidiosis-associated mortality following a massive waterborne outbreak in Milwaukee, Wisconsin, *American Journal of Public Health* **87**(12), 2032–2035.
- [8] Pipes, W.O., Bordner, R., Christian, R.R., El-Shaarawi, A.H., Fuhs, G.W., Kennedy, H., Means, E., Moser, R. & Victoreen, H. (1983). Monitoring of microbiological quality, in *Assessment of Microbiology and Turbidity Standards for Drinking Water*, EPA 570/9-83-001, P.S. Berger & Y. Argaman, eds, Office of Drinking Water, U.S. EPA, Washington, DC, pp. II-1–III-45.
- [9] Olivieri, V.P., Ballesterio, J., Cabelli, V.J., Chamberlin, C., Cliver, D., DuFour, A., Ginsberg, W., Healy, G., Highsmith, A., Read, R., Reasoner, D. & Tobin, R. (1983). Measurements of microbiological quality, in *Assessment of Microbiology and Turbidity Standards for Drinking Water*, EPA 570/9-83-001, P.S. Berger & Y. Argaman, eds, Office of Drinking Water, U.S. EPA, Washington, DC, pp. II-1–II-42.
- [10] Pipes, W.O. (1990). Microbiological methods and monitoring of drinking water, in *Drinking Water Microbiology*, G.A. McFeters, ed, Springer-Verlag.
- [11] El-Shaarawi, A.H. & Marsalek, J. (1999). Guidelines for indicator bacteria in waters: uncertainties in applications, *Environmetrics* **10**, 521–529.
- [12] Carbonez, A., El-Shaarawi, A.H. & Teugels, J.L. (1999). Maximum microbiological contaminant levels, *Environmetrics* **10**, 79–86.
- [13] Cabelli, V.J. (1983). *Health Effects Criteria for Marine Waters*, EPA-600/1-80-031, US Environmental Protection Agency, Research Triangle Park.
- [14] El-Shaarawi, A.H. (2003). Lognormal distribution, in *Encyclopedia of Environmetrics*, A.H. El-Shaarawi & W. Piegorsch, eds, John Wiley & Sons, Chichester, Vol. 2, pp. 1180–1183.
- [15] El-Shaarawi, A.H. (2003). Negative binomial distribution, in *Encyclopedia of Environmetrics*, A.H. El-Shaarawi & W. Piegorsch, eds, John Wiley & Sons, Chichester, Vol. 3, pp. 1383–1388.
- [16] El-Shaarawi, A.H., Esterby, S.R. & Dutka, B.J. (1981). Bacterial density in water determined by Poisson or negative binomial distributions, *Applied and Environmental Microbiology* **41**, 107–116.

Related Articles

Air Pollution Risk

Cumulative Risk Assessment for Environmental Hazards

Environmental Risk Assessment of Water Pollution

What are Hazardous Materials?

ABDEL H. EL-SHAARAWI

Wear

In material sciences, wear is defined as a dimensional loss of a solid resulting from its interaction with its environment. Wear processes include several different physical phenomena, e.g., erosion, corrosion, fatigue, adhesive wear, and abrasive wear. In risk and reliability engineering, the term *wear* (see **Imprecise Reliability; Reliability Data; Bayesian Statistics in Quantitative Risk Assessment**) is used in a broader sense and often refers to the deterioration of an item or equipment with usage and age, see, e.g. [1–3]. In this broader framework, wear represents a larger array of physical phenomena due to, e.g., mechanical and/or chemical actions such as fatigue, fatigue crack growth [4, 5], fracture, creep, erosion, corrosion, alteration of material properties with time, temperature, and light. Cutting tools [6], hydraulic and concrete structures [7, 8], brake linings, airplane engine compressor blades [9], corroding pipelines [10], and rotating equipments are examples of items that suffer increasing wear with usage and age.

For a given item, wear accumulates over time, contributes to the aging process and can be seen as the failure-generating mechanism: the original properties of the item change, which may lead to failure (when the wear level exceeds a given limit) if proper maintenance action is not carried out, to higher operating costs, and to a more intensive maintenance activity. To assess the reliability of a wearing system, to evaluate its life-cycle cost, or to optimize its design and maintenance, it is of primary interest to be able to predict its wear behavior, and we need models for this purpose. However, since the physical phenomena involved are very complex, it is almost impossible to derive a complete wear model on the basis of well-known physical laws, except for very local phenomena and it is thus necessary to use phenomenological models (see **Experience Feedback**). Since the wear model should also take into account the uncertainty, the variations around a mean behavior between items, and, if possible, the temporal variability of the deterioration, we have to use probabilistic phenomenological models [1, 11, 12]. Such models can then be used to estimate the time at which a given wear level is reached or to characterize the residual life of the wearing item. We can distinguish two approaches in probabilistic wear modeling: the lifetime modeling approach and

models based on time-dependent stochastic processes. The sections titled “Wear Modeling Using Lifetime Distributions” and “Wear Modeling by Stochastic Processes” propose synthetic, critical descriptions of these two wear-modeling options. The section titled “Application in Reliability and Maintenance” gives some examples of application of wear models in reliability and maintenance.

Wear Modeling Using Lifetime Distributions

In reliability theory [13, 14], we use a lifetime distribution (see **Risk Classification/Life; Mathematics of Risk and Reliability: A Select History**) to represent the random behavior of the time to failure T of an item. If $f(t)$ denotes the probability density function of the time to failure, and $F(t) = \mathbb{P}(T \leq t)$ the probability distribution function, the failure-rate function $\lambda(t)$ of the item is given by

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{\bar{F}(t)} \quad (1)$$

The survival (or reliability) function of the item is directly obtained from equation (1) by integration and exponentiation, and we have

$$\bar{F}(t) = \exp \left\{ - \int_0^t \lambda(s) \, ds \right\} \quad (2)$$

The quantity $\lambda(t) \, dt$ can be interpreted as the conditional probability that an item that has survived until time t will fail in the next time interval $[t, t + dt]$. Following this interpretation, it is natural that for wearing items, the failure rate is increasing. The classical “bathtub” curve (see **Accelerated Life Testing**) sketches the typical evolution of an item as a function of time that can be divided into three periods: the burn-in period, the useful life, and the wear-out period during which the failure rate increases under the effect of cumulative wear (see Figure 1). At this stage, assumptions on the wear behavior of the item are used to choose the general form of its failure-rate function. Lifetime distributions with an increasing failure rate (IFR) (see **Systems Reliability; Repair, Inspection, and Replacement Models**), known as *IFR lifetime distributions* [13, 14], are adapted to model the reliability behavior of wearing items. The Weibull

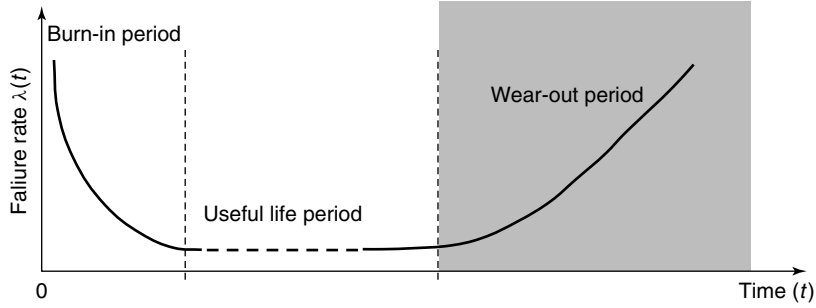


Figure 1 The bathtub curve

distribution (for $\alpha > 1$) and the gamma distribution (for $\alpha > 1$) both belong to the IFR distribution family.

Weibull Distribution

The Weibull distribution is one of the most widely used lifetime distributions because of its flexibility: an appropriate choice of parameters permits one to model different types of failure-rate behaviors. The failure-rate function given by

$$\lambda(t) = \alpha\beta^\alpha t^{\alpha-1} \quad (3)$$

is increasing for a shape parameter $\alpha > 1$; hence, the Weibull distribution with $\alpha > 1$ is often chosen to model the failure behavior of a wearing item.

Gamma Distribution

The probability density function for a gamma distribution is given by

$$f(t) = \frac{1}{\Gamma(\alpha)} \beta^\alpha t^{\alpha-1} e^{-\beta t} \quad (4)$$

It can be shown that the gamma distribution is IFR for a shape parameter $\alpha > 1$. The gamma distribution is also particularly adapted to the failure modeling of an item subject to wear because, for an integer value of α , it corresponds to an item whose wear level increases under the effect of a series of shocks occurring according to a homogeneous Poisson process with rate β , leading to a failure after shock α .

This modeling procedure has the main drawback of relying on a static binary “black box” approach.

The considered item is assumed to be either in the running state or in the failed state, regardless of its intermediate wear level and the dynamic failure-generating mechanism resulting from accumulated wear is not explicit in the model. A more dynamic and flexible approach to reliability modeling of wearing items is thus needed to take into account the temporal variability of deterioration and to elaborate wear-dependent reliability models: wear models based on stochastic processes offer such an alternative.

Wear Modeling by Stochastic Processes

For a number of applications, data are available from sensors or monitoring devices. This allows one to obtain measures describing the changes in the state of an item over time or with usage. From a modeling point of view, the engineering approach tries to describe by theory and testing, the physics of the dynamics to failure of the system. Such a methodology requires exact description of behavior and has already proved its worth in many cases, but makes it difficult to capture fluctuating operating environment. Two items of the same type with the same load can still show different deterioration rates. The engineering approach might work for tests in controlled laboratory environments. The results, *in situ*, will have a certain scatter and noise. At that instant, a stochastic approach that is presumed to describe the failure-generating mechanism under study might provide interesting results. Stochastic processes are particularly relevant to such a description, especially for items operating under a dynamic environment. They jointly incorporate the evolution of degradation with the temporal uncertainty and have proved to be better wear models than a random variable model

when the variability in degradation, thus in lifetime, is high [15].

Assuming the degradation of the system is measured on an increasing scale, let X_t denote the state of an item at time t . The internal wear of the item over the time interval $[s, t]$ is defined as an increment $X_t - X_s$ describing the degradation caused by continuous use over $[s, t]$. In most applications (e.g., erosion, corrosion), the wear of a nonmaintained system is a nondecreasing phenomenon. However, there are many situations in which the wear need not be increasing (e.g., because of possible healing of fatigue cracks). Various models that have been considered are based on the concept of accumulated damage, viewing the degradation and resulting failure of the item as the result of accumulation of wear over the lifetime. Shocks may occur to the system, and each shock may cause a random amount of wear, which accumulates additively. Most of the stochastic processes used for wear modeling proceed from the class of Lévy processes (see **Lévy Processes in Asset Pricing**), which contains tractable continuous-time processes with independent and stationary increments [16]. Using the class of Lévy processes retains the Markov property. Three Lévy processes can be considered as the main basic elements for stochastic wear modeling, each of them having explicit descriptions and possible extensions: the compound Poisson process (jump process [17, 18]), the gamma process (jump process [11, 19, 20]) and the Wiener process (with continuous sample paths [21–23]). A short description is given below.

Compound Poisson Process

The compound Poisson process (see **Ruin Probabilities: Computational Aspects; Volatility Smile**) can be used to model wear that accumulates in discrete amounts at isolated points in time, due to shocks. The deterioration $(X_t)_{t \geq 0}$ is described as a sum of wear increments W_i occurring at discrete times t_i :

$$X_t = \sum_{i=1}^{N_t} W_i \quad (5)$$

where $(N_t)_{t \geq 0}$ is a Poisson process and $(W_i)_{i=1,2,\dots}$ is a family of independent and identically distributed random variables, independent of the Poisson process. One of the first attempts to model wear as a stochastic process may have been with compound

Poisson Process [17]. Further studies can be found in [24] with extensions to continuous-wear processes.

Gamma Process

The gamma process (see **Large Insurance Losses Distributions; Repairable Systems Reliability**) is appropriate to model a nondecreasing deterioration. It is a stochastic continuous-time process with independent nonnegative increments having a gamma distribution [19]. The amount of wear $X_t - X_s = W_{t-s}$ over the time interval $[s, t]$ follows a gamma distribution $\text{Ga}(\alpha(t-s); \beta)$ with a shape parameter depending on $t-s$. The mean and variance are respectively

$$\mu_w = \frac{\alpha(t-s)}{\beta} \text{ and } \sigma_w^2 = \frac{\alpha(t-s)}{\beta^2} \quad (6)$$

As a limiting case of the compound Poisson process, the gamma process implies a wear increasing with frequent occurrences of very small increments, leading to practical uses in modeling erosion or corrosion, for example.

Wiener Process

The Wiener process (see **Lifetime Models and Risk Assessment**) (Brownian motion with drift) is a nonmonotonic process with continuous sample paths. The degradation at time t is described by a Wiener process with drift μ and variance σ if

$$X_t = \sigma B_t + \mu t \quad (7)$$

where B_t is the standard Brownian motion. A large drift and small variance parameters are often used to approximate monotonicity. One of the main advantages of the Wiener process for wear modeling is that it has been extensively studied in the mathematical literature. For example, the first passage time distribution is known to be the generalized inverse Gaussian distribution [25].

Covariates

Multiple sources of information can be used to characterize the wear of an item in time. Additional environmental and explanatory factors such as temperature, pressure, light, etc. may influence the degradation and lead to substantial heterogeneity

between the degradation paths. Describing failure-generating mechanisms by stochastic processes with covariates is particularly relevant to items operating under dynamic environment when the stresses are partly induced by environmental factors (or covariates) varying over time [2]. Wear stochastic processes with covariates can be used for internal stresses that change the rates and the modes by which the item degrades to failure. In these models, the degradation is described by the process $(X_{t,\theta})_{t \geq 0}$, where t is the time and θ is some possibly multidimensional random variable.

A lot of wear stochastic processes with covariates are extensions of the gamma process or of the Wiener process with linear drift. The consequences of such extensions are mainly an increasing number of parameters and possible related estimation troubles. In a gamma process, the covariates are mostly incorporated through the shape parameter α [11]. In [26], an accelerated life model is used, replacing αt by $\alpha(t \exp(\theta^T x))$, whereas in [20], the scale parameter is considered as a function of covariates $\beta(\theta)$ and can be replaced by $z\beta(\theta)$ where z is a random effect.

Multivariate stochastic processes like the Markov additive processes (MAP) [18] provide an extended framework for modeling degradation of items, given the stochastic behavior of covariates. Covariate processes that drive the wear are referred to in the engineering literature as “excitation processes”. MAP can be considered as extensions of compound Poisson processes [27].

Application in Reliability and Maintenance

Application in Reliability

One of the advantages of using a stochastic wear process is to develop a degradation-based reliability model instead of a failure-based reliability model [2, 28–30]. If one considers that the item fails as soon as the wear level X_t exceeds a given threshold L , the lifetime of the system is described by the random variable

$$\sigma_L(x) = \inf\{t : X_t > L\} \quad (8)$$

i.e., the first hitting time of the level L for the wear process X_t . The main difficulty with this approach is, then, to derive the lifetime distribution for $\sigma_L(x)$,

i.e., to establish an expression for the item reliability at time t :

$$\bar{F}^x(t) = \mathbb{P}(\sigma_L(x) > t | X_0 = x_0) \quad (9)$$

The evaluation of equation (9) requires calculation of the distribution of the first passage time of the process $(X_t)_{t \geq 0}$ above the level L , which is, in general, difficult for arbitrary stochastic processes. Closed form solutions or explicit results for the properties of the lifetime distribution in equation (9) can be obtained under simplifying assumptions [16, 19, 24, 31–33], but in the general case, for the estimation and characterization of the lifetime in equation (9) we have to resort to simulation-based techniques. As an example, considering the gamma process and given $X_0 = 0$, the usual distributions are given as follows:

- Density function of the wear level at time t :

$$f_w(u) = f_{\alpha t, \beta}(u) = \frac{\beta^{\alpha t} u^{\alpha t - 1}}{\Gamma(\alpha t)} \exp(-\beta u) \mathbb{I}_{u > 0} \quad (10)$$

- Cumulative lifetime distribution at time t :

$$F_\sigma(t) = \mathbb{P}(X_t \geq L) = 1 - F_{\alpha t, \beta}(L) \quad (11)$$

- Density function of the lifetime:

$$\begin{aligned} f_\sigma(t) &= - \frac{dF_T(t)}{dt} \\ &= \int_0^L \alpha f_{\alpha t, \beta}(s) \left(\frac{\Gamma'(\alpha t)}{\Gamma(\alpha t)} - \log(\beta s) \right) ds \end{aligned} \quad (12)$$

where $F_{\alpha t, \beta}$ is the distribution function of the gamma density probability function $f_{\alpha t, \beta}$ and $\Gamma(u)$ is the gamma function. The probability density function of the hitting time σ_L , given by equation (12), has no closed form expression and has to be computed numerically when required.

Traumatic events causing the failure of the item occurring, for example, as a Poisson process with a “killing” rate k depending on the item’s level of wear can be considered to make these models even more general [2]. Finally, this stochastic process-based approach, even if mathematically more demanding, leads to reliability models with a greater flexibility to fit failure data and to include environmental explicative variables as covariates. These models

thus offer the possibility to integrate the available information on dynamic environment and changing operating conditions of the considered item. They extend the range of possible applications of reliability models; for example, they can be used for a prognostic purpose, i.e., to predict the future wear behavior and to estimate the reliability and residual life of an item from wear measurements [28, 34]. Moreover, these wear-based reliability models include the classical failure-based lifetime models (such as Weibull or gamma distributions) as particular cases [2, 3]. To illustrate this last point, consider the “limit case” example given in [2] of an item subject to a wear, modeled by a deterministic diffusion process $X(t)$ satisfying the integral equation

$$X(t) = \int_0^t \mu(X(s)) ds + x_0 \quad (13)$$

The item reliability is then obtained as

$$\bar{F}^x(t) = \exp \left\{ - \int_0^t k(X(s)) ds \right\} \mathbb{I}_{\sigma_L(x) > t} \quad (14)$$

which shows that for a deterministic diffusion wear, the failure rate of the item at time t equals the wear-dependent killing rate at this time.

Application in Maintenance

On a wearing item, maintenance is necessary to control the wear level and to restore the item to a state in which it can perform correctly its required function. For classical time-based and age-based maintenance policies, the times between maintenance

actions (inspections, repairs, replacements, and so on) are determined using a prior knowledge given by the lifetime distribution of the maintained item. However, for a wearing item, when the wear level can be directly or indirectly monitored, it is preferable to choose a condition-based maintenance strategy and to base the maintenance decision on the observed wear level. The decision problem to solve is to determine sequentially the optimal inspection schedule and the optimal maintenance action to be carried out on the item at a given time. Here, a wear model is necessary to correctly interpret the raw wear measurements, to forecast the future behavior of the wearing system after a maintenance action, to assess the effects of a maintenance decision, and to optimize the condition-based maintenance policy parameters. Among the main advantages of a condition-based maintenance approach based on wear modeling, we mention the following:

- it has the capacity to adapt to the temporal variability of the wear evolution, e.g., an aperiodic inspection schedule (a more deteriorated system is inspected more frequently) or a proportioned wear level reducing maintenance operation;
- it has the ability to represent general maintenance actions (whose effect is between as-good-as-new replacements and as-bad-as-old minimal repairs) to control the wear of the item around a level where both preventive and corrective maintenance costs and operating costs reach an optimal balance.

As an example of such a condition-based maintenance policies, Figures 2 and 3 show the stochastic

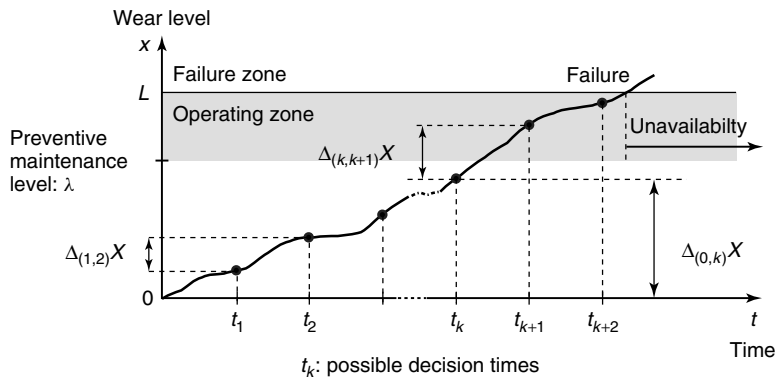


Figure 2 Wear model and condition-based maintenance policy

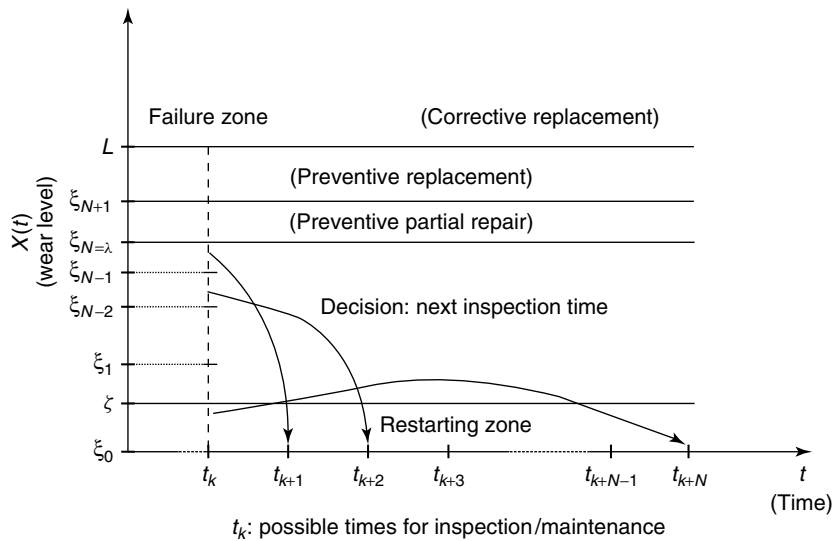


Figure 3 Structure of the condition-based maintenance policy

wear model and the maintenance policy structure used to schedule inspections, partial preventive repairs, and as-good-as-new preventive replacements in the condition-based maintenance model presented in [35]. The developed maintenance cost model allows the optimal tuning of the wear thresholds of the policy structure (multilevel control limit rule). See [36–42] for other examples and [11] for a complete review of the use of gamma wear processes in condition-based maintenance models.

References

- [1] Bogdanoff, J.L. & Kozin, F. (1985). *Probabilistic Models of Cumulative Damage*, John Wiley & Sons.
- [2] Singpurwalla, N.D. (1995). Survival in dynamic environments, *Statistical Science* **10**(1), 86–103.
- [3] Wenocur, M.L. (1989). A reliability model based on the gamma process and its analytic theory, *Advances in Applied Probability* **21**, 899–918.
- [4] Ray, A. & Tangirala, S. (1997). A nonlinear stochastic model of fatigue crack propagation, *Probabilistic Engineering Mechanics* **12**, 33–40.
- [5] Gillen, K.T. & Celina, M. (2001). The wear-out approach for predicting the remaining lifetime of materials, *Polymer Degradation and Stability* **71**, 15–30.
- [6] Jeang, A. (1999). Tool replacement policy for probabilistic tool life and random wear process, *Quality and Reliability Engineering International* **15**, 205–212.
- [7] van Noortwijk, J.M. & Klatter, H.E. (1999). Optimal inspection decisions for the block mats of the Eastern-Scheldt barrier, *Reliability Engineering and System Safety* **65**, 203–211.
- [8] Redmond, D.F., Christer, A.H., Rigden, S.R., Burley, E., Tajelli, A. & Abu-Tair, A. (1997). O.R. modelling of the deterioration and maintenance of concrete structures, *European Journal of Operational Research* **99**, 619–631.
- [9] Hopp, W.J. & Kuo, Y.-L. (1998). An optimal structured policy for maintenance of partially observable aircraft engine components, *Naval Research Logistics* **45**, 335–352.
- [10] Hong, H.P. (1999). Inspection and maintenance planning of pipeline under external corrosion considering generation of new defects, *Structural Safety* **21**, 203–222.
- [11] van Noortwijk, J.M. (2007). A survey of the application of gamma processes in maintenance, *Reliability Engineering and System Safety*. DOI:10.1016/j.ress.2007.03.019 (In Press).
- [12] Virkler, D.A., Hillberry, B.M. & Goel, P.K. (1979). The statistical nature of fatigue crack propagation, *ASME Journal of Engineering Materials and Technology* **101**(2), 148–153.
- [13] Rausand, M. & Høyland, A. (2004). *System Reliability Theory – Models, Statistical Methods and Applications*, Wiley Series in Probability and Statistics, Wiley-Interscience.
- [14] Barlow, R.E. & Proschan, F. (1996). *Mathematical Theory of Reliability, Classics in Applied Mathematics*, SIAM, Vol. 17. Previously published by John Wiley & Sons, New York (1965).
- [15] Pandey, M.D., Yuan, X.X. & van Noortwijk, J.M. (2007). The influence of temporal uncertainty of

- deterioration in life-cycle management of structures, *Structure and Infrastructure Engineering* DOI: 10.1080/15732470601012154.
- [16] Abdel-Hameed, M. (1984). Life distribution properties of devices subject to Lévy wear process, *Mathematics of Operations Research* **9**, 606–614.
- [17] Mercer, A. & Smith, C.S. (1959). A random walk in which the steps occur randomly in time, *Biometrika* **46**(1–2), 30–35.
- [18] Çinlar, E. (1977). Shock and wear models and Markov additive processes, in *The Theory and Applications of Reliability*, I.N. Shimi & C.P. Tsokos, eds, Academic Press, pp. 97–115, Vol. 1.
- [19] Abdel-Hameed, M. (1975). A gamma wear process, *IEEE Transactions on Reliability* **24**, 152–153.
- [20] Lawless, J. & Crowder, M. (2004). Covariates and random effects in a gamma process model with application to degradation and failure, *Lifetime Data Analysis* **10**, 213–227.
- [21] Lawless, J., Hu, J. & Cao, J. (1995). Methods for the estimation of failure distribution and rates from automobile warranty data, *Lifetime Data Analysis* **1**, 227–240.
- [22] Whitmore, G.A. (1995). Estimating degradation by a Wiener diffusion process subject to measurement error, *Lifetime Data Analysis* **1**, 307–319.
- [23] Whitmore, G.A., Crowder, M.J. & Lawless, J.F. (1998). Failure inference from a marker process based on a bivariate Wiener process, *Lifetime Data Analysis* **4**, 229–251.
- [24] Esary, J.D., Marshall, A.W. & Proschan, F. (1973). Shock models and wear processes, *The Annals of Probability* **1**(4), 627–649.
- [25] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, Wiley Series in Probability and Mathematical Statistics, 2nd Edition, John Wiley & Sons.
- [26] Bagdonavicius, V. & Nikulin, M. (2001). Estimation in degradation models with explanatory variables, *Lifetime Data Analysis* **7**(1), 85–103.
- [27] Çinlar, E. (1972). Markov additive processes II, *Probability Theory and Related Fields* **24**(2), 95–121.
- [28] Kharoufeh, J.P. & Cox, S.M. (2005). Stochastic models for degradation-based reliability, *IIE Transactions* **37**, 533–542.
- [29] Mercer, A. (1961). Some simple wear-dependent renewal processes, *Journal of the Royal Statistical Society B* **23**(2), 368–376.
- [30] Reynolds, D.S. & Savage, I.R. (1971). Random wear models in reliability theory, *Advances in Applied Probability* **3**, 229–248.
- [31] Kharoufeh, J.P. (2003). Explicit results for wear processes in a Markovian environment, *Operations Research Letters* **31**(3), 237–244.
- [32] Lu, C.J. & Meeker, W.Q. (1993). Using degradation measures to estimate a time-to-failure distribution, *Technometrics* **35**(2), 161–174.
- [33] Lu, C.J., Park, J. & Yang, Q. (1997). Statistical inference of a time-to-failure distribution derived from linear degradation data, *Technometrics* **39**, 391–400.
- [34] Chinnam, R.B. (1999). On-line reliability estimation of individual components using degradation signals, *IEEE Transactions on Reliability* **48**(4), 403–412.
- [35] Castanier, B., Bérenguer, C. & Grall, A. (2003). A sequential condition-based repair/replacement policy with non periodic inspections for a system subject to continuous wear, *Applied Stochastic Models in Business and Industry* **19**, 327–347.
- [36] Grall, A., Dieulle, L., Bérenguer, C. & Roussignol, M. (2002). Continuous-time predictive-maintenance scheduling for a deteriorating system, *IEEE Transactions on Reliability* **51**(2), 141–150.
- [37] Hontelez, J.A.M., Burger, H.H. & Wijnmalen, J.D. (1996). Optimum condition-based maintenance policies for deteriorating systems with partial information, *Reliability Engineering and System Safety* **51**, 267–274.
- [38] Grall, A., Bérenguer, C. & Dieulle, L. (2002). A condition-based maintenance policy for stochastically deteriorating systems, *Reliability Engineering and System Safety* **76**, 167–180.
- [39] Christer, A.H. & Wang, W. (1995). A simple condition monitoring model for a direct monitoring process, *European Journal of Operational Research* **82**, 258–269.
- [40] Barker, C.T. & Newby, M.J. (2007). Optimal non-periodic inspection for a multivariate degradation model, *Reliability Engineering and System Safety*. DOI:10.1016/j.ress.2007.03.015 (In Press).
- [41] Park, K.S. (1988). Optimal continuous-wear limit replacement under periodic inspections, *IEEE Transactions on Reliability* **37**(1), 97–102.
- [42] Park, K.S. (1988). Optimal wear-limit replacement with wear-dependent failures, *IEEE Transactions on Reliability* **37**(1), 293–294.

CHRISTOPHE BÉRENGUER AND ANTOINE
GRALL

Weather Derivatives

Historical Perspective

Weather risks are unpredictable fluctuations in cash flows caused by noncatastrophic weather events such as temperature, rainfall, snowfall, and humidity. Weather risks are in general low loss but high occurrence events and are covered by weather derivatives. In contrast, catastrophic (CAT) risks such as tornadoes, wind storms, and tsunamis are usually covered by weather insurance (*see Nonlife Insurance Markets; Large Insurance Losses Distributions; Risk Classification/Life; Ruin Probabilities: Computational Aspects*).

Weather risks can have large impacts on business activities. It is estimated that about 30% of the US economy faces the challenge of weather risks. For example, agricultural outputs and earnings of energy industries are affected by weather conditions. If a summer month is unusually cool, then profits of energy companies are adversely affected owing to the decreased demand of air conditioning. It is therefore necessary to hedge weather risks due to unexpected weather conditions (*see Statistical Arbitrage*). In July 1996, the first weather derivative product was introduced when Aquila Energy structured a dual-commodity hedge for Consolidated Edison (ConEd) Company. The transaction involved ConEd's purchase of electric power from Aquila for the month of August in 1996. A weather clause was embedded into the contract. Aquila would offer a discount to ConEd if August turned out to be cooler than expected. Subsequently, weather derivatives slowly began trading over the counter in 1997. The Weather Risk Management Association (WRMA) was found in 1999. It plays an important role in developing the weather market. WRMA coordinates with the National Oceanic and Atmospheric Administration (NOAA) and the National Weather Service (NWS) to improve weather reporting and data collection (*see Natural Resource Management*). It extends the market to different commercial areas, such as banking, agriculture, energy and trading, and organizes regular meetings and conferences in Europe, Asia, and North America (*see WRMA homepage: <http://www.wrma.org>*).

As the market for weather-related products expanded, Chicago Mercantile Exchange (CME) introduced the first exchange-traded weather futures contracts (and corresponding options) in 1999 and launched weather contracts in October 2003. CME now launches weather contracts in more than a dozen United States cities and in several European and Japanese cities.

The weather market in 1998 was estimated at US \$500 million and it has grown to about US \$5 billion today. The weather derivatives market has become the fastest growing derivatives market (*see CME homepage: <http://www.cme.com/>*).

Weather Derivatives

Definition of Weather Derivatives

Weather derivatives are contingent securities that offer prespecified payments based on how an underlying weather index differs from a prespecified strike value. A typical weather derivative consists of several components as follows:

- *Weather station*
All weather contracts are based on the observations of one or more specific weather stations.
- *Period*
It includes specific start and end date of contract. The length is usually either a month or a season.
- *Underlying index*
It is used to define the payoff of a weather contract. Popular indexes include heating degree days (HDD), cooling degree days (CDD), average temperature, rainfall, snowfall, and humidity. Among them, most trading index is temperature-related derivatives and most contracts are written on either HDD or CDD. Specifically, the HDD index is defined as the sum of the extent to which daily degree average temperatures below the 65°F benchmark while CDD index is the sum of the extent to which degree daily average temperatures above the 65°F benchmark. They are defined as follows:

$$\text{HDD} = \sum \max\{65 - T_i, 0\}, \quad i \in \text{prespecified period}$$

$$\text{CDD} = \sum \max\{T_i - 65, 0\}, \quad i \in \text{prespecified period}$$

where T_i is the average temperature of the i th day.

Table 1 Atlanta February HDD swap

Reference weather station	Atlanta Peachtree Airport (WBAN: 53863)
Period	February 1–28, 2006
Tick size	\$500 per HDD
Fixed rate	1000
Actual accumulated HDD	1300
Payoff at maturity for buyer	$(1300 - 1000) \times 500 = \$150\,000$

- *Tick size*

Prespecified amount attached to each unit of index value.

As an example, consider the contract listed in Table 1.

Weather Derivatives versus Traditional Derivatives

Weather derivatives differ from traditional derivatives for several reasons. While traditional derivatives can be priced using Black–Scholes model, pricing of weather derivatives is more trickier.

First of all, underlying assets such as stocks or currencies are easily available in the spot market for traditional derivatives. In contrast, temperature, rainfall, and snowfall are all nontradable assets. Second, although the weather derivative market has been growing gradually, it is unlikely that it will be as liquid as traditional markets due to its location-specific nature. According to US-based Futures Industry Association (FIA), over 90% of global derivatives trading is in financial derivatives while less than 0.02% is in weather derivatives.

Valuation of Weather Derivatives

People are interested in valuation of weather derivative contracts for three main reasons. The first one is to determine an appropriate strike or an appropriate option premium. The second one is that once a contract has been traded, it is important to know the current value of holdings based on the latest weather conditions. The third one is that external regulators often need to monitor the risk a weather trading organization faced due to the traded contracts. To value weather contracts, there are several possibilities.

Actuarial Pricing

Weather contracts can be valued on the basis of the probabilities of all possible outcomes. This is usually the starting point of all weather option valuation and is commonly known as *actuarial pricing* (see **Insurance Pricing/Nonlife**). Actuarial pricing is performed by using historical meteorological data and meteorological forecasts to predict the distribution of the possible outcomes of the index.

Market-Based Pricing

As in the equity market, market-based pricing consists of using the market price of the equity along with an appropriate model to derive the no-arbitrage price (see **Risk-Neutral Pricing: Importance and Relevance**). In the case of weather contracts, we could consider using the observable market price for the weather options as the value. But this will only be possible if there is an observable market.

At first sight, the payoff of a weather option depends on an index derived from meteorological variables, which is not a traded quantity, and as a result, dynamic hedging may not be possible. But if there is a swap contract defined on the index, then under certain assumptions, dynamic hedging of the risk in a weather option is still possible using swaps in the same way as equities is used to hedge equity options. See [1] for a no-arbitrage pricing method using weather swaps/forwards to hedge weather options. For such dynamic hedging to be possible, the underlying swap market has to be fairly liquid, otherwise the costs involved are huge. Another method along the same line was proposed in [2]. By assuming that the temperature follows a fractional Ornstein–Uhlenbeck process (see **Lévy Processes in Asset Pricing**), Benth [2] derived an arbitrage-free pricing formula for weather derivatives using the idea of quasi-conditional expectation.

Marginal Value Approach

Davis [3] valued weather derivative using the “shadow price” or the “marginal substitution value” approach from mathematical economics (see [4]). He modeled HDD and commodity prices by a geometric Brownian motion (see **Stochastic Control for**

Insurance Companies) and deduced expressions for option values and swap rates.

Equilibrium Valuation Approach

Cao and Wei [5] proposed to value weather derivative by means of the Lucas model. Lucas model is a consumption-based asset pricing model, where an economy consists of a large number of identical individual consumers. The Lucas model then relates the aggregate output of an economy to exogenous output variables (see [6]). Cao and Wei treated temperature as a fundamental uncertainty in the economy. After modeling the daily temperature by an autoregressive time series, they related it with aggregate dividends through correlations. They used this model to test the significance of market price of risk of temperature. Surprisingly, their empirical results showed that the market price of risk associated with temperature is insignificant. Their results also suggested that historical simulations will generally lead to overpricing of weather options.

Richards *et al.* [7] developed a general valuation method of weather derivatives based on the approach of Cao and Wei. They used a mean-reverting Brownian motion model with lognormal jumps and time-varying volatility to model the daily temperature. They concluded that the assumption of zero market price of risk is problematic when risk-neutral valuation principle cannot be applied. Further details about pricing of weather derivatives can be found in [8, 9].

In summary, weather option pricing is a mixture of actuarial and market-based techniques, with more emphasis on the actuarial side in most cases. For locations where the swap market is ill-liquid, actuarial valuation may be the only means. For locations where the swap market is sufficiently liquid, no-arbitrage pricing may have some relevance because dynamic

hedging using swaps may be possible. Finally, for locations where options are actively traded, one may be able to take the option valuation directly from the market.

Use of Weather Modeling in Pricing Weather Derivatives

Weather modeling cannot be separated from weather derivatives pricing. Weather forecast plays an indispensable role in weather derivatives market. Buyers need to evaluate as to how much weather risk is needed to be hedged in order to determine the correct number of weather contracts to enter. Sellers need to decide an appropriate price for weather derivatives. As price movements of traded contracts are usually driven by weather forecasts, a good weather forecasting model is instrumental in accurate pricing of weather contracts. Among different types of weather forecasting models, temperature forecasting is the most commonly studied one.

Modeling Temperature

Most weather forecasts are produced by dynamical models of the atmosphere known as *atmospheric general circulation models (ACGMs)*; see Section 9.1.3 of [8] and the references therein for further details about ACGMs. A lucid discussion on meteorology is given in [10]. Time series is another common approach used to model and forecast daily average temperatures. Figure 1 shows the daily average temperature time series from 1/1/2000 to 31/12/2004 for four US cities. All temperature data are obtained from the global summary of the day (GSOD) database archived by the National Climatic Data Center (NCDC).

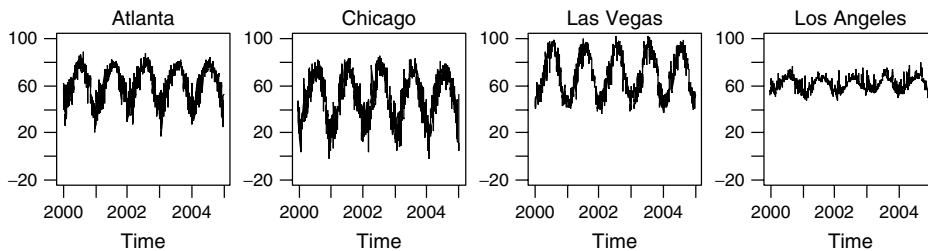


Figure 1 Time series plots of daily temperature for four US cities

Global warming and solar activities may produce a deterministic trend for the temperature, though it is minor for this sampling period. The daily average temperature exhibits strong and stable seasonal trends in both mean and variance and oscillates regularly between high temperature (summer) and low temperature (winter) in a period of 1 year. Standard deviation also displays higher values in winter months but lower values in summer months. Each series has similar patterns although they are different in peak values across cities.

After removing the deterministic trends using standard methods as given in chapter one of [11], the remaining data is the anomaly. Unlike the cyclic trend observed in Figure 1, anomaly is the unpredictable component in temperature. In other words, it represents the temperature risk. Cabellero *et al.* [12] studied the behavior of anomaly and detected long-range dependence in Central England, Chicago, and Los Angeles. Moreno [13] introduced an ARFIMA-FIGARCH process to capture the long-memory and heteroskedasticity properties.

Another approach is to model the underlying index directly. But this requires an additional normality assumption for the distribution of the temperature index. Validity of the normality assumption was challenged in [14].

Weather Forecasting

As weather contracts usually last for months or seasons, weather models for pricing weather derivatives should provide good prediction power over an extended period of time. However, most of the meteorological weather forecasting literature only focuses at short-horizon forecasts using physical models. Meteorological models may not be the best means for weather contracts. Having a reasonable long-horizon weather forecast (or even a complete density forecast) is thus crucial for pricing weather derivatives.

To this end, Campbell and Diebold [15] estimated the unconditional distributions of daily temperature by using a nonparametric kernel density estimate. They accessed the performance of long-horizon density forecasts and obtained encouraging results. Another method of long-horizon weather forecasting is ensemble prediction. Ensemble prediction derives a more sophisticated density estimate for weather variables. The prediction systems

generate many realizations of weather variables using different initial conditions of numerical models. These are then used to construct an estimate of the density function (*see Copulas and Other Measures of Dependency*). Taylor and Buizza [16] compared the ensemble-based forecast with the univariate time series approach and concluded that the ensemble-based forecast outperformed the time series approach.

Conclusion

In this article, we present the history of the weather market, and discuss weather risks, weather derivatives, and their valuations. Since the weather derivative market is an incomplete market, traditional complete market pricing method may not be directly applicable. Further research is needed to develop an effective pricing method for incomplete markets of weather derivatives.

References

- [1] Jewson, S. & Zerovs, M. (2003). *The Black-Scholes Equation for Weather Derivatives*, <http://ssrn.com/abstract=436282>.
- [2] Benth, F.E. (2003). On arbitrage-free pricing of weather derivatives based on fractional Brownian motion, *Applied Mathematical Finance* **10**, 303–324.
- [3] Davis, M. (2001). Pricing weather derivatives by marginal value, *Quantitative Finance* **1**, 1–4.
- [4] Foldes, L. (2000). Valuation and martingale properties of shadow prices: an exposition, *Journal of Economic Dynamics and Control* **24**, 1641–1701.
- [5] Cao, M. & Wei, J. (2004). Equilibrium valuation of weather derivatives, *Journal of Futures Markets* **24**, 1065–1089.
- [6] Lucas, R.E. (1978). Asset prices in an exchange economy, *Econometrica* **46**, 1429–1445.
- [7] Richards, T.J., Manfredo, M.R. & Sanders, D.R. (2004). Pricing weather derivatives, *American Journal of Agricultural Economics* **86**, 1005–1017.
- [8] Jewson, S. & Brix, A. (2005). *Weather Derivative Valuation*, Cambridge University Press, Cambridge.
- [9] Banks, E. (2002). *Weather Risk Management: Markets, Products and Applications*, Macmillan, New York.
- [10] Thompson, R. (1998). *Atmospheric Processes and Systems*, Routledge.
- [11] Chan, N.H. (2002). *Time Series: Applications to Finance*, John Wiley & Sons, New York.
- [12] Caballero, R., Jewson, S. & Brix, A. (2002). Long memory in surface air temperature detection, modeling

and application to weather derivative valuation, *Climate Research* **21**, 127–140.

- [13] Moreno, M. (2003). Weather derivatives hedging and swap illiquidity, *WRMA*, Coral Gables, Florida.
- [14] Jewson, S. (2004). *Weather Derivative Pricing and the Distributions of Standard Weather Indices on Us Temperatures*, <http://ssrn.com/abstract=535982>.
- [15] Campbell, S. & Diebold, F.X. (2005). Weather forecasting for weather derivatives, *Journal of the American Statistical Association* **100**, 6–16.
- [16] Taylor, J.W. & Buizza, R. (2006). Density forecasting for weather derivative pricing, *International Journal of Forecasting* **22**, 29–42.

Related Articles

Environmental Security

Global Warming

Hazards Insurance: A Brief History

Insurability Conditions

Multiattribute Modeling

NGAI H. CHAN AND CHUN Y. WONG

What are Hazardous Materials?

A hazardous substance is a material with a substantial potential to pose a danger to living organisms, materials, structures, or the environment. In a risk assessment context, this is called the *hazard identification phase*. Throughout history, humans have suffered from excessive exposure to hazardous substances. Exposure incidents range from inhalation of the noxious volcanic gases, death due to exposure to carbon monoxide from fires in caves, to lung diseases caused by using mineral asbestos fibers when weaving cloths in ancient Greece. Adverse health effects were observed in Europe during the 1700–1800s as a result of human exposure to hazardous material and toxicants in runoff from mine spoils piles and from poisoning due to exposure to coal tar by-product [1]. In the late 1900s, it became evident that environmental pollution and exposure to hazardous materials, particularly in industrialized countries, were becoming major problems (*see Environmental Hazard*).

According to Paracelsus, all chemicals and metals from synthetic to naturally occurring ones have the potential to be poisons, and only the dose differentiates a poison from a remedy. Depending on the dose, it is possible for any material to cause harm, no matter how benign or essential it is to life [2]. Recently, during fraternity hazing events, several college students drank excessive quantities of water and died owing to the resulting electrolyte imbalance.

Hazard is also described as an intrinsic capability of material to cause harm. The hazard posed by material can be a function of material toxicity, mobility, environmental persistence, bioaccumulation rate, and/or methods used to contain the material [3]. Hazards can be either natural (geophysical and biological) or technological (industrial emissions, accidents, and chemical spills). The technological hazards are those causing human injury, illness, or death, and ecological, environmental, and property damage (*see Environmental Hazard*). Hazard may also be associated with the toxicological, nuclear, chemical, or biological properties of a substance [4].

Legally, when a chemical is present in a consumer product or other media such as waste streams, the term *hazardous material* can have a specific meaning

and certain expectations may be imposed. Under most laws or regulations, a hazardous material is defined as an item or agent (biological, chemical, or physical) that has the potential to cause harm either by itself or through interaction with other factors [2]. Accordingly, any material, element, compound, or combination of materials that poses a threat to human health and the environment owing to its toxic, corrosive, ignitable, explosive, or chemically reactive nature is labeled a hazardous material [5–7].

Most regulatory programs have identified plans and procedures to evaluate quantitative estimate of risks associated with exposure to the hazardous material. The risk assessment program defined by the US National Academy of Sciences and the United States Environmental Protection Agency (USEPA) involves processes such as hazard identification, exposure assessment, toxicity assessment, and risk characterization [3, 8].

Regulatory Framework

Hazardous Material Characterization

Hazardous material is often measured and expressed in terms of its chemical properties. Accordingly, a fundamental knowledge of some basic concepts and principles of chemical basics are often needed in order to classify chemicals. Matters are generally divided into three physical states: solids, liquids, and gases. These terms, however, can take on a special significance when they are found in regulatory language [2].

Hazardous substances, a subset of hazardous materials, are identified and listed by regulatory agencies such as the USEPA and the Occupational Safety and Health Administration (OSHA) under various regulations. The main difference between the USEPA and OSHA is that OSHA is an enforcement organization and has major responsibility over safety and health issues in the workplace [7].

United States Environmental Protection Agency (USEPA)

Protection of the human health and the environment often requires development of regulatory frameworks and guidelines for proper characterization, prevention, handling, and management of hazardous material (*see Environmental Health Risk*). According

to the USEPA, hazardous materials encompass any material, including substances and wastes that may pose an unreasonable risk to health, safety, property, or the environment. In specific quantities and forms, these materials can often pose a danger of fire, become explosive, flammable, or undergo a reaction, which releases excessive heat. Regulatory citations are provided only for the statutes dealing most directly with pesticides and industrial chemicals [9, 10].

The USEPA further classifies hazardous materials as carcinogenic or toxic depending on the concentration, duration, and type of exposure, whether the material can cause sickness or death, and if exposure to those induces reactions in tissues causing cellular damage or death [3, 9]. For the purpose of quantifying human health risk, chemicals are characterized as carcinogens, which are those chemicals with demonstrated propensity for cancer induction, and noncarcinogens. If chemicals behave as both carcinogens and noncarcinogens, they appear in both types of calculations of potential human health risk assessments [3]. On the basis of tests conducted by the National Toxicology Program (NTP) of the National Institute of Environmental Health Sciences (NIEHS) and according to the weight of evidence from animal experiments, clinical experience, and epidemiologic studies, the USEPA also recommend the following classifications for chemicals: class A (human carcinogens), class B (probable human carcinogen), class C (possible human carcinogen), class D (not classified as human carcinogen), and class E (evidence of non-carcinogenicity for humans) [11].

Under some regulations, the USEPA also uses four criteria to list discarded material (waste) as hazardous: (a) the waste typically contains certain chemicals, and certain factors are considered, which decide whether it could pose a threat to human health and the environment in the absence of special regulation (toxic listed wastes); (b) the waste contains such dangerous chemicals that it could pose a threat to human health and the environment even when properly managed (acutely hazardous wastes); (c) the waste typically exhibits one of the four characteristics of hazardous waste described in the hazardous waste identification regulations (ignitability, corrosivity, reactivity, or toxicity); and (d) when EPA has a cause to believe, for some other reason, the waste typically fits within the statutory definition of hazardous waste developed by the US Congress [7, 12, 13].

The following sections present an overview of the four main criteria used in characterization of hazardous material by the USEPA regulations [13]. A comprehensive review of the regulations and the guidelines are found in the literature [9, 11–13].

Ignitability. Ignitable materials are wastes that can readily catch fire and sustain combustion. Most ignitable wastes are liquid in physical form. The USEPA selected a flash point test as the method for determining whether a liquid waste is ignitable and, therefore, can be classified as hazardous. The flash point test determines the lowest temperature at which a chemical ignites when exposed to flame ($<140^{\circ}\text{F}$). A nonliquid waste is ignitable and hazardous if it can spontaneously catch fire under normal handling conditions and can burn so vigorously that it creates a hazard. Certain compressed gases and chemicals called *oxidizers* can also be ignitable. The regulations describing the characteristics of ignitability are codified with the Code of Federal Register (CFR) (40 CFR part 261.21) [13].

Corrosivity. Materials with acidic or alkaline (basic) characteristics and capability of corroding steel at a rate greater than 6.35 mm yr^{-1} at a temperature of 55°C , or dissolving flesh, metal, or other materials (chemicals with pH between 2 and 12.5) are characterized as corrosive. The regulations describing the characteristic of corrosivity can be found in section 40 CFR part 261.22 [13].

Reactivity. A reactive material can readily explode or undergo violent reactions. Common examples are discarded munitions or explosives. In many cases, it is difficult to do tests for the material reactivity under normal conditions. Reactive materials undergo violent change without detonating, react violently with water, form explosive mixtures with water, and generate toxic gases, vapors, or fumes. Cyanide or sulfide-bearing materials (wastes) are considered reactive, since when present in solutions having a pH between 2 and 12.5 they generate gases, vapors, or fumes, and are capable of detonation or explosive reactions. The characteristic of reactivity is described in the regulations at 40 CFR part 261.23 [13].

Toxicity. Toxicity is another term commonly used in discussions of hazardous material. It refers to the

capacity to cause toxic effects in the living organisms, including human and ecological receptors [9]. The degree of toxicity varies with the particular chemical. Upon exposure, some chemicals can produce asphyxiation, causing oxygen concentrations in the blood to go below the level that is necessary to support life. Others, following ingestion, can damage vital organs such as liver or heart, and/or cause adverse effects on the nerves and ability of the nerves to transfer impulses [6, 14].

Toxicity is often characterized as chronic or acute. Chronic toxicity is the occurrence of symptoms, diseases, or other adverse health effects that develop or persist over time following exposure to a single dose or multiple doses of a chemical over a long period of time. Acute toxicity is the development of symptoms of poisoning or the occurrence of adverse health effects after exposure to a single dose or multiple doses of a chemical within a short period of time. The risk is the statistical probability that a toxic response will be observed [2].

USEPA suggested a toxicity characteristic leachate procedure (TCLP) for determination of the toxicity level of the material containing hazardous chemicals. A substance is determined to be toxic if the results of TCLP tests indicate that it contains chemicals with concentrations exceeding published levels of contaminants listed in 40 CFR 261.24. Examples of toxic material listed by the USEPA include arsenic, barium, cadmium, chromium, lead, mercury, selenium, silver, endrin, lindane, methoxychlor, and toxaphene [12, 13].

Occupational Safety and Health Administration (OSHA)

OSHA defines hazardous materials and chemicals in terms of the “health hazards” and “physical hazards”.

Health hazards include chemicals that are carcinogens, toxic agents, irritants, corrosives, and sensitizers; agents that act on the hematopoietic system; agents that damage the lungs, skin, eyes, or mucous membranes; chemicals that are combustible, explosive, flammable, oxidizers, pyrophorics, unstable-reactive, or water-reactive; chemicals that, in the course of normal handling, use, or storage may produce or release dusts, gases, fumes, vapors, mists, or smoke, which may have any of the previously mentioned characteristics (29 CFR 1910.1200) [5, 12]. Physical hazards are an entirely separate

category under OSHA. These include heat stress, exposure to nonionizing and ionizing radiation, vibration, excessive or inadequate lighting, ergonomic stresses, and others.

Hazard Impact Analysis and Toxicity

Quantitative measures of hazard consequences of hazardous material can be determined in terms of risk, which are conditional probabilities of experiencing harm, its likelihoods, and severity of the harm. Hazard impact analysis is a science-based estimate of the human health risk due to exposure to hazardous material. Risk assessment can estimate the probability that a particular substance will pose a certain adverse effect to humans and the environment. Risk is estimated as a range of probabilities and characterized by both uncertainty within the data and assumptions made during the assessment processes. Risk starts with the exposure, which is defined as the contact of an organism’s exterior with a chemical.

Toxicologically, exposure is the means by which the living organism acquires a dose, which is defined as the total amount of chemicals received through contact with various substances (i.e. toxic material). Exposure can be categorized as external and/or internal. External exposure is linked closely to the chemical’s environmental dissipation, contact rate, the quantity of the chemical available, and the time. The internal exposure, however, is the amount of the chemical absorbed into the systemic circulation [14].

The USEPA recommends a four-step procedure for assessing human health risk. The procedure consists of (a) hazard identification, (b) dose–response assessment, (c) exposure assessment, and (d) risk characterization [8]. Hazard identification demonstrates whether exposure to an agent can cause or increase a particular adverse health effect [15]. The process involves gathering and evaluating data on the types and extent of the consequences produced by hazardous situations or substances. Toxicity assessment emphasizes establishment of dose–response characteristics, a graphical representation of the relationship between the degree of exposure to a chemical substance, and the hazard predicted by the biological effects or responses [15].

When the extent of any particular response is plotted against dose (exposure level), a characteristic

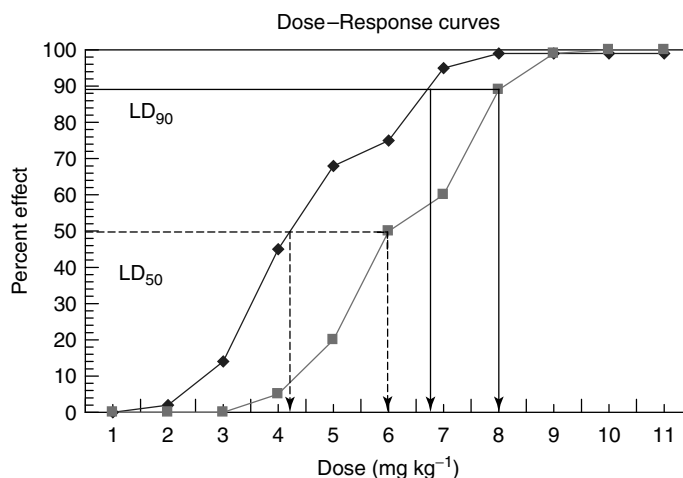


Figure 1 Illustration of dose–response curves, LD₅₀, and LD₉₀, identified for two different chemicals used in a toxicological study

S-shaped curve is produced. As shown in Figure 1, the response of the test organism (identified as percent effect) varies according to the magnitude of the administered dose and the potency of the material used in the study. The LD₅₀ and LD₉₀ represent lethal dose affecting 50 and 90% of the test population, respectively. In cases where the observed response is death, the dose–response curve (*see Dose–Response Analysis*) can be used to estimate the LD₅₀ (the dose that will kill about 50% of the animals). Similarly, depending on the classification scheme that is used, a nonlethal response may be quantified as an effective dose (ED) or a toxic dose (TD) [14].

The highest dose for which no response is observed is identified as the no observed effect level (NOEL), while the lowest dose to give the response may be termed the *toxic threshold*, again, depending on the regulatory or classification system used. Some chemicals, like the genotoxic carcinogens, can be assumed to have no threshold dose for regulatory purposes [15].

Dose–response curves can also be used to identify toxicological criteria such as lowest observed adverse effect level (LOAEL), lowest observed effect level (LOEL), and no observed adverse effect level (NOAEL) in units of milligram per kilogram of body weight per day. The slope factor (SF) can also be determined from the dose–response curve. On the basis of regulatory convention, the SF is the plausible upper-bound estimate (estimate not likely to be

lower than the true value) of the probability of an individual developing cancer as a result of a lifetime of exposure to a particular level of a potential carcinogen [2, 15].

The dose–response estimation process takes into account the uncertainty within the process, defined as the lack of confidence in the estimate of a variable’s magnitude or probability of its occurrence. Factors such as genetic fitness, metabolic differences, body fat, the level and extent of antagonism, bioaccumulation, and bioconcentration of chemicals in body and cell tissues control experimental outcome for the administered doses [11]. Hypersensitive organisms respond at a low dose end and resistant organisms at the other. If the test population were homogeneous, the dose–response line would be essentially vertical. The steepness of the slope helps to determine the degree of homogeneity of the population tested [14].

Hazardous Material Management

Mitigating the risk associated with exposure to hazardous materials requires the application of safety precautions during their production, handling, transportation, storage, use, and disposal. Most countries have recognized the need for regulating hazardous material. In the United States, proper management of hazardous materials has been defined through numerous administrative requirements. These have

been identified through regulations promulgated by various agencies such as USEPA, Nuclear Regulatory Commissions (NRC), National Institute for Occupational Safety and Health (NIOSH), Department of Transportation (DOT), and several acts such as Toxic Substances and Control Act (TSCA), Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), Federal Food, Drug, and Cosmetic Act (FFDCA), Emergency Planning, and Community Right-to-Know Act (EPCRA) (mandates a national inventory of toxic chemical releases (Toxics Release Inventory (TRI)), Clean Air Act (CAA), Clean Water Act (CWA), Safe Drinking Water Act (SDWA), Resource Conservation and Recovery Act (RCRA), Comprehensive Environmental Response, Compensation and Liability Act (CERCLA), Hazardous Materials Transportation Act (HMTA), Federal Hazardous Substances Act (FHSA), Consumer Product Safety Act (CPSA), Poison Prevention Packaging Act (PPPA), and Occupational Safety and Health Administration [10].

The exposure mitigation methods include the use of material safety data sheet, promulgation of exposure limits, testing for toxicity characteristics to medical surveillance programs, and dozens of other miscellaneous administrative and engineering control initiatives [2, 5, 12, 13]. In order to assist hazardous material management programs nationwide, a list of hazardous material is published by numerous governmental, trade, and professional agencies. Those include but are not limited to (a) 189 hazardous air pollutants in Title III of CAA amendments, (b) Registry of Toxic Effects of Chemical Substances (RTECS) (NIOSH), (c) list of toxic and hazardous substances (OSHA), (d) Pocket Guide to Chemical Hazards (NIOSH), and (e) Threshold Limit Values for Chemical Substances and Physical Agents in the Workroom Environment with Intended Change (published and updated yearly, American conference of governmental and industrial hygienists (ACGIH)) [16].

References

- [1] Manaham, S.E. (1990). *Hazardous Waste Chemistry, Toxicology and Treatment*, Lewis Publishers, Chelsea, pp. 1–2.
- [2] Leonard, J.E. & Robinson, G.D. (2002). *Managing Hazardous Material*, Institute of Hazardous Materials Management, Rockville, pp. 2, 64.
- [3] LaGrega, M.D., Buckingham, P.L. & Evans, J.C. (2001). *Hazardous Waste Management*, McGraw-Hill Companies, New York, pp. 866–870.
- [4] Chechile, R.A. & Carlisle, S. (1991). *Environmental Decision Making A Multidisciplinary Perspective*, Tuft University Center for the Study of Decision Making, Van Nostrand Reinhold, New York, pp. 92–95.
- [5] WoodSide, G. (1993). *Hazardous Materials and Hazardous Waste Management: A Technical Guide*, John Wiley & Sons, New York, pp. 22–25, 53–123.
- [6] Isman, W.E., & Carlson, G.P. (1980). *Hazardous Material*, Glencoe Publishing Company, Encino; Collier McMillan Publishers, London, pp. 2, 57.
- [7] Sullivan, T.F.P. (2003). *Environmental Law Handbook*, 17th Edition, ABS Consulting, Government Institutes, Rockville, pp. 133, 784.
- [8] United States Government Accountability Office (GAO), Report to the Chairman, Committee on Environment and Public Works, U.S. Senate (2006). *Human Health Risk Assessment: EPA Has Taken Steps to Strengthen Its Process, but Improvements Needed in Planning, Data Development, and Training* (GAO-06-595), www.gao.gov/special.pubs/gao-06-637sp/index.html, p. 9–11.
- [9] Schumacher, A. (1988). *A Guide to Hazardous Materials Management: Physical Characteristics, Federal Regulations, And Response Alternatives*, Quorum Books, Blum, Schumacher, & Associates, Green Wood Press, New York, pp. 2, 5.
- [10] United States Environmental Protection Agency (1997). *United States National Profile on Management of Chemicals*, Office of Prevention, Pesticides and Toxic Substances, pp. 29–31.
- [11] Kolluru, R., Bartell, S., Pitblado, R. & Stricoff, S. (1996). *Risk Assessment and Management: Hand Book for Environmental, Health, and Safety Professionals*, McGraw-Hill, New York, pp. 4–14.
- [12] Lee, K.W. (1995). *Practical Management of Chemical and Hazardous Waste, An Environmental and Safety Professional's Guide*, Prentice Hall, Englewood Cliffs, pp. 68–73.
- [13] Introduction to Hazardous Waste Identification (40 CFR Parts 261) (2005). *Solid Waste and Emergency Response* (5305 W), EPA530-K-05-012, pp. 5–7.
- [14] Crosby, D.G. (1998). *Environmental Toxicology and Chemistry*, Oxford University Press, New York, pp. 123, 143.
- [15] Asante-Duah, D.K. (1993). *Hazardous Waste Risk Assessment*, Lewis Publishers, CRC Press, Boca Raton, pp. 283–293.
- [16] Heinsohn, R.J. & Kabel, R.L. (1999). *Sources and Control of Air Pollution*, Prentice Hall, Englewood Cliffs, p. 12.

NASRIN R. KHALILI

Zero Failure Data

This article considers estimation of the reliability of a “system” when no failures have been observed (*see Reliability Demonstration*). The term *system* refers to the hardware under consideration, which may be as simple as a fuse or as complex as an entire nuclear power plant. The “zero-failure-data” case is not the same as the “no-data” problem. In other words, we assume here that the system has been observed for some time T (or demanded N times) but that no failures have been observed. In the no-data situation, the system would never have been observed.

Estimating risk when no failures have been observed is inherently extremely difficult. Unfortunately, the situation is also quite common: any time a first of a kind prototype system or plant is built, the risk prediction is based on no data, and even after operation starts there is a period of zero failures for many components. Yet failure rates or failure probabilities need to be estimated.

Several approaches are possible.

- Be content with a “pessimistic” estimate, an upper bound only.
- Find data for similar equipment, and construct an estimate with the aid of the other data.
- Analyze the parts, ways, or processes by which the system can fail, and use data on these various parts to construct an estimate for the system as a whole.

These three topics form the main sections of this article. A final section discusses the situation when few failures, rather than no failures, are observed.

Pessimistic Estimates

The zero failure data give no information on how small the risk might be. They only give information on how large the risk might be. In other words, the observed absence of failures enables one essentially to rule out very large values of the failure probability or failure rate. Formally, various upper bounds – confidence upper bounds or upper percentiles of Bayesian distributions can be constructed (*see Imprecise Reliability; Lifetime Models and*

Risk Assessment; Mathematics of Risk and Reliability: A Select History; Meta-Analysis in Nonclinical Risk Assessment). However, any small value of the probability of failure or failure rate, no matter how small, is in principle consistent with the observed lack of failures.

Basic Pessimistic Upper Bound and Point Estimate

The upper limit can be defined formally in several equivalent ways: as the smallest value that would be rejected by a hypothesis test, or as a confidence limit, or as the upper end of a Bayes credible interval based on some suitably pessimistic prior. Formulas are simple in two common cases given here.

1. Suppose first that the failures occur according to a Poisson process, or equivalently, the times between failures are independent and exponentially distributed, with failure rate λ . The probability of zero failures in time T is $e^{-\lambda T}$. A hypothesis test would reject a λ value at the significance level 0.05 if this probability is ≤ 0.05 , that is, if $\lambda \geq -\ln(0.05)/T$. Alternatively, a 95% confidence upper bound based on x failures in time T is $\chi^2_{0.95}(2x+2)/2T$ [1], where $\chi^2_{0.95}(df)$ denotes the 95th percentile of a χ^2 distribution with df degrees of freedom. This can also be expressed as $\text{gamma}_{0.95}(x+1, T)$, the 95th percentile of a gamma distribution. Therefore, the 95% confidence upper bound based on zero failures in time T is $\text{gamma}_{0.95}(1, T)$. All these formulas can be shown to give the same numerical limit, $3.00/T$. The upper limit can be found similarly for other significance levels/confidence levels, simply by changing 0.05 and 0.95 to the desired values.

A Bayesian distribution that corresponds exactly to this upper limit for all confidence levels is found as follows (*see also Bayesian Statistics in Quantitative Risk Assessment*). The conjugate prior for λ is $\text{gamma}(\alpha_{\text{prior}}, \beta_{\text{prior}})$, and the posterior distribution based on zero failures in time T is $\text{gamma}(\alpha_{\text{prior}} + 0, \beta_{\text{prior}} + T)$ [2]. Therefore, the prior distribution that yields posterior percentiles that match confidence limits of the above form is uniform [formally, $\text{gamma}(1, 0)$]. The resulting posterior 95th percentile is $\text{gamma}_{0.95}(1, T)$, agreeing exactly with

the above 95% confidence limit, and all the limits for other confidence levels also agree. The mean of this distribution is $1/T$, a value that we might choose as a point estimate.

Another way to think of conservative estimation is as follows: base the decision on the exponentially distributed times to failure. When you observe no failures up to time T , the most conservative assumption imaginable, still consistent with the observation, is to assume that the first failure would occur at the next moment $T+$. The resulting confidence bound is conservative, and happens to be exactly the same as the upper confidence limit given above for the Poisson model [3].

2. On the other hand, failures may occur on demand. The binomial model assumes that failures on distinct demands are independent, and that the probability of failure on any demand is some number, denoted q . The probability of zero failures in N demands is $(1 - q)^N$. A hypothesis test would reject a value of q at the 0.05 significance level if this probability is less than 0.05, or equivalently if $q \geq 1 - 0.05^{1/N}$. Alternatively, the classical 95% confidence upper bound for q , based on x failures in N demands, is the 95th percentile of a $\text{beta}(x + 1, N)$ distribution [1]. Therefore, the 95% confidence upper bound based on zero failures is $\text{beta}_{0.95}(1, N)$. These two approaches can be shown to give the same numerical limit. The limit for significance levels other than 0.05 can be found in the same way.

A Bayesian distribution that corresponds exactly to these upper limits for all confidence levels is easily found. The conjugate prior for q is $\text{beta}(\alpha_{\text{prior}}, \beta_{\text{prior}})$, and the posterior distribution based on zero failures in N demands is $\text{beta}(\alpha_{\text{prior}} + 0, \beta_{\text{prior}} + N - 0)$ [2]. Therefore, the prior distribution that yields posterior percentiles matching the upper confidence limits given above is $\text{beta}(1, 0)$. The resulting posterior 95th percentile is $\text{beta}_{0.95}(1, N)$, which agrees with the 95% confidence limit found above.

The mean of this distribution is $1/(N + 1)$, an estimate that we might use if we assume that the first failure will occur on the very next demand after the N observed demands. This is a pessimistic assumption that may be used in point-value risk assessments.

These bounds and distributions are based on the question, “How large could the parameter be, and how confident can I be about it according to the data alone?” The procedure does not use any external information that might indicate that the parameter actually is substantially smaller than the conservative distributions suggest.

When the observation period is small – a short observation time or a few demands – the conservative upper bound can be unrealistically large, and unacceptable as an estimate. But there are cases when the upper bound is acceptable, especially for components that are not so important to risk.

Less Pessimistic Estimates

If one or more failures are present, we can obtain both upper and lower confidence limits. There is no Bayes prior distribution whose posterior distribution matches both the lower and upper confidence limits for discrete Poisson or binomial observables, but a compromise is the “Jeffreys noninformative prior”, $\text{gamma}(0.5, 0)$ for failures in time and $\text{beta}(0.5, 0.5)$ for failures on demand [2–4] (*see also Bayesian Statistics in Quantitative Risk Assessment*). The $\text{gamma}(0.5, 0)$ distribution is an example of an “improper” distribution, having a density with an infinite integral; however, if it is manipulated formally, the resulting posterior distribution is proper. Therefore, improper priors are allowed (and regarded, if necessary, as limiting approximations of proper distributions).

On the basis of x failures observed in time T or in N demands, the posterior distribution is $\text{gamma}(x + 0.5, T)$ or $\text{beta}(x + 0.5, N - x + 0.5)$. Heuristically, these results correspond to assuming an extra half failure. Depending on the viewpoint, this half failure may be imagined as occurring just after time T or demand N , or prior to the data, in zero time or on one prior demand. The results are less pessimistic than with the priors given above in the section titled “Basic Pessimistic Upper Bound and Point Estimate”.

Consider first the Jeffreys prior for failures in time. If zero failures are observed in time T , the posterior distribution for the failure rate λ is $\text{gamma}(0.5, T)$. The posterior mean is $1/(2T)$, half the pessimistic mean. The 95th percentile is 0.64 times the 95th percentile when the most pessimistic prior is used. With the Jeffreys prior for failures on demand, the posterior mean is $1/(2N + 1)$, approximately half the

pessimistic mean. The 95th percentile is smaller by a factor 0.81 than the 95th percentile when the most pessimistic prior is used and when $N = 1$, and by a factor 0.64 when N is large.

Another way to the above conclusions is as follows. In case of at least some failures ($x > 0$), it is also possible to estimate the failure rate using *failure truncated* data (see **Reliability Growth Testing; Repairable Systems Reliability**), in which case the observable is the time to the x th failure, T_x , a continuous variable. This has an Erlang (or Erlangian) distribution $\text{gamma}(x, \lambda)$, and all confidence bounds for λ are given by the distribution $\text{gamma}(x, T_x)$, or equivalently $\chi^2(2x)/(2T_x)$. The prior density that yields this as a posterior in the Bayesian process is $\text{gamma}(0, 0)$, the inverse of the failure rate. Instead, when the observations are *time truncated* at some fixed T , with x failures observed, one can say for sure that T is longer than the time to the x th failure and shorter than the time to the $(x + 1)$ th failure. As a heuristic compromise, T is formally the observed time to the $(x + 1/2)$ th failure. This yields confidence bounds based on $\chi^2(2x + 1)/(2T)$, consistent with the Jeffreys prior for Poisson data, $\text{gamma}(0.5, 0)$ [3].

Estimates Based on Data from Similar Equipment

When failure data are available on similar components under reasonably similar conditions, the methodology depends strongly on the degree of similarity, and on the detail of the data. If data can be assumed to be from identical systems and conditions as the system of interest, one may pool the numbers of failures and observation times (or numbers of trials) and use well-known formulas to estimate λ or q . This highest degree of similarity is rarely the case. The next two subsections consider two cases with equipment that is similar to, but not identical to, the equipment of interest.

Detailed Data from Similar Systems

Suppose that it is possible to find failure data on a family of equipment whose members are not identical but are “similar enough” to the equipment of interest, in the sense that one can consider the latter as a sample from the same family. For example, steam generators used at WWER (water-cooled

water-moderated energetic reactor) nuclear power plants are quite similar, but maintenance practices and water chemistry vary among the plants, causing variation in tube rupture frequencies. It is reasonable to assume each plant to be a sample from the same family. This allows use of empirically based Bayesian methods, with either “empirical Bayes” (see **Repairable Systems Reliability; Common Cause Failure Modeling; Cross-Species Extrapolation**) or “hierarchical Bayes” implementation (see **Bayesian Statistics in Quantitative Risk Assessment**) [3, 5]. This approach can be used in estimating tube rupture frequencies for each plant, including the plant of interest. It satisfies those who want to use available data to the maximal extent possible. This procedure also yields proper values for plants that have themselves experienced no failures or even have no data at all.

Of course, there is subjectivity in judging which units form a family of “similar enough” units to form a sampling/prior distribution. But there are ways to experiment with alternative families to sort this out. For example, one can try failure data from *all* pressurized water reactor steam generators and see how the family distribution differs from that obtained for WWER units alone. Presumably, the latter is narrower because the units are the same type and more similar. Consequently, the posterior for the unit of interest is more accurate. In any case, even a wider sampling/prior distribution is usually better than any noninformative prior distribution.

Even in the empirical Bayes or hierarchical Bayes cases, the lower bounds cannot necessarily be estimated accurately. The values can depend strongly on the assumed model. An example is given in [5] of failure to start, with 68 plants, hundreds of demands, and only six failures. The problem is to estimate q , the probability of failure to start, at each plant. Two hierarchical Bayes analyses (see **Common Cause Failure Modeling; Meta-Analysis in Nonclinical Risk Assessment**) are performed, one assuming that q follows a beta distribution between plants and the other assuming that q follows a logistic-normal distribution. For plants with zero failures, the 95th percentiles agree for the two models to within a factor of 2, and the means agree within a factor of 3. The fifth percentiles, on the other hand, disagree by over 10 orders of magnitude. Similar behavior can be noticed in other comparisons [6].

For additional real applications of these methods in risk assessment, the reader is referred to [6–9].

Summary Data from Similar Systems

Sometimes, failure data or rate estimates for certain component types are available from various branches of industries, even if not exactly for the plant or industry under investigation. Such data also may be available in summary form, such as a value from a handbook. This kind of generic data may be too scattered and diverse for formal manipulation. “Engineering judgment” has been used for years to obtain meaningful central estimates and uncertainty bounds, accounting for similarities and differences of systems and operating conditions.

One less subjective way would be to use a so-called constrained noninformative prior [10]. This is a prior distribution that maximizes the entropy (in a suitable parameterization) subject to the constraint that the mean equals a known value. It is a generalization of the Jeffreys prior, adding the constraint of the mean. For highly reliable equipment, the Jeffreys prior is typically overly pessimistic, and the constrained noninformative prior seems to be inherently more realistic, at least with respect to the mean. This prior can then be updated using the data from the system of interest.

Analysis and Synthesis of the Failure Processes

A final method used for systems with no observed failures is to analyze the ways that failure can occur, possibly decompose the system into more primitive pieces, and use available data on these processes or contributors to construct an estimate for the system. Several of the many applications of this method are pointed out here.

A notable success in this regard is probabilistic safety assessment (PSA) in the nuclear power industry. The methodology is also known as *probabilistic risk assessment* (PRA) (see **Probabilistic Risk Assessment**). The first well-known major study in the United States [11] was completed in 1975. That study was largely based on generic failure probability estimates for components such as pumps, valves, diesel generators, etc. based on median estimates of the type described in the section titled “Summary

Data from Similar Systems”, and similarly for the rates of initiating events like pipe breaks and loss of electric power. Uncertainty studies were performed by propagating the uncertainty distributions of the input parameters through the model.

Since then, most commercial plants have performed wide-scope risk studies and established extensive failure and maintenance data collection systems. Consequently, the methodologies of the section titled “Detailed Data from Similar Systems” can be widely used. Although zero core-damage accidents have occurred in plants with modern design and upgraded safety features, the accident frequency can be estimated with ever-increasing accuracy.

Analysis of the processes contributing to failure is also the basis of load-strength interference. For example, practically all external event PSAs (seismic, extreme winds, snowstorms, seaweeds, floods, etc.) consider events that have never happened (at the site under study). The observed intensities (loads) of these events need to be extrapolated to such values that they could cause failures. The strength measures (or fragilities) also have distributions that need to be estimated. Convolution integrals of the loads and strengths then yield failure probabilities for components and structures [12].

Similarly, very large pipe breaks have never occurred in commercial nuclear reactors. However, degradation may be based on mechanistic structural failure phenomena (fatigue, wear, stress-corrosion cracking, flow-induced thinning, etc.), which in turn are based on material properties and system characteristics (flow, pressure, temperature, coolant chemistry, etc.). Interference with strength characteristics then yields failure probabilities [13].

Comparison of “Few” Failures with No Failures

If a few failures have been observed, rather than none, some of the above discussion remains essentially unchanged:

- Experience from outside the system may still be very useful in estimating risk, using the Bayesian approach. In particular, the methods given above in the section titled “Estimates Based on Data from Similar Equipment” are still useful.
- Decomposing the system into more primitive pieces, with data available for each piece, may

also be useful, as in the section titled “Analysis and Synthesis of the Failure Processes”.

In other respects, however, having at least one observed failure is much more useful than having none:

- We no longer need to be content with an upper bound only.
- A Bayesian conclusion is less sensitive to details of the prior. If a conjugate gamma or beta distribution is used, and x failures are observed, the posterior first parameter is $(\alpha_{\text{prior}} + x)$. This posterior distribution is much less dependent on α_{prior} for $x \geq 1$ than for $x = 0$. As an example of the latter, [14] assumes that zero failures occur in some testing period, and this leads to a struggle with the sensitivity of the prediction results to the value of α_{prior} .
- The sensitivity to the prior is further magnified when the prior is used repeatedly in an analysis, for example, when several failure modes are considered in a single analysis.

References

- [1] Johnson, N.L., Kotz, S. & Kemp, A.W. (1992). *Univariate Discrete Distributions*, 2nd Edition, John Wiley & Sons, New York.
- [2] Box, G.E.P. & Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Reissued 1992 as Wiley Classics Library Edition.
- [3] Vaurio, J.K. (1987). On analytic empirical Bayes estimation of failure rates, *Risk Analysis* **7**, 329–338.
- [4] Jeffreys, H. (1961). *Theory of Probability*, 3rd Edition, Clarendon Press, Oxford.
- [5] Atwood, C.L., LaChance, J.L., Martz, H.F., Anderson, D.J., Englehardt, M., Whitehead, D. & Wheeler, T. (2003). *Handbook of Parameter Estimation for Probabilistic Risk Assessment*, NUREG/CR-6823, U.S. Nuclear Regulatory Commission, at <http://www.nrc.gov/reading-rm/doc-collections/nuregs/contract/cr6823/>.
- [6] Vaurio, J.K. & Jänkälä, K.E. (2006). Evaluation and comparison of estimation methods for failure rates and probabilities, *Reliability Engineering and System Safety* **91**, 209–221.
- [7] Vaurio, J.K. (2005). Uncertainties and quantification of common cause failure rates and probabilities for system analyses, *Reliability Engineering and System Safety* **90**, 186–195.
- [8] U.S. Nuclear Regulatory Commission (1998–2001). *Reliability Studies*, NUREG/CR-5500, An eleven-volume report, reprinting various system studies, at <http://nrcoe.inl.gov/results/index.cfm?fuseaction=SysStudy.showMenu>.
- [9] Poloski, J.P., Marksberry, D.G., Atwood, C.L. & Galyean, W.J. (1999). *Rates of Initiating Events at U.S. Nuclear Power Plants: 1987–1995*, NUREG/CR-5750, U.S. Nuclear Regulatory Commission, at <http://nrcoe.inl.gov/results/index.cfm?fuseaction=InitEvent.showMenu>.
- [10] Atwood, C.L. (1996). Constrained noninformative priors in risk assessment, *Reliability Engineering and System Safety* **53**, 37–46.
- [11] U.S. Nuclear Regulatory Commission (1975). *Reactor Safety Study: An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants*, WASH-1400, NUREG-75-014.
- [12] Fullwood, R.R. & Hall, R.E. (1988). *Probabilistic Risk Assessment in the Nuclear Power Industry*, Pergamon Press, New York.
- [13] Lewis, E.E. (1987). *Introduction to Reliability Engineering*, John Wiley & Sons, New York.
- [14] Coolen, F.P.A., Coolen-Schrijner, P. & Rahrouh, M. (2005). Bayesian reliability demonstration for failure-free periods, *Reliability Engineering and System Safety* **88**, 81–91.

CORWIN L. ATWOOD AND JUSSI K. VAURIO